# Survival of Defeat

## Evolution, Moral Objectivity, and Undercutting

Michael Klenk

# Survival of Defeat

Evolution, Moral Objectivity, and Undercutting

Cover art:

A duck billed platypus. Colour lithograph after W. Kuhnert (CC BY)

Cover design by Rosalie Schneegaß

ISBN 978-94-6103-068-9

# Survival of Defeat

## Evolution, Moral Objectivity, and Undercutting

## Weerlegging Overleven

### Evolutie, Morele Objectiviteit en Ondermijning

(met een samenvatting in het Nederlands)

## Widerlegen Überleben

### Evolution, Moralische Objektivität und Untergraben

(mit einer Zusammenfassung in deutscher Sprache)

**Proefschrift**

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. H.R.B.M. Kummeling, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op vrijdag 29 juni 2018 des middags te 12.45 uur

door

**Michael Bernhard Otto Theodor Klenk**

geboren op 10 november 1989

te Heilbronn-Neckargartach, Duitsland

Promotor: Prof.dr.mr. H. Philipse
Copromotor: Dr. H.C. Sauer

This work is part of the research programme 'Evolutionary Ethics? The (Meta-) Ethical Implications of Evolutionary Explanations of Morality' with project number 360-20-340, which is financed by the Netherlands Organisation for Scientific Research (NWO).

# Acknowledgements

During my early years at school, attention to my academic prospects was high, though too often for the wrong reasons. Only optimists would have seen it in a positive light: with so many enforced over-hours spent at school, something good had to come of me! I dare say the optimists were right. Finishing this dissertation feels like the erstwhile culmination point of a formative and exciting journey, and I wish to express my gratitude to the many people who have put their trust in me along the way and who gave me a chance to thrive.

Herman Philipse, my thesis supervisor, opened the door to academia for me. Herman believed in my academic and professional abilities from the start and meticulously and patiently commented on my research output. While he claims that his first principle in academia is to 'Have fun!' I believe he really means 'Have fun – and perform!' He helped me achieve both aims by giving me the freedom to develop my own views, and by holding me to a pristine standard when he challenged me to defend them. Becoming a philosopher under Herman's tutelage was much fun indeed, and I am deeply grateful for his guidance and exceptional support. Hanno Sauer, my co-supervisor, provided thoughtful criticism, sound pragmatic advice, and, especially in the later stages of my research, encouragement. He helped me set aside doubts and hesitation, and to successfully wrap up this project.

I conducted my research within the 'Evolutionary Ethics' research programme. The team, consisting of Julia Hermann, Jeroen Hopster, Wouter Kalf, and Joeri Witteveen, provided feedback and, due to the many conferences we organised, one of the best environments to study evolutionary ethics worldwide. Wouter greatly helped me to develop academically with his sharp philosophical mind in many discussions, especially at the beginning of my PhD. With Jeroen, I shared a house during our time at Harvard. It was perhaps the most intellectually fruitful and stimulating time I ever had, filled with talks, seminars, and writing from early till late. Jeroen proved not only to be an excellent housemate and discussant, but also a good friend.

I had the good fortune of being able to spend time at research institutions abroad while working on this thesis. I am thankful to Selim Berker, Justin Clarke-

Doane, and Guy Kahane for taking time out of their busy schedules to advise me during my research stays at the universities of Harvard, Columbia, and Oxford, respectively. Our discussions sharpened my research questions and helped me to look for answers in the right places. I also learned much from the incredibly open academic communities at all three universities.

None of this would have been possible without the generous funding provided by the Netherlands Organisation for Scientific Research (NWO). I am immensely grateful to NWO for enabling me to carry out my research without worry, and to the Department of Philosophy and Religious Studies for providing a fantastic place to work. Marcus Düwell, in particular, offered help when I needed it and Biene Meijerman and Suzanne van Vliet kept things running organisationally with the utmost supportiveness. I found a place for stimulating conversations and refreshing banter in the department's attic, where most PhDs and Postdocs work. With Hein Duijf, I discussed just about every chapter in this thesis, and much else besides. He always believed that he was right, and I very much tried to prove him wrong. Of course, I am agnostic as to who was right, but confident that our discussions made writing this dissertation much more fun and the result better. Our games of foosball were instrumental, too, as they offered respite especially in the final phase of work on my dissertation. Some even rumour that having advocated successfully for a foosball table in the attic will have been my most significant achievement. I do hope that it brings joy to many future generations of attic dwellers, as it did to us, but also that the rumours will be wrong and that I will move on to even higher things in life.

Looking back again to the time before I came to Utrecht to write this dissertation, there are many more Zeit-Genossen that I met in sports clubs, universities or firms, who made the journey exciting and meaningful. I thank my parents Waltraud and Bernhard and my sister Mona Klenk, as well as my friends from Oberstenfeld, the 'AV Hütte', from near and far, for making my life colourful, grounded, and worthwhile. Johanna Neubauer, with you by my side I discovered philosophy and the Netherlands with the greatest joy; we grew together, and you supported me beyond compare.

# Table of Contents

This page intentionally contains only this sentence.

# Introduction

We can't escape our past. Our upright posture, the size of our brain, and the position of our thumbs are consequences of our evolutionary history.[1] The influx of evolutionary forces does not stop at our bodies – the way we behave and our psychology have been shaped by evolution too. Charles Darwin, the great biologist, boldly suggested that the theory of evolution by natural selection also explains the human capacity to think in moral terms and the content of our moral rules:

> If … men were reared under precisely the same conditions as hive-bees, there can hardly be a doubt that our unmarried females would, like the worker-bees, think it a sacred duty to kill their brothers, and mothers would strive to kill their fertile daughters; and no one would think of interfering. (Darwin 1871 [2004]: 70)

Evolution shapes our norms, and this suggests that an alternative evolutionary trajectory would have left us with different moral norms, or so Darwin argued. The magnitude of the implications of Darwin's thesis about the origin of humans and morality were judged to be "impossible to overestimate".[2] The *Edinburgh Review* wrote of Darwin's *The Descent of Man*:[3]

> If our humanity be merely the natural product of the modified faculty of brutes, most earnest-minded men will be compelled to give up those motives by which they have attempted to live noble and virtuous lives, *as founded on a mistake*; our moral sense will turn out to be a mere developed instinct, identical in kind with those of ants or bees; and the revelation of God to us, and the hope of a future life, pleasurable daydreams invented for the good of society. *If these views be true, a revolution in thought is imminent, which will shake society to its very foundations.* (1871: 195–6 emphasis added)

Initially, the moral consequences of Darwin's thesis were of quite a different kind. Rather than turning people away from morality, as the *Edinburgh Review* feared, new moral theories were erected on the basis of Darwin's theory. That is,

[1] Cf. Young (2003).
[2] Cf. Ruse and Richards (2017); Secord (2003).
[3] See also James (1902 [2002]: 13–4).

evolutionary *explanations* were taken to *justify* particular moral theories. A Cambridge student's recollection of the reception of Darwin's theory nicely illustrates the enthusiasm with which many adopted the new way of thinking about ethics:

> We seemed to ride triumphant on an ocean of new life and boundless possibilities. Natural Selection was to be the master-key of the universe; we expected it to solve all riddles and reconcile all contradictions. Among other things it was to give us a new system of ethics, combining the exactness of the utilitarian with the poetical ideals of the transcendentalist. We were not only to believe joyfully in the survival of the fittest, but to take an active and conscious part in making ourselves fitter. (Clifford 1879: 33)

Most academic philosophers were not impressed. In the wake of Sidgwick and Moore, philosophers rebuked such prescriptive evolutionary ethics as fallacious and deeply misguided.[4] Moore concluded "that Evolution has very little indeed to say to Ethics" (Moore 1903 [1988]: 2). Many years later, Ludwig Wittgenstein expressed a common sentiment of philosophers working in the analytic tradition when he wrote that "Darwin's theory has no more to do with philosophy than any other hypothesis in natural science" (Wittgenstein et al. 1921 [2006]). While many analytic philosophers might have evinced deep respect for the empirical and formal sciences in general[5], they were nevertheless committed to a deep anti-psychologism, the idea that facts about psychology (and thus about evolutionary psychology) are irrelevant to questions of moral truth and justification (Kusch 2006).[6]

In recent years, analytic philosophers have shed their scepticism towards the philosophical relevance of genealogical data and have begun to consider seriously the idea that revealing the causal origins of a belief may have (negative)

[4] I will discuss this in greater detail in chapter 1, section 1.3.

[5] Rorty even sought to define analytic philosophy in terms of its respect for science; see Rorty (2008: 220).

[6] Though largely absent from the debate in analytic philosophy, the suspicion that the contingencies of language, culture, and upbringing have consequences for what we *ought* to believe has loomed large in the history of European intellectual thought (Leiter 2004). Friedrich Nietzsche coined the term 'genealogy' as applied to concepts, beliefs, and values in his endeavour to discredit Christian values based on an analysis of their origins (Nietzsche 1887 [2013]). Following Nietzsche, Freud predicted that "our attitude to the problem of religion will undergo a marked displacement" once we learn about the psychological dispositions that make us religious (Freud 1927: 215).

implications for the epistemic credentials of holding that belief.[7] Genealogical scepticism about philosophy, or at least one or more of its domains, is increasingly common (Srinivasan 2015).

In particular, many ethicists are deeply impressed by the availability of evolutionary explanations of morality, which have improved dramatically in quality and scope since Darwin's time.[8] Not only is the theory of evolution by natural selection one of the most explanatorily powerful theories of science (Jones 2001; Thompson 2015),[9] and confirmed beyond serious doubt (Endler 1986), the evolution of the human capacity to think in moral terms is now better understood than ever (e.g. Verplaetse et al. 2009). Such evolutionary explanations of morality have tempted many ethicists to draw conclusions about the *justification, truth,* or the *construal* of the contents of moral judgements, both on the local level of specific types of moral judgements (Greene 2008; Lazari-Radek and Singer 2012) and on the global level, concerning *all* moral judgements (Joyce 2006; Ruse and Wilson 1986; Street 2006).[10] This is by no means an exhaustive list, and there are important differences between these alleged implications for (meta)ethics and the way in which philosophers have argued for them.[11] The underlying thought, however, is similar: we can't escape our evolutionary past – and this is supposed to have significant metaethical consequences.

My focus in this thesis will be on arguments that imply that *all* moral judgements are *unjustified* in light of evolutionary explanations of morality. If such arguments succeed, they would *defeat* our justification for holding moral

[7] For example, genealogically motivated arguments have been raised against metaphysics (Ladyman and Ross 2010), logic, naturalism (Plantinga 1993: ch. 12), and theism (Dennett 1995). Experimental philosophy, often concerned with pointing out how philosophical intuitions vary with seemingly irrelevant background factors, may also be seen as an example of genealogical criticism, see Doris and Stich (2012); Knobe and Nichols (2008); Knobe and Nichols (2014).

[8] E.g. Bogardus (2016); Crow (2016); Gibbard (2003), Greene (2008, 2013), Joyce (2001, 2006, 2016a, 2016d), Kitcher (2011), Ruse (1995a, 2006), Ruse and Wilson (1986); Singer (2005), Street (2006, 2008a, 2008b, 2016).

[9] See Klenk (2016b) for a review.

[10] Cf. Philipse (2015) for a discussion of the *global* vs *local* distinction amongst evolutionary challenges.

[11] This is by no means an exhaustive overview, and there are significant differences between these alleged implications for (meta)ethics and the way in which philosophers have argued for them. I will carefully tease them apart in chapter 1, section 1.4.

judgements. However, many have argued that these arguments do not succeed. These dissenting voices claim that there is no sound way in which evolutionary explanations of morality defeat the justification of all our moral judgements. Thus, the 'survival of defeat' based on evolutionary considerations of morality concerning all moral judgements is in jeopardy. My specific aim is to evaluate whether, and how, all moral judgements could be shown to be unjustified in light of evolutionary explanations of morality and thus to show what the conditions are for the survival of evolutionary defeat.[12] To fully introduce my research question, I will provide a first impression of (1) evolutionary explanations of morality, (2) illustrate why they seem to raise a challenge for the justification of our moral beliefs, and (3) suggest why we should be sceptical about such arguments.

## Evolution and Morality

The leading hypothesis about the evolution of morality is that effective cooperation was evolutionarily advantageous and that the capacity to make normative judgements, including moral ones, made our ancestors more effective co-operators that were able to coordinate actions in mutually beneficial ways (cf. Gibbard 1990: 107–8).[13] This picture has two essential parts: the (evolutionary) advantageousness of cooperation and the role of moral judgements in making cooperation more effective. Various lines of research suggest that cooperation played an important role in our evolutionary history and that effective cooperators were more likely to survive and reproduce than less effective cooperators.

Consider the allocation of scant resources as an example. A successful hunter might share the spoils of a hunt with the less fortunate members of his or her group. Though this incurs some costs to the successful hunter of today, he will probably be unsuccessful, and hungry, in the future and then stand to benefit when others reciprocate and share their spoils too (Boehm 2012: ch. 7; Hill 2002). Individuals might thus gain from cooperating and groups consisting of cooperators

[12] Except in headings, I will henceforth drop the qualification ‚evolutionary' when writing about the survival of evolutionary defeat for ease of expression.

[13] E.g. Axelrod (2006); Axelrod and Hamilton (1981); Bowles and Gintis (2011); Curry (2016); Kitcher (2011); Machery and Mallon (2010); Trivers (1971). See also Joyce (2006: 117), Kitcher (2011: 5–6), and Barkhausen (2016: 664–72) for discussions in philosophy.

may be more successful than groups of egoists, or so the theory goes.[14] The problem is that effective cooperation is threatened by free-riders: individuals enjoying the collaborative gains (e.g. receiving a share of the spoils of a hunt) without contributing themselves (e.g. they would not share their own game).

The second part of the picture of the evolution of morality is that a capacity to make moral judgements functions, in rough analogy, like a tool to enhance cooperation. For example, a capacity to grasp and to internalise norms can be shown to have various effects on motivation and behaviour, making it less likely that those who make relevant moral judgements free-ride in cooperative endeavours (Bicchieri 2006; Joyce 2006: ch. 1). The *capacity* to make normative judgements, including moral ones, can thus be considered as a tool that helps people to be cooperative. The ability to cooperate effectively seems to have been conducive to evolutionary fitness because it allowed our ancestors to, for example, hunt larger game, transmit knowledge more effectively, or persist in inter-group competition (cf. Bowles and Gintis 2011; Sterelny 2012; Tomasello 2016).

Moreover, the *content* of some normative judgements is plausibly influenced by evolutionary forces too. For example, some basic evaluative tendencies, such as the urge to care for one's offspring and aversions to sexual relations with siblings, are plausibly seen as having a biological basis because having these tendencies helped our ancestors to get their genes into the next generation (Low 2015).[15] So, both the capacity to think in normative terms and the content of some of our moral beliefs may reflect their 'evolutionary past', just like Darwin suggested.[16]

## Evolutionary Defeat of Moral Judgements

Such *evolutionary explanations of morality raise an epistemological challenge for the justification of our moral judgements*.[17] If something like the above story is

[14] This can be shown through game-theoretic models and it bears out in the ethnographic record; see Bowles and Gintis (2011) and Baumard (2016) as well as Klenk (2016d) for a review. I will discuss this more in chapter 1.

[15] See Klenk (2015b) for a review.

[16] I will say more about this in chapter 1, section 1.2.

[17] This is but one version of a metaethical challenge that can be raised based on evolutionary explanations of morality. I will precisify the relation of the problem that I address to extant arguments in the literature, specifically those of Street (2006, 2008b),

correct, then at least some of our moral judgements appear to be contingent on our evolutionary history, because they are aimed at detecting 'factors that aided cooperation' in our evolutionary past. The problem is twofold. First, the 'factors that aid cooperation' might be different, had we taken a different evolutionary path and so we might have been prone to make different moral judgments today. Second, the 'factors that aid cooperation' are not per se morally good. For example, a culture of honour in some Mediterranean societies prescribed that a man who 'dishonoured' another man's unmarried daughter or sister had to marry her or otherwise pay for his rashness with his life, irrespectively of whether sex was consensual and amongst adults.[18] Making moral judgements in accordance with such norms might have upheld social order in these societies, and thus might be said to fulfil their 'evolutionary function', but endorsing them seems morally wrong nonetheless: it is nobody's business to intervene in the romantic lives of consenting adults, let alone by murdering them. In consequence, evolutionary theory shows us that we might be disposed to endorse moral claims because doing so was or is helpful for cooperation, and *not* because these claims are correct. This implies that our moral judgements might be off track in regards to the moral truth and this raises a problem about the justification of our moral judgments.[19] That is, the evolutionary story does not threaten the *existence* of moral norms or facts, but whether we are capable of having justified beliefs about them: the problem is epistemological, rather than metaphysical. There are many ways to explain why the justificatory problem arises precisely, but often they have in common their final destination: evolutionary considerations *defeat* (to wit, take away or reduce) the justification of our moral judgements.[20] How this works, precisely, is of course the main question of this thesis. I want to illustrate the alleged problem with an analogy and an example.

---

Joyce (2006), and Ruse (2006), in section 1.4. Doing so now would unnecessarily distract from describing the main problem addressed in my thesis.

[18] See Appiah (2011); Bicchieri (1990); Frank (1988); Nisbett and Cohen (1996) for relevant discussion.

[19] I will use 'moral judgments' and 'moral beliefs' interchangeably throughout this thesis. As will become clear in the section on Method and Presuppositions in this introduction.

[20] I will disentangle different proposals about how evolutionary explanations have an effect on the justification of our moral judgments in chapter 1, section 1.4.

Consider first an analogy that illustrates the problem. Suppose you are on a factory visit and a conveyor belt carries what look like red wedges to you. Under normal circumstances, you seem entirely justified in believing that the wedges are red. Later the foreman tells you that, for technical reasons, the conveyor belt is illuminated by a red light. The foreman's testimony implies that you should not be sure that the wedges *seeming* red to you is a good ground for thinking that they *are* red.[21] Evolutionary considerations about our moral judgements might play a role similar to the foreman's testimony. Though many moral judgements *seem* correct to us, particularly those based on reflection and deliberation, they might seem correct to us because it was evolutionarily advantageous for us to think so and not because these judgements *are* correct.

Moving beyond an analogy, one of the ways in which evolution is supposed to defeat the justification of moral judgements is by raising a problem about the *reliability* of moral judgements.[22] This 'reliability view' of evolutionary defeat goes roughly as follows. First, we learn that evolutionary factors influenced our moral judgements and so we have to say why we can still expect that our moral judgements are reliable, in the sense of being correct more often than not. Second, we cannot justifiably maintain our moral judgements if it is in principle impossible to explain their reliability (cf. Field 1989). Third, evolutionary explanations of morality take away one possible explanation of our moral reliability (they show that we have no reason to think that evolutionary forces are likely to lead to true moral beliefs) and so they might show that it is impossible to explain the reliability of our moral judgements. The details of the reliability view are of course more complicated, but it illustrates one concrete way in which evolutionary explanations of our moral judgements are thought to have justificatory, metaethical relevance.

[21] In this case, I rely on an intuitive distinction between the colour of an object seen in normal circumstances and the colour the object has in virtue of abnormal lighting conditions.

[22] See Schechter (2010), Tersman (2016), Dogramaci (2017), Bedke (2014) Fraser (2014), Gibbard (2003: ch. 13), or Talbott (2015) for exemplary discussions of the reliability view.

## Endagered: The 'Survival of Defeat'

However, there are *reasons to be sceptical about attempts to defeat the justification of moral judgements based on evolutionary explanations of morality,* and they have to do with the plausibility of the epistemic principle invoked by these arguments. An epistemic principle about justification specifies conditions relevant for the justification of a (moral) judgement. For example, the reliability view on evolutionary defeat relies on the epistemic principle that <the judgement that p is justified if and only if it is in principle possible to explain the reliability the judgement that p>. It has been argued that worries about reliability cannot defeat our moral judgements because the supporting epistemic principle is false: we need not give up our moral beliefs even if we cannot in principle explain their reliability.[23] Others accept the principle invoked by the reliability view but claim that evolutionary explanations do not rule out such explanations: there are legitimate ways to explain the reliability of our moral judgements in light of evolutionary explanations and so we need not give up our moral judgments.[24] Either way, defeating moral judgements according to the reliability view is supposed to fail. Of course, alternatives to the reliability view exist, and there are other proposals about how to get from evolutionary explanations of morality to a justificatory loss for moral judgements.[25] When we consider these proposals in greater detail, beginning in chapter 1, we will see that there are good reasons to question whether there is a sound, plausible, or interesting way in which metaethical conclusions can be drawn from evolutionary explanations of morality. Thus, the survival of defeat is in jeopardy.

## Aims of the Thesis

I aim to find out in this thesis how we get from evolutionary explanations of morality to a justificatory loss for moral judgements. On that path, we have to pursue two investigations for each proposal about the justificatory implications of evolutionary explanations of morality: is the proposal based on a true epistemic

[23] E.g. Bogardus (2016), Clarke-Doane (2015, 2016a).
[24] E.g. Enoch (2010); Skarsaune (2011); Wielenberg (2010).
[25] E.g. Joyce (2006); Mogensen (2016a); Street (2006); Vavova (2014a).

principle and is that principle shown to be violated by evolutionary explanations of morality? The question mark in Figure 0.1 illustrates the basic question about whether there are good epistemic reasons to get from evolutionary explanations of morality to a justificatory loss for moral judgements:[26]

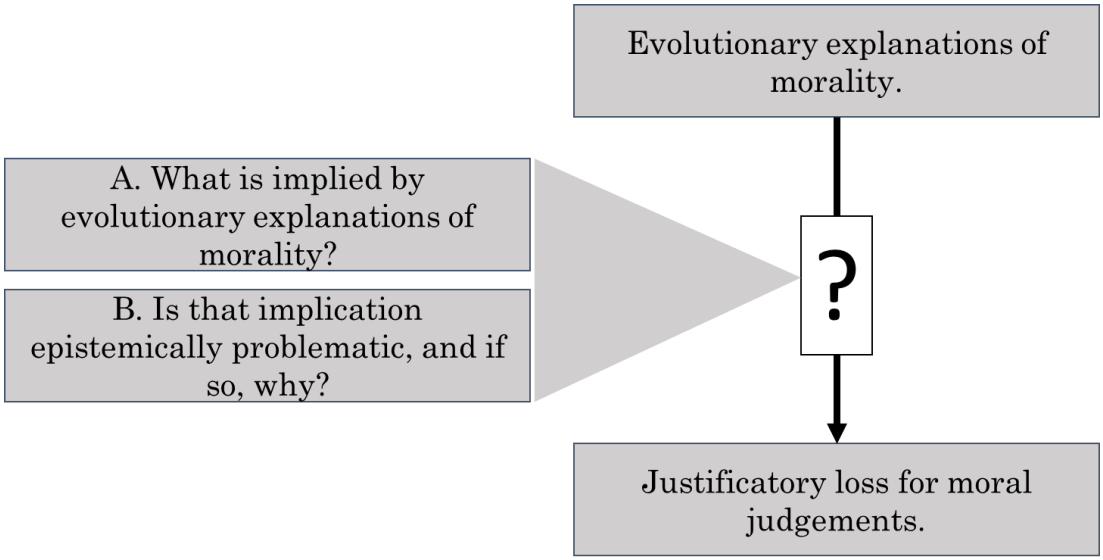


*Figure 0.1 How to get from evolution to loss of justification?*

In order to make my research question more specific, I want to introduce three qualifications to that broad question. First, an obvious reason for giving up all our moral judgements would be if evolutionary explanations of morality show that all our moral judgements are false.[27] As I show in section 1.5 of chapter 1, evolutionary explanations do not show that our justified moral beliefs are false, and thus I assume that the survival of defeat depends on inferring a justificatory loss for moral judgements *without showing that they are false*. Moreover, I assume that at least some moral judgements are *defeasibly* and *non-empirically* justified. A belief is defeasibly justified if its justification can be lost, for example, upon learning new information. A belief is non-empirically justified if it is justified and its justification is not implied by the best explanation of our observations.[28] Good candidates for

[26] We first have to answer question A and then, to answer question B, determine whether whatever is found to be implied by evolutionary explanations of morality is epistemically problematic.

[27] See Bicchieri (2017: 11) and Sinnott-Armstrong (2011: 20) for endorsements of this view.

[28] This is, of course, the category of a priori justified beliefs; see BonJour and Sosa (2003: 5–6). Many moral objectivists claim that at least some moral beliefs are non-empirically justified (Clarke-Doane 2016a: 25; Enoch 2011b; Shafer-Landau 2003). See BonJour (1998) for a defence of non-empirical justification in general.

defeasible, non-empirically justified moral judgements are fundamental moral judgments about, for example, the value of life, the equality postulate of treating persons with equal concern and respect, the view that we ought to pursue the means to our ends, pro tanto reasons to avoid harm, and so on. I can now state my research question in full detail:

> How can evolutionary explanations of morality *defeat* the justification of all *defeasibly* and *non-empirically justified* moral beliefs without showing that these beliefs are *false*?

For ease of expression, I will say that giving a positive answer to the question 'Can evolutionary explanations of morality undercut the justification of all moral beliefs?' constitutes an *evolutionary defeat challenge*. So, my project can also be understood as an inquiry into whether an evolutionary defeat challenge, thus understood, succeeds. This is means that I *assess the philosophical merits of the evolutionary defeat challenge (a particular epistemological challenge)* to *all moral judgements.* [29] Will defeat survive? There is both a yea and a nay response to this question, and neither has won the upper hand on this hotly debated issue in metaethics. I will briefly preview my thesis before giving a more detailed overview at the end of this introduction.

In seven chapters, I explore various questions related to this project. My strategy is to consider in detail the nature of undercutting defeat and to ask whether current accounts support the evolutionary defeat challenge. In the first four chapters of the thesis, I discuss in greater detail some threats to the survival of defeat, the conditions of defeat in general, the specific nature of evolutionary defeaters, and whether the justification of moral judgements could be reinstated if the evolutionary defeat challenge succeeds. In the remaining three chapters, I discuss in detail the two most promising accounts of the evolutionary defeat challenge and argue that both fail (that is, they do not show that all our moral judgements are defeated). The reliability view, discussed above, fails because our beliefs can be shown to be reliable in the relevant sense. The 'disagreement view',

[29] To be precise, the challenge is whether non-logically-tautologous moral judgments are defeated. Hence, beliefs with contents such as <Mother Teresa was a good person or she was not a good person> would of course remain justified.

says that evolutionary defeat comes by way of an epistemically troubling kind of counterfactual disagreement, which is 'revealed' by evolutionary explanations of morality, and for that reason undercuts moral beliefs. I will argue that the disagreement view fails, because evolutionary explanations do not imply defeating counterfactual disagreement about all our moral beliefs. This is not the end of evolutionary defeat, however. I argue that one interpretation of the challenge is still alive, but it succeeds only on a novel and broad construal of undercutting defeat, which I call the 'etiquette conception' of defeat. According to the etiquette conception, we ought to give up a belief when we learn that the belief does not qualify as knowledge, even though we might have no reason to doubt that the belief is justified and true. I call this the etiquette conception because it may seem as if a belief qualifying as knowledge is a superfluous extra if all that we want from a belief is the belief to be justified and true.

The most significant upshot of my thesis is that it shows that the best current accounts of defeat, which are often invoked but rarely discussed in detail in discussions of evolutionary challenges, are inadequate to support an evolutionary defeat challenge and that it points out what kind of account of defeat we would have to adopt to make the challenge succeed.

The focus of my contribution is to support my overarching claim about the relevance of the etiquette conception. The individuals chapters together show that the etiquette conception of defeat is required for the evolutionary defeat challenge to succeed and thus to keep open the 'survival of defeat'. Put simply, the overarching implication of my thesis is the following view:

> Evolutionary explanations of morality undercut all defeasibly and non-empirically justified moral beliefs only if the etiquette conception of defeat is true.

The etiquette conception of defeat would mark a radical departure from orthodox conceptions of defeat, which portray epistemic defeat as the result of failing to (non-accidentally) believe the truth. Though evolutionary defeat is in jeopardy, we will see that it is not dead yet and how it could be revived by defending the etiquette conception of defeat and by showing that moral beliefs violate epistemic etiquette. In comparison with previous discussion of evolutionary

challenges in metaethics, I focus much more on the conception of defeat and show where our conception of defeat is crucial for our understanding of the implications of evolutionary explanations of morality. Others may have put the challenge in terms of defeat but rarely considered that the current conception of defeat is inadequate to support the challenge that they sought to raise.

It must be clear that my thesis is concerned with establishing the conditional claim and not the claim that all moral beliefs are undercut. Consequently, there are two crucial limitations that must be pointed out. First, I do not investigate *whether* evolutionary explanations of morality show that our fundamental moral beliefs fail to qualify as knowledge and so I do not show whether or not moral beliefs are defeated, assuming that the etiquette conception of defeat is true. Second, I do not aim to vindicate the etiquette conception of defeat, though I suggest reasons in its favour in the final chapter of the thesis. Both projects would require a discussion of the norms of belief and the constituents of (moral) knowledge that is beyond the scope of this thesis. Before giving a more detailed overview of the thesis, I discuss the most crucial assumptions.

## Method and Presuppositions

My investigation is first and foremost a philosophical investigation. I touch on two related fields of philosophical inquiry: epistemology and ethics. I make noteworthy assumptions in both fields, which I clarify in this section.[30]

## Epistemology

Epistemology, narrowly understood, is the study of knowledge and justified belief (Steup 2017). My main concern will be epistemic justification. Epistemic justification is intimately connected to normative questions about what we ought to believe and what we are permitted to believe. I will assume that being justified in believing that p is a necessary condition for being permitted to believe that p and for it to be the case that you ought to believe that p. The kind of epistemic

[30] Further presuppositions are clarified in the introductions to individual chapters where they are relevant.

justification that I am interested in is *positive*, *adequate* epistemic justification.[31] *Positive* justification demands that the believer has positive grounds for holding the belief as opposed to remaining agnostic (Chisholm 1989). For a belief to be positively justified, the ground for holding the belief must be related to the belief's truth. All epistemologists agree that something more than mere absence of conflicting reason is necessary for a belief to be justified, so all will accept the requirement of being positively justified (cf. Sinnott-Armstrong 2006b; Swinburne 2001). *Adequate* justification demands that the overall balance of reason (or: all-things-considered reason) for holding a belief meets the threshold required for positive justification. For example, several newspaper reports might each contribute positively or negatively to your justification for believing that your favourite party will win the next election. Even with very little evidence in favour of a win by your favourite party, you might be *slightly* justified in believing that they will win, but not *adequately* justified, because your justification does not meet the required threshold. How high the threshold is depends, amongst other things, on the context and how much is at stake. I will not make any particular assumption about how high the threshold for adequate justification is in the case of moral beliefs.

I am investigating whether the justification of *all* moral judgements is threatened by evolutionary explanations of morality because I assume that the justification of moral beliefs depends in an important way on fundamental moral beliefs such as beliefs about the value of life and the fundamental equality of persons, the view that we ought to pursue the means to our ends, and pro tanto reasons to avoid harm, and so on. These beliefs might be fundamental because they are the 'bedrock' of a foundationalist view of justification or because they are part of any recognisably moral belief system on a coherentist view of justification. On the assumption that the support for all other moral beliefs depends on the justificatory support for the fundamental moral beliefs, all moral beliefs will lose their support once the fundamental moral beliefs are shown to be unjustified.

[31] For ease of expression, I will say that someone is *unjustified* in believing that p when I mean that he or she does not have adequate justification for believing that p.

I assume that at least some of these fundamental moral beliefs are *defeasibly justified*. That is, we can take them to be justified in the first place. Focusing on defeasible justification is one crucial way in which the evolutionary challenge differs from a radical sceptical challenge.[32] In general, in pursuit of an answer to my main research question, I demand that the constraints on answering the evolutionary defeat challenge must be different from the constraints on answering a radical sceptical challenge. Otherwise, the challenge would not retain philosophical interest *as a separate, novel challenge*. For example, a radical sceptic about perceptual knowledge will ask how we can trust our senses and demand that an answer to his sceptical challenge be given without relying on the outputs of our senses. To answer such a radical sceptical challenge to the sceptic's satisfaction, we would have to produce *independent* reasons, that is, reasons not based on perceptual input, that demonstrate that our perception is reliable. Philosophers reject or ignore the radical sceptical challenge, but nobody has succeeded in answering it on the sceptic's terms. Hence, for the evolutionary defeat challenge to offer a novel contribution to philosophy, it has to be different from the radical sceptical challenge (when applied to morality). One way in which this requirement might be violated is if the epistemic principle that could support the evolutionary defeat challenge gave rise to scepticism about the external world. Moreover, I assume that some method of belief formation, such as moral perception or intuitionism, allows us to arrive at prima facie justified beliefs (Audi 2013; Huemer 2005). If the justificatory status of objectivist moral beliefs is doubtful on independent grounds, there is little point in worrying about the justificatory effects of evolutionary explanations of morality.[33]

By asking how defeasibly justified beliefs lose their justification, I am asking a question about the more general phenomenon of so-called *epistemic defeat*. An epistemic defeater is a prima facie reason to withhold belief about some proposition that one would have otherwise been justified in believing.[34] Typically, two kinds of defeaters are distinguished: rebutters and undercutters. A rebutter is a reason to

[32] Further differences are discussed in the reader's guide to chapter 3.

[33] Another author who grants the default entitlement of moral beliefs is Locke (2014: 220).

[34] See Pollock and Cruz (1999), Pollock (1995), Pollock (1970), Chisholm (1964), and Hart (1948).

believe that the target belief is *false*. An undercutter is a reason to believe that the grounds on which one holds the belief fail to guarantee its truth. Losing your justification for believing that wedges are red when illuminated by a red light is an example of undercutting defeat. Learning that your judgements that you morally ought to help out in the local soup kitchen or that you ought not to betray your partner are influenced by evolutionary forces might be another undercutting defeater, insofar as the explanation of your belief might be independent of the truth of the belief.

In asking about 'evolutionary defeat', I will remain neutral between epistemic internalism and externalism regarding justified belief. The former holds that facts relevant to the epistemic justification of your beliefs must supervene on your mental states while the latter is the denial of epistemic internalism.[35] Externalists, like reliabilists, will be troubled by the possibility that moral beliefs lack a reliable connection to moral truths, while any internalist would also find it troubling to *learn* about the lack of an evidential connection between moral truths and moral facts. Thus, though the evolutionary 'reliability challenge' is stated in particular terms, such as 'reliability', that are congenial to an externalist account of epistemic justification, it can be 'internalised' to put pressure on internalists about epistemic justification too.

My final epistemological assumption is to focus my investigation on *non-empirically justified beliefs*. A belief is non-empirically justified only if it is justified and the truth of its content is not implied by the best explanation of any of our observations. Only some metaethical views rely on the possibility that moral beliefs can be non-empirically justified, and these are the views that have commonly been considered the primary target of various evolutionary challenges (Ruse 1995a; Street 2006; Vavova 2015; Warren 2017).

[35] Whether or not one needs to be *aware* of the facts relevant for epistemic justification or whether *access* to these facts is sufficient is a matter of some dispute amongst internalists regarding justification. See Alston (1989) for a classic statement of the awareness criterion and Conee and Feldman (2001) for a defence of what might be called 'access internalism' or 'mentalism'.

## Metaethics

Metaethics as a field of research is concerned with questions about the conditions needed to have moral knowledge and justified moral beliefs, the extent to which we have moral knowledge and justified moral beliefs, the metaphysical aspects of morality, and the semantics of moral language and thought. My main metaethical presupposition is to accept 'moral objectivism' for the sake of argument. I stipulate that moral objectivism is a view about the constituents of moral truth and our abilities to come to have justified moral beliefs and knowledge.

Moral objectivism has six theses, which should be taken as both specifying the model of morality presupposed in this thesis and specifying the *target* of the evolutionary defeat challenge. It makes sense to focus on objectivism because answering the challenge seems particularly problematic for this view, and thus it is a good test case for the 'survival of defeat'. Insofar as the metaethical view of choice encompasses the following six theses, it will be subject to the evolutionary defeat challenge.[36]

- TRUTH-APTNESS: Moral judgements can be assessed as being correct or incorrect.

TRUTH-APTNESS specifies that moral judgements are not mere expressions of sentiments but are like factual judgements. All moral realists should accept this thesis.[37] Anti-realists who argue that moral commitments do not express beliefs but that they can be assessed as beliefs nonetheless can also accept it (cf. Blackburn 1984; Gibbard 2003).[38] Insofar as anti-realists are committed to the correctness of moral judgements, they fall within the scope of the evolutionary defeat challenge.[39]

- NON-PLENITUDE: Of the many possible moral views, only a few are correct. Not all moral views are on a par.

[36] My exposition of the six theses follows, with exceptions, the exposition in Schechter (2018).

[37] Though see Kahane (2012) and Skorupski (1999).

[38] COGNITIVISM merely excludes early versions of non-cognitivism, like that of Hare (1963), Stevenson (1937), and Ayer (1971 [1936]).

[39] See Golub (2017) and Rovane (2013: 211ff) for an assessment.

NON-PLENITUDE excludes the view that there are multiple internally consistent moral systems, each with a set of moral truths that might be incompatible with the moral truths of other moral systems. For example, even if there are different groups with internally coherent but mutually incompatible sets of norms, only some of them can be correct. This excludes various forms of relativism from counting as objectivist in my stipulated sense and is responsible for the intuition that it would be problematic if evolution shows that we could have adopted a different moral belief system, had we taken a different evolutionary path (cf. Harman and Thomson 1996; Putnam 2013; Rovane 2013; Velleman 2013; Wong 2006).

- STANCE-INDEPENDENCE:[40] The truth-makers of fundamental moral principles are independent of ratification within any given or hypothetical perspective.

STANCE-INDEPENDENCE is a thesis about the fundamental moral principles that determine, together with non-moral facts, the remaining moral truths (Schechter 2018). Of course, moral facts that derive from fundamental moral principles often crucially involve a particular perspective or facts about our psychology, culture, or language. For example, the goodness of bravery is surely constituted partly in one's facing danger without fear. So, psychological facts play a role in determining correct use of moral judgements, but they do not constitute the fundamental moral principles. I do not mean to bestow particular weight to the term 'principle' as distinct from facts. I use the term 'principle' to signify that postulating fundamental moral principles need not incur metaphysical commitment (Cuneo 2014; Cuneo and Shafer-Landau 2014; Parfit 2011a; Scanlon 2014).[41] This way of framing moral objectivism makes it prima facie independent of ontological questions about what properties exist. This is why I call the position moral objectivism rather than moral realism. This is stipulative, of course, but there are good reasons for drawing a distinction.[42] One is that many self-

[40] I borrow this from Shafer-Landau (2003: 15).
[41] See Klenk (2015a) for a review.
[42] Though others often use both terms interchangeably; see Rachels (1998: 10).

proclaimed moral realists do not count as objectivists in my sense, because they reject STANCE-INDEPENDENCE but instead adopt a view according to which moral principles are true relative to the human life-form.[43] Moreover, STANCE-INDEPENDENCE makes it clear that the evolutionary issue that I am focusing on is not the causality of moral facts and their reducibility to non-moral facts, contrary to what some believe to be at the heart of the evolutionary issue (e.g. Artiga 2015; Crow 2016). Any view that holds that the truth values of the fundamental moral principles are stance-independent is subject to the evolutionary defeat challenge and qualifies as moral objectivism in the stipulated sense.[44] It excludes response-dependence views (e.g. Pettit 1991) and constructivism (e.g. Street 2010) from counting as objectivist in my sense. So much for moral objectivism's view on what makes moral judgements correct.

The final three theses of moral objectivism concern the epistemic status of moral judgements.

- JUSTIFICATION: A significant proportion of the moral judgements that we accept upon reflection and discussion are epistemically justified.

The JUSTIFICATION thesis is different to the epistemological presupposition, introduced above, that we are justified in taking at least some of our moral judgements to be prima facie justified. JUSTIFICATION specifies an epistemic commitment held by moral objectivism, and showing it to be violated (e.g. by showing that evolutionary considerations defeat the epistemic justification a significant proportion of our reflective moral judgements) would show that objectivism in my stipulated sense fails. The next requirement also specifies an epistemic commitment of objectivism:

- RELIABILITY: The moral judgements that we accept upon reflection and discussion are true significantly more often than chance would predict.[45]

[43] Cf. Railton (1986); Boyd (1988 [1995]); Brink (1989); Sayre-McCord (1986).

[44] The causal point might be inessential. Oddie (2009: ch. 5) and Wedgwood (2007: ch. 8) are thus also targeted. See Barkhausen (2016) and Setiya (2012: 114).

[45] A similar interpretation of the reliability thesis is adopted by Dogramaci (2017: 72); Enoch (2011b: 155); Field (1989: 230); Setiya (2012: 68).

The RELIABILITY thesis[46] entails that error theories are false, according to which moral judgements purport to be correct but *no* (positive) moral judgement that attributes a moral property to some act, person, or event is correct (Kalf 2013; Mackie 1977; Olson 2014; Streumer 2017). The reliability condition is vague in that it does not specify precisely what the threshold for reliability is. Specifying the threshold will not be crucial, however, because it suffices for objectivists to demonstrate that moral judgements are not likely to be false or correct only by chance.

- KNOWLEDGE: A significant proportion of the moral judgements that we believe upon reflection and discussion qualify as knowledge.

The KNOWLEDGE thesis demonstrates a commitment of moral objectivists to the possibility of moral knowledge (Enoch 2011b; Shafer-Landau 2003). This package of theses is accepted by a number of philosophers, most obviously by proponents of robust moral realism (Enoch 2011b; FitzPatrick 2008; Kramer 2009; Shafer-Landau 2003; Wielenberg 2014) and the moral intuitionists (Audi 2004; Huemer 2005; Ross 1930 [2007]; Sidgwick 1981 [1874]). However, there are others who fall under its scope too. For example, realists who claim that moral facts are causally efficacious (Oddie 2009; Wedgwood 2007) and realists who claim that their view does not commit them to the existence of moral facts per se (Cuneo 2007; Parfit 2011a; Scanlon 2014). Some moral naturalists also fit my characterisation of objectivism (Jackson 1998; Jackson and Pettit 1995). Thus, a significant number of important positions in metaethics is in the target area of the evolutionary defeat challenge.

I have clarified the commitments of moral objectivism, the target of the evolutionary defeat challenge. Two final questions need to be addressed, at least in outline. First, why accept moral objectivism in the first place? Second, what happens if we have to reject moral objectivism? Of course, a full answer to either question would be a metaethical treatise on its own. However, a brief sketch of an answer should help to lend support to my assumption that moral objectivism is a view worth discussing. The chief reason for taking moral objectivism seriously is

[46] The reliability thesis as a commitment of moral objectivism must not be confused with the reliability view about the way in which the evolutionary challenge works.

that it vindicates the intuition that some acts are morally wrong and some persons and perhaps events are morally bad in a counterfactually robust way. I assume that vindicating this intuition is a desideratum and that objectivism fulfils it if it is a tenable view.[47] Other metaethical views might vindicate the desideratum too, though I take no stance on this question.[48]

What if the evolutionary defeat challenge succeeds and objectivist moral judgements are undercut? As the individual theses of objectivism are formulated, there is no logical entailment between them. The evolutionary defeat challenge is an epistemic challenge and thus threatens JUSTIFICATION, RELIABILITY, and KNOWLEDGE. Rejecting these theses is consistent with maintaining, for example, the semantic theses STANCE-INDEPENDENCE or TRUTH-APTNESS. For proponents of a form of objectivism in my sense, however, this is a view that seems to have just about "zero appeal" (Shafer-Landau 2012: 1).[49] Thus, assuming that the evolutionary defeat challenge succeeds, one option is to construe differently the determinants of the truth values of moral propositions (cf. Street 2006). For example, one could drop STANCE-INDEPENDENCE and accept a form of constructivism according to which the truth-value of a moral judgement depends on the attitudes of the person making the judgement. In such a view were adequate (which I don't assess in this thesis), the rejection of moral objectivism would have purely metaethical implications. However, the challenge might cut deeper into moral discourse if we cannot easily change metaethical commitments. That is, for example, if a view like constructivism is not tenable. Construing differently the truth-makers of moral propositions might mean switching topics from morality to something else (Joyce 2006). Which option is open to us depends on whether objectivists are right about the contents of moral judgements. Having assumed that moral objectivism is true for the sake of argument, I will not discuss this question in what follows.

[47] Some argue that the metaethical claim that the truth assignment of moral propositions is counterfactually robust (in the sense specified above) is itself a moral claim and that there are good moral reasons to defend it; see Dworkin (1996); Kramer (2009).

[48] E.g. Maagt (2017).

[49] Error theorists might embrace it, of course, but they do not qualify as objectivists in my sense.

For this reason, when I write about the evolutionary defeat challenge for *moral beliefs or judgements*, I will assume that objectivists are right about the contents of moral beliefs and thus when I write of moral beliefs or moral judgments being defeated, I mean that moral beliefs or moral judgment *objectively construed* are being defeated. Of course, by showing that all moral beliefs are defeated (and that their justification could not be reinstated) one would show that the JUSTIFICATION and KNOWLEDGE theses of moral objectivism would be false and thus would defeat moral objectivism as a metaethical view. It all starts with showing that moral beliefs or judgements are defeated by evolutionary explanations of morality.

## Overview of the Thesis

In this section, I provide an overview of the thesis as a whole and point out relationships between the chapters. The main body of the thesis consists of seven chapters. Except for chapter 1, which is predominantly expository in nature, each chapter begins with a brief 'Reader's Guide' in which I relate the chapter's content to the main question of the thesis and clarify crucial assumptions. Each reader's guide comes in addition to an introductory section that specifies the particular research question(s) addressed in the chapter. There is some overlap between the introductory sections of some chapters because all chapters (except for chapter 1 and minus the respective reader's guides) were conceived of as self-standing articles. For this reason, each chapter can be read independently of the others (beginning with each chapter's introduction), whereas each chapter's reader's guide serves to connect the chapters to each other.

In **chapter 1**,[50] I establish the theoretical background of my thesis in greater detail. The most relevant background of my thesis can be thought of as involving four major parts: (1) empirical accounts of the evolution of morality, which gave rise to (2) substantive normative conclusions, which I want to set apart from my thesis, and (3) metaethical conclusions that met with (4) criticism of the arguments used to derive metaethical conclusions which ultimately threaten the survival of defeat. Providing a more detailed sketch of (1), setting aside (2), introducing the

[50] Parts of this chapter are based on Klenk (forthcoming).

most prominent accounts of (3), and explaining the motivation for (4) are the goals of chapter 1. I begin by providing a more detailed account of how morality evolved as a 'tool' for cooperation, which aims to make it plausible that both the capacity for moral thought and the content of some moral beliefs can be explained as influenced by evolutionary forces. I then distinguish my approach from the infamous 'evolutionary ethics' of the 20th century, which is important to avoid misconceptions about my research project.

Turning to metaethics, I relate my question to the most widely discussed 'evolutionary debunking arguments' in metaethics, which are those of Michael Ruse, Sharon Street, and Richard Joyce. Ruse argues that evolutionary explanations of morality imply that our moral beliefs are *false*, and I show that such arguments rely on controversial metaphysical questions that are beyond the scope of this thesis and for this reason I set them apart. I argue that both Street's and Joyce's evolutionary debunking arguments rely on an answer to my research question. Both need to answer how and why defeasibly and non-empirically justified beliefs ought to be given up in light of evolutionary explanations of morality without showing that they are false. Of course, Street, Joyce, and many commentators have proposed answers, but a review of the recent literature shows that extant proposals face serious obstacles. In particular, I address the views that evolutionary explanations show our moral beliefs to be in *error*, *coincidentally true*, or violating epistemic principles such as *safety* and *sensitivity*, as well as the aforementioned *reliability* and *disagreement* views. The result of my assessment of the literature is that the survival of evolutionary defeat is in jeopardy. This is why I address the nature of evolutionary defeat in greater detail in chapters 2, 3, and 4, and then turn to a discussion of the best currently available accounts of evolutionary defeat, the reliability view and the disagreement view, in chapters 5, 6, and 7.

In **chapter 2**,[51] I ask about the conditions needed for a belief to be undercut and argue that there have to be objective conditions for undercutting defeat but that currently available accounts are lacking. The reader's guide elaborates on the basic idea of defeat and points out the shortcomings of the best currently available

[51] This chapter is based on a paper that is currently under review.

accounts of defeat that are relevant for this thesis. I then raise two questions that need to be answered to make progress in the debate about undercutting defeat and, consequently, the evolutionary defeat challenge: are the conditions of defeat objectivist or subjectivist, and what are they, specifically? I argue for objectivist conditions of defeat. By 'objectivist conditions' I mean that the question of whether defeat is instantiated does not depend on whether the subject takes a belief to be defeated. This contrasts with subjectivist conditions, which are defended, amongst others, by Michael Bergmann (2006) and Alvin Plantinga (1993, 1994, 2002), who want to make room for defeat in cases where the subject (whose belief or beliefs are potentially defeated) firmly believes that he or she has a defeater. Though there may well be a concept of 'subjective defeat', I argue that this cannot be the concept of defeat relevant in assessments of the evolutionary defeat challenge. The upshot of the chapter is that defeat depends on good epistemic reasons, and this raises the question of what those reasons could be.

In **chapter 3**,[52] I investigate whether evolutionary explanations of morality make for a special kind of defeat and argue that evolutionary defeat depends on a priori considerations. As we will see, it has often been suggested that the evolutionary challenge is special in terms of the kind of challenge it is or in terms of the magnitude of its potential conclusion. I argue that evolutionary explanations of morality might point to an a priori problem with having non-empirically justified moral beliefs and I identify this problem to be the so-called Benacerraf-Field challenge, which relies on the same epistemic principle as the reliability view. It can be shown that the problem suggested by the Benacerraf-Field challenge is at the heart of the evolutionary defeat challenge. I argue that any defeater caused by the evolutionary challenge will depend on whether there are good a priori grounds to withhold belief in non-empirically justified beliefs.

This conclusion is relevant for two reasons. First, it clarifies the nature of the evolutionary challenge: it is really an a priori challenge in disguise. Second, it shows that to address the merits of the evolutionary challenge, the merits of the a

[52] This chapter is based on Klenk (2017c).

priori challenge have to be addressed and that moral objectivism cannot be defeated on empirical grounds alone.

In **chapter 4**,[53] I presuppose that there is *some* way in which the evolutionary defeat challenge works and then show that moral objectivists would not have the resources to resist defeat, which shows us something about the potential strength of the evolutionary defeat challenge. I show that defeat can be resisted by defeating it, but that a defeater-defeater would be unavailable if the evolutionary challenge succeeds because the evolutionary defeat challenge would defeat *all* moral beliefs. My main contribution in chapter 4 relates to a novel proposal about how to resist defeat. Objectivists might *deflect* defeating information from moral beliefs, as suggested by Moon (2017). Moon suggests that there are situations in which a believer might have information that will prevent him or her from having a defeater that he or she would otherwise have had. I argue that Moon's proposed way of deflecting defeat is unavailable for moral objectivists by considering carefully the examples of defeat deflection discussed by Moon and comparing them with the evolutionary defeat challenge. I conclude that *if* the evolutionary defeat challenge raises a defeater, it could neither be defeated nor deflected and the result would be devastating for the justification of our moral beliefs.

In **chapter 5**,[54] I turn to the disagreement view, which is the view that the evolutionary defeat challenge defeats our moral beliefs in virtue of the epistemological significance of counterfactual disagreement. In light of the problems for extant interpretations of the evolutionary challenge, some have suggested that, roughly, evolutionary explanations suggest to us that, given a different evolutionary past, we would have disagreed with our own present moral beliefs (Bogardus 2013, 2016; Mogensen 2016a). Assuming a conciliatory view about the epistemology of disagreement (cf. Elga 2007), these authors argue that the hypothetical disagreement implied by the 'evolutionary hypothesis' defeats our moral beliefs.

I argue that we should reject the disagreement view on the grounds that the counterfactual moral disagreement implied by the evolutionary challenge will not be between epistemic peers and thus will not be epistemically relevant enough to

[53] This chapter is based on Klenk (2017a).

[54] This chapter is based on a paper that is currently under review.

undercut our moral beliefs. My argument takes the form of a dilemma. I argue that either we consider nearby alternative evolutionary trajectories or far away ones. In the nearby trajectories, there will be agreement rather than disagreement about our moral beliefs, as implied by evolutionary explanations of morality. In the far away trajectories, there will be no disagreement such that it would undercut the justification of all moral beliefs. This is a stark claim and I defend it by leaning on the notion of epistemic peerhood. Peerhood, in one way or another, is a necessary condition for being conciliatory if faced with a disagreement. The concept of epistemic peerhood plays a central role in contemporary epistemological discussions of disagreement, but it is rarely analysed on its own. The two most frequent proposals are that a peer is someone who is in equal evidential possession (narrow conception) and that a peer is someone who is equally likely to get (the disputed) matters right (broad conception). I show that evolutionary explanations of morality imply that there is *no* disagreement with peers on a narrow conception of peerhood precisely because such accounts imply that we would have different evidence on alternative evolutionary paths. Turning to a broad conception of peerhood, I distinguish total from partial moral disagreement. In cases of total disagreement, we have no common ground about morality (not even concerning its function or methods) which would give us reason to think that our interlocutors are *not* equally likely to get things right. Partial disagreements might defeat some moral beliefs, but not all. Hence, I conclude that evolutionary explanations of morality do not defeat all moral beliefs even on a broad conception of peerhood.

The upshot for an answer to my main question is that the disagreement view fails to show how the evolutionary defeat challenge might succeed and so the prospects of the evolutionary defeat challenge depend on finding an alternative explanation. The chapter also makes a point about how the evolutionary challenge relates to the debate about peer disagreement.

In **chapter 6**,[55] I connect the discussion of the disagreement view with a topic relevant for the reliability view. So-called third-factor explanations are attempts to explain the reliability of moral beliefs by postulating a substantive moral claim, such as 'survival is good', which would then explain both why humans tend to form

[55] This chapter is based on a paper that is currently under review.

such beliefs and why these beliefs are true. A third-factor account along these lines could be the basis of explaining our moral reliability and thus provide a way to resist the claim that evolutionary explanations show that there is *no* way to explain the reliability of our moral beliefs. This thesis is relevant both for the reliability view and the disagreement view about evolutionary defeat. The epistemic significance of disagreement might yet turn out to be relevant for the evolutionary defeat challenge (in whatever sense it might work) if it destabilises third-factor accounts, as argued by Tersman (2017). Put differently, concerns about moral disagreement might show that what many take to be a promising answer to the evolutionary defeat challenge (on the reliability view) fails, and because it is the epistemic significance of disagreement that seals the fate of moral objectivism, the disagreement view might turn out to get it right after all.

I argue that disagreement cannot play the role envisaged by Tersman on the terms of the evolutionary defeat challenge. Roughly, Tersman envisions that the substantive moral claims required for third-factor accounts might be subject to defeating disagreement and thus unavailable to set up a third-factor account. I argue that the assumptions granted in the context of the reliability challenge entail that there cannot be defeating disagreement. Once we allow that there is a set of moral beliefs that can legitimately be used to answer the reliability challenge, the moral beliefs in that set should be immune from defeating disagreement.

The upshot of this chapter for my main question is that disagreement also cannot indirectly support the disagreement view. The broader implications are that third-factor accounts, which are an interesting topic independently of the evolutionary defeat challenge, are not threatened by the epistemic significance of disagreement.

In **chapter 7**,[56] I address the major challenge for the reliability view: general considerations about defeat suggest that learning new information E can only undercut a belief if E implies that the belief fails to be epistemically safe or sensitive. Very roughly, a belief is sensitive if it would not be held if it were false, and a belief is safe if it is mostly true when it is held (properly speaking, safety and

[56] This chapter is based on a paper that is currently under review.

sensitivity apply to the beliefs of thinkers, as I make clear in the chapter). The challenge for the reliability view is that evolutionary explanations of morality imply that at least some moral beliefs are safe, and a standard semantic of counterfactual conditionals such as 'sensitivity' implies that those moral beliefs are sensitive. Hence, at least some moral beliefs cannot be undercut. This challenge has been defended most prominently by Justin Clarke-Doane (2015, 2016a, 2017c), and many others have taken up his claim, suggesting that the reliability view falters in light of this problem (e.g. Baras 2017; Barkhausen 2016; Dogramaci 2017; Warren 2017). An important consequence of this view is that moral objectivism would be immune from undercutting defeat as well as several related but controversial views such as Platonism in the philosophy of mathematics or modal realism in metaphysics.

I argue that moral objectivism is not immune from undercutting defeat by presenting a counterargument to the epistemic principle that supports the view of Clarke-Doane and his followers. First, in the reader's guide to this chapter, I clarify some crucial assumptions about the nature of Clarke-Doane's challenge (including the assumption that some moral beliefs are metaphysically necessary and the standard method of assigning truth values to conditionals) and then turn to my rebuttal of Clarke-Doane's challenge. I show that Clarke-Doane is right that the reliability view does not succeed: at least some of our moral beliefs can be shown to be reliable, where a belief is taken to be reliable insofar as it is sensitive and safe. However, I argue that this does not immunise moral objectivism from the threat of undercutting defeat. My strategy involves two steps. First, I show that a belief that is all but guaranteed to be true (thus fulfilling Clarke-Doane's criteria) can nonetheless fail to be knowledge. Second, I show that epistemic value is not exhausted by a belief being justified and true and that learning that a belief fails to qualify as knowledge might thus give us reason to give up the belief. Hence, Clarke-Doane's defence of moral objectivism is based on a mistaken assumption about the nature of undercutting defeat.

I then show that my counterargument commits us to the etiquette conception of defeat, which is the view that new information can undercut a belief by showing that the belief does not qualify as knowledge even though it gives us no reason to doubt that the belief is justified and true. Even though we have, by this point,

rejected the disagreement view and the reliability view of the evolutionary defeat challenge, evolutionary defeat is not dead yet. I clarify my choice of terminology for the etiquette conception, provide reasons in its favour, and contrast it with the orthodox conception of defeat, which is, as chapter 2 has shown, at the heart of virtually all current discussions of the evolutionary defeat challenge (insofar as the challenge is put in terms of defeat). My argument in chapter 7 shows that the reliance on the orthodox conception is a mistake. Though I do not provide a full defence of the etiquette conception, I point out some reasons in its favour.

By outlining the conditions needed for the evolutionary defeat challenge to succeed, and by rejecting the most promising alternatives, I will thus have answered the main question of my thesis: the evolutionary defeat challenge succeeds if the etiquette conception of defeat is true (as I suggest) and if moral beliefs violate epistemic etiquette.

These seven chapters are rounded off by a **conclusion**, which summarises my argument and sets out the key questions which arise in its wake. In particular, I outline how we should, based on my assessment of the conditions for the survival of defeat, focus on the norms for moral belief and the question of whether there is a virtue requirement on knowledge to assess whether evolutionary defeat is not only possible, but whether it will also be found in the wild.

# 1 Evolution and Morality: Science, Pitfalls, and Metaethics

## 1.1 Introduction[1]

In this chapter, I provide the empirical, historical, and systematic background against which we can begin to assess the main question of my thesis. My main aim in this chapter is to provide the reader with the background information that serves as the starting point of my thesis. Readers familiar with the debate on evolutionary explanations of morality and the debate about their (meta)ethical implications may want to skip sections 1.2 to 1.4. In section 1.2, I will sketch out the evolutionary origins of morality and argue that the capacity for normative thinking was selected by evolutionary forces, plausibly on the group level as well as on the individual level, to make our ancestors better cooperators. As we will see in chapter 3, the specifics of the empirical account turn out to be inessential for the evolutionary defeat challenge, but it helps to illustrate one such account to make plausible how morality could have evolved. I also discuss what the evolution of a normative faculty implies for the evolution of a moral faculty. In section 1.3, I distinguish my thesis from *prescriptive* evolutionary ethics, whose practitioners try to derive first-order normative implications from evolutionary explanations of morality and which is mainly responsible for the "dreadful reputation" (Ruse 1995b: 196) of evolutionary ethics today.

Turning to the metaethical implications of evolutionary explanations of morality in section 1.4, I introduce the most prominent metaethical evolutionary debunking arguments, which are those of Michael Ruse, Sharon Street, and Richard Joyce, and show how my research question relates to these arguments.

In sections 1.5 and 1.6, I discuss the extant criticism of the evolutionary defeat challenge. The overview of the extant literature that I will provide indicates

[1] Sections 1.2 to 1.4 are based on content from Klenk (forthcoming).

that evolutionary explanations of morality do not undercut defeasibly justified moral beliefs. Apart from providing the empirical and historical background for my thesis, the upshot of this chapter is thus that the survival of defeat is in jeopardy.

## 1.2 Evolutionary Explanations of Morality

A descriptive account of human moral beliefs from an evolutionary perspective takes into account the findings of human sociobiology, human behavioural ecology, evolutionary psychology, and gene-culture co-evolution to answer two primary questions about ethics (cf. Laland and Brown 2011). Evolutionary challenges for morality start from the claim that two questions, which we can call CAPACITY and CONTENT, can be answered in evolutionary terms:[2]

- CAPACITY: Why do people grasp norms, make normative judgements, and behave according to norms, such as norms about the permissibility of killing?
- CONTENT: Why do people have the norms they do about, for example, when it is permissible to kill?

The remainder of this section aims to make sufficiently concrete and plausible how CAPACITY and CONTENT can be answered in evolutionary terms before we delve into the philosophical discussion. We will see that an ability to grasp, understand, and adhere to *norms* in general, to have a normative capacity, is plausibly seen as an evolutionary adaptation (cf. Cummins 1996; Machery and Mallon 2010). Norms are understood here as shared beliefs and expectations of the members of a group that effect observable behaviour in the group (Bicchieri 1993: 232, 2006: 11). Although an account of the evolution of norms does not show that *moral* norms are also adaptations, the story about norms is enough to support an evolutionary answer to CAPACITY and CONTENT, which is enough to start the evolutionary defeat challenge.

To begin with, we should note that "all phenotypes [including some cultural artefacts] are to some extent the products of the process of evolution by natural selection" (Brandon 1990: 41). After all, neither we nor our norms just popped into

[2] See Schloss (2014: 85ff) for more detailed discussion and further relevant distinctions of the explanandum in evolutionary explanations of morality.

existence. But to show that some *trait*, such as a capacity to think in normative terms or a tendency to value some things rather than others, evolved is not enough to show that it is an *adaptation* or a product of evolution in a stricter sense. A trait in the biological sense is a characteristic or attribute of an organism that is expressed by genes and/or influenced by the environment. Traits include the physical attributes of an organism, such as hair colour, but also behavioural characteristics, such as nesting patterns in birds. Some trait T is an *adaptation* for doing X if and only if there was selection for T because having T promoted doing X (Sober 1984b: 208; West-Eberhard 1999).[3] For example, pale skin in humans is an adaptation that aims to produce more vitamin D in dim climates. The paleness of human bones, in contrast, is a by-product of natural selection that produces sturdier bones, increasing their calcium content, which in turn accounts for their whiteness. In a loose sense, this is an evolutionary explanation of bone colour (there was selection *of* this trait) even though it does not show that whiteness was selected *for* (Sober 1984b).[4] Thus, we can see the human capacity to grasp norms and the contents of specific norms as an adaptation if they are heritable, innate dispositions, though such an explanation need not exclude the influence of the environment, for example through culture or upbringing.

The prominent philosophical discussions of evolutionary challenges in moral philosophy of Ruse (1986) and Joyce (2001, 2006) see a disposition to engage in normative thought as an innate, heritable trait favoured by individual-level selection (to wit, where evolutionary forces are thought to work exclusively on individual organisms).[5] The evolution of cooperation that is mutualistic (i.e. it incurs a net-benefit for all involved parties) or that involves only close relatives is certainly easily explained by a model of individual-level selection (where prosocial

[3] It's important to distinguish traits that are adaptations from traits that are adaptive. The concept 'adaptation' looks to the past. To say that a trait is an adaptation is to make a claim about its history. 'Adaptive' looks to the future. To say that a trait is adaptive is to say that it promotes reproductive success and survival (Sober and Wilson 2011).

[4] Apart from being an adaptation or the by-product of an adaptation, a trait might also be an evolutionary accident. I ignore this possibility here because it is not well supported by the evidence; see Williams (1988) for a discussion of norms as evolutionary accidents.

[5] So does Street (2006); see also Hauser (2006: 53) and Dwyer (2006).

dispositions, such as dispositions to cooperate, are seen as an adequate proxy of morality).[6] Amongst close kin, cooperation could have evolved by natural selection because the benefits of cooperative actions are conferred on the genetic relatives of the cooperator, thereby helping to proliferate alleles associated with the cooperative behaviour.[7] Amongst repeatedly reciprocal partnerships, cooperation could also have evolved because one individual's costly contribution to the benefit of another individual is reliably reciprocated at a future date, thereby making cooperation mutualistic.[8] These explanations of the evolution of cooperation (which, to repeat, is often taken as a proxy of the evolution of morality), which might indeed be called 'enlightened self-interest', are popular amongst biologists and economists and explain many forms of human cooperation, particularly those that take place in families, frequently repeated dyadic relations, or minimal group interactions.

However, there are two problems with a restricted focus on individual-level selection. First, its commitment to moral nativism, the claim that certain moral norms are innate, weakens the evolutionary case for answering the CONTENT question because it is implausible that norms with specific content were selected *for* on the individual level. Of course, some fundamental evaluative tendencies in humans might be innate (such as an aversion to snake-like objects), but more readily recognisable moral norms, such as norms concerning harm avoidance, are less easily explained on the individual level. For example, a moral rule commonly cited to be universal in the sense of being endorsed by individuals of all cultures is 'don't harm innocent people of the in-group'. However, this rule is certainly not universal, and the universality of a norm would be an important indicator that it is indeed innate, which would lend some support (though not conclusive support)

[6] Of course, much depends on whether operationalisations of 'moral judgments' used in evolutionary explanations of morality, of which prosocial behaviour is the most common one, are indeed adequate; see Pölzler (2017). The thicker our account of morality, the less likely it is that we will find that morality evolved. I am aware of this difficulty, but because my aim here is to present a common view about the evolution of morality, I adopt the common operationalisation in terms of prosocial, cooperative behaviour.

[7] This process is known as kin selection; see Hamilton (1963).

[8] This process is known as reciprocal altruism; see Trivers (1971). A related framework, called indirect reciprocity, includes reputation and is able to account for the evolution of cooperative behaviour in more scenarios than simple reciprocal altruism; see Alexander (1987) and DeScioli and Kurzban (2013).

to it being an adaptation. Many societies harmed, for example, women of the in-group, and ritualistic practices such as initiation rituals often involved harm for in-group members (e.g. Abarbanell and Hauser 2010). Of course, counterexamples on the ethnographic record can be avoided by, for example, interpreting 'in-group' widely as 'those that one ought not harm', but then the allegedly universal moral rule becomes a triviality and thus uninteresting as a test case of whether some moral norms are universal (cf. Prinz 2009). Some are therefore suspicious of the claim that there are substantive, non-trivial universal moral norms. Second, human cooperation takes place in groups far larger than the immediate family and in both field and laboratory experiments that are unlikely to be repeated and where no reputational gains are to be expected (Bowles and Gintis 2011). The processes of individual-level selection have difficulties explaining such behaviour, mainly because they cannot readily explain how humans became steady cooperators, not opportunistic cheaters (Boehm 2001, 2012). Luckily, the processes of individual-level selection, such as kin selection and reciprocal altruism, are only two of the "four paths to cooperation" (Dugatkin 2000). Thus, an adequate evolutionary explanation of norms departs from the current philosophical focus on individual-level selection and takes into account a more encompassing view of the levels of selection.[9]

Irrespective of the debate about the unit or the level of selection that accounted for the evolution of morality, the leading hypothesis about the evolution of norms is what we might call the 'cooperation hypothesis', according to which the origins of the human normative capacity lie in the development of cooperation (Axelrod 2006; Axelrod and Hamilton 1981; Curry 2016; Kitcher 2011).[10] One point in favour of the cooperation hypothesis is that cooperation is common in many species, but *Homo sapiens* is exceptional because in humans cooperation extends beyond close genealogical kin to include even total strangers, and occurs on a much larger scale than in other species except for the social insects (Bowles and Gintis 2011; Dugatkin 2000). Another point is that a normative capacity affords a solution to many problems that early hominids had to solve to efficiently cooperate (Kitcher

[9] Cf. Richerson and Boyd (2006); Sober and Wilson (1998).
[10] See Smyth (2017) for criticism of this view.

2011: ch. 1; Tersman 2006: 124ff). The cooperation hypothesis gains support on both the proximate and the ultimate level of explanation.

A proximate explanation explains how a trait works in the current (socio-cultural) environment (Mayr 1977).[11] The proximate explanation of cooperation beyond kin and reciprocal partners is that humans exhibit strong social preferences, which include a concern for the well-being of others, as well as a desire to uphold ethical norms. Countless experiments and field studies have shown that the "critical role of social preferences in sustaining altruistic cooperation is ubiquitous" (Bowles and Gintis 2011: 3; compare Fehr and Fischbacher 2003; Henrich et al. 2001). Selfishness is an important motive, but other motives, such as social ones, are no less important (Forgas et al. 2016).[12] In experimental studies, many subjects were found to be fair and generous towards those with similar inclinations and nasty towards those who violate prosocial rules (Bowles 2016).[13] Developmental studies offer further support for the view that humans are built to cooperate (Tomasello 2016).[14] At least concerning their behaviour towards their in-group, humans are not the nasty brutes of Hobbesian repute, nor purely selfish maximisers of their own interest. This raises a question about the ultimate explanation of these tendencies.

An ultimate explanation explains how a trait evolved and thus whether a trait can be seen as an adaptation, a by-product, or noise. The ultimate explanation of how humans became 'cooperation machines' required for the evolutionary defeat challenge must show how being a good cooperator can be seen as an adaptation.[15] As suggested above, such an explanation will be most plausible if it is based on a view of the evolution of cooperation that takes into account the role of group selection and gene-culture co-evolution.[16] According to this view, natural selection also exerts pressure on groups, not just on individuals. Group selectionists do not

[11] Hume and the British sentimentalists were well aware of the importance of human social preferences in shaping moral behaviour; see Schneewind (1998) and Darwall (1995).

[12] See Klenk (2016e) for a review.

[13] See Klenk (2017e) for a review.

[14] See Klenk (2016a) for a review.

[15] Natural selection is just one of the processes that can produce evolutionary change; see Sober (1994: 95ff) for a relevant discussion in the context of descriptive evolutionary ethics.

[16] For a defence of the possibility of group selection and its relevance in human evolutionary history, see Sober and Wilson (1998).

deny the role of selection on the level of the individual, but they emphasise that group selection is a crucial ingredient in the evolution of morality (Bowles and Gintis 2011). Whether this account is true will depend on the environment that our ancestors inhabited. A plausible hypothesis comes from Michael Tomasello. As competition for resources between groups of hunter-gatherers intensified around 150,000 years ago, group cohesion and cooperation amongst group members became an essential factor (Tomasello 2016). Groups with more cooperative individuals would be better able to compete with other groups for resources, and social preferences would have been a way to increase cooperation in this struggle. Humans later created novel social and physical environments in which cooperation yielded similar, or even greater, benefits, and culture probably played a crucial role in this process (Henrich 2016; Richerson and Boyd 2006). Cooperation must have been beneficial to the members of the groups that practised it, and our early ancestors were able to erect institutions that minimised the disadvantages to cooperative members (thus shielding them from exploitation by selfish fellows) and at the same time heightened the group-level advantages associated with the higher levels of cooperation (Bowles and Gintis 2011: 4).

Therefore, though the issue is far from settled, there are plausible explanations of how the human normative capacity was selected for. Provided that we take into account the influence of culture on genetic evolution, it also seems plausible that the content of some normative beliefs, where at least some rudimentary ones are recognisably moral, were influenced by evolutionary forces.

The chief reason that it is unclear whether a 'moral capacity' or the content of moral norms are adaptations, by-products, or noise is the difficulty of drawing a neat conceptual distinction between moral norms and non-moral norms (Pölzler 2017; Sinnott-Armstrong and Wheatley 2013; Sripada and Stich 2006).[17] For example, pointing the bottom of your feet towards another person is considered immoral in some Middle Eastern cultures, though perhaps impolite, but not immoral, in most European countries. Some who are impressed by the difficulty of demarcating moral from non-moral norms, such as prudential, conventional, or aesthetic norms, dismiss the idea that *morality* as a distinctive type of normative

[17] Cf. Machery and Mallon (2010).

cognition is an adaptation (Machery and Mallon 2010: 36). This point is well taken. For a proper evolutionary explanation of morality, we should have a clear idea about what the trait is that we want to explain. The problem is that one person's moral norm might be another's conventional (or prudential, or aesthetic, etc.) norm. As long as we are unclear about whether trait M counts as a moral norm, as opposed to some other kind of norm, it will be difficult to fully convincingly establish that morality is an adaptation. However, insofar as moral norms are the output of an evolved normative capacity, these discussion points can be set aside for present purposes. It suffices to show that the human normative capacity plausibly evolved, while the content of those norms, moral in some cases, is the product of culture (Henrich 2016).

We have seen a sketch of how modern descriptive evolutionary ethics might answer CAPACITY and CONTENT. The empirical basis of the evolutionary defeat challenge is now sufficiently established for the purposes of this thesis. Let's turn to its purported implications for moral theory next.

## 1.3 Normative Implications?

The historical perspective of this section provides a context within which we can begin to discuss contemporary discussions about the (meta)ethical relevance of evolutionary hypotheses.[18] As we have seen above, the poor repute of evolutionary ethics stems from the attempt to derive *moral justifications* from evolutionary *explanations* to answer questions such as

- JUSTIFICATION QUESTION: When is killing morally permissible?

The justification question is a question of ethics. Ethics involves systematising, defending, and recommending concepts, rules, or norms of right and wrong conduct. To fix ideas, I will make use of a paradigmatic understanding of ethics, which introduces the content of ethics by way of examples rather than by explicit definition (Joyce 2006: 70; Kalf 2013).[19] Failing to keep a promise is a paradigmatic moral act of negative valence, caring for our offspring is a paradigmatic act of positive moral valence, that it is horrible that almost 11% of

[18] For a more detailed account see Laland and Brown (2011: ch. 2).
[19] Like Enoch (2011b), I use the terms morality and ethics interchangeably.

the world's population was starving in 2016 is a moral fact, and that one ought to be as selfless as Mother Teresa or that one ought not to steal are moral judgements. Saying much more about what ethics or morality is at this point would risk restricting the scope of my investigation to particular, substantive theories about what ethics is. In line with common practice, I will avoid this and trust that a paradigmatic understanding of ethics and morality will suffice (cf. Sauer 2017: 16–8).[20]

Some have thought that evolutionary explanations of morality can have direct import in ethics by playing a vindicating role for morality in general or for particular moral theories. These ideas took off immediately after Darwin published his book *The Descent of Man* in 1871. Darwin's book came at a time when "old religious verities were decaying and crumbling" in the West, and people sought a new basis for morality (Ruse and Richards 2017: 1). A theory that explained the origins of life and the battle for survival was considered to be a proper ground for morality by many. An early proponent of prescriptive evolutionary ethics was Herbert Spencer, who is often seen as the paradigmatic defender of prescriptive evolutionary ethics (Spencer 1879, 1893).[21] Spencer is often mentioned in connection with the claim that 'more evolved' means 'morally better' and the idea that this supports an ethics of progress.[22]

The philosophical problems with prescriptive evolutionary ethics have been pointed out since its inception, yet the horrific real-world consequences of some prescriptive evolutionary thinking and the fact that attempts at reinstating the project are repeatedly made (e.g. Richards 1986, 2017) merit a brief discussion at this point.

Spencer's followers thought that evolutionary theory could show that particular moral principles or rules are justified. In other words, instead of answering the CONTENT or CAPACITY question, they sought to tackle the

[20] Though see Kumar (2015), for an attempt to capture moral judgement as a natural kind as well as Jackson (1998) for a related proposal of how to go about fixating the extension of the concept 'morality'. Of course, an adequate understanding of this question is important, as pointed out in footnote 6 of this chapter, in trying to give a descriptive account of the evolution of morality.

[21] As Lillehammer (2010) points out, Spencer's views are more nuanced than usually portrayed.

[22] Lillehammer (2016) provides a good discussion of reactions to Spencer's views.

JUSTIFICATION question. For example, recall the central role played by inter-individual fitness differences in processes of natural selection. There is an evolutionary explanation that shows why 'fit' phenotypes survive and reproduce more or better than unfit phenotypes. Many have taken this *explanation* to argue that *the strong are morally justified to dominate the weak*. Comprehensive eugenics programmes in many countries were just one of the acts allegedly sanctified by this norm (Paul 2006), as were nationalistic ideologies (Pichot 2009; Weikart 2006).[23]

However, such thinking betrays an argumentative fallacy known as an 'appeal to nature'. For example, arguing that the use of contraceptives is morally wrong because they prevent the 'natural' outcome of sexual intercourse or that men should not do household chores because it is not in their nature are appeals to nature. Appeals to nature can be fallacious in mistakenly identifying something as 'natural' – the example of men not doing household chores illustrates this.[24] Appeals to nature are fallacious in a logical sense too, because they make an illegitimate step from is to ought. This is, of course, one of the "shibboleths of contemporary philosophy" (Dennett 1995: 467), David Hume's famous claim that you cannot derive an 'ought' from an 'is':

> In every system of morality, which I have hitherto met with, I have always remark'd, that the author proceeds for some time in the ordinary way of reasoning, and establishes the being of a God or makes observations concerning human affairs; when of a sudden I am surpriz'd to find, that instead of the usual copulations of propositions, is, and is not, I meet with no proposition that is not connected with an ought, or an ought not. This change is imperceptible; but is, however, of the last consequence. (Hume 1738 [2007]: 302)

Hume deplores the unexplained, imperceptible change from 'is' to 'ought' in a moral argument, and such imperceptible and illegitimate steps were plentiful in

[23] See Paul (2006) for a comprehensive study of the relation of Darwinism and Eugenics. Flew (1970) and Farber (1994) provide timely overviews of the debate about prescriptive evolutionary ethics.

[24] What 'natural' means is unclear. The story of John D. Rockefeller and Andrew Carnegie is telling in this respect. The former grew rich from his company Standard Oil and advanced the view that the 'natural' way is for the powerful to dominate the less powerful, from which he sought to derive justification for his ruthless business practices. The latter founded US Steel and grew rich too, but thought that it would be most 'natural' to help fellow humans to improve themselves, which he took to justify his philanthropic engagement; see Russett (1976).

some of the arguments of early proponents of evolutionary ethics. The problem was that Spencer and his followers sought to derive moral principles directly from evolutionary insights. Many added normative principles to their evolutionary arguments, of course, such as that the *fittest ought to survive,* but they thought they derived crucial support for these ethical principles from Darwinism.[25]

Historically, the demise of early normative evolutionary ethics is most closely associated with the British philosopher G.E. Moore, who effectively put an end to the heydays of Spencer's project in 1903.[26] Moore charged normative evolutionary ethics with committing a "naturalistic fallacy".[27] Moore was interested in the definition of 'good' and argued that since 'good' is a simple property, it cannot be defined by outlining its more basic properties. Thus, identifying 'good' with 'higher evolved' as Spencer did was to commit the naturalistic fallacy. The impact of Moore's challenge has been devastating for normative evolutionary ethics. More than ninety years after Moore, Michael Ruse observes that "it has been enough for the student to murmur the magical phrase 'naturalistic fallacy' and then he or she can move on to the next question, confident of having gained full marks thus far on the exam" (Ruse 1995b: 220).[28]

The dismissal of *prescriptive* evolutionary ethics by philosophers since Moore and up to Ruse in the late 1990s not only created one of the most controversial academic disputes when E.O. Wilson (1975 [2002]) tried to revive evolutionary ethics in the mid-1980s but it has led to undue neglect of *metaethical* evolutionary ethics. Evolutionary metaethics is concerned with the impact of descriptive

[25] Logically, it is straightforward to derive an ought from an 'is' even though the word 'ought' does not appear in any premise of the argument by way of disjunctive addition or by inferring an ought from a contradiction Prior (1960). To use Prior's example: Paris is the capital of France and it is not the capital of France. So, you ought not steal bananas. Whether there is an interesting sense in which this can be done, however, is another question (cf. Pidgen 1989).

[26] Cf. Allhoff (2003).

[27] Some philosophers illegitimately equate Moore's naturalistic fallacy with the is-ought challenge (e.g. Wright 1995: 330). Moore's challenge, however, concerns the analysability of moral properties, not the logical relation of is and ought. Frankena (1939) has a penchant discussion of the merits of Moore's argument in support of the naturalistic fallacy.

[28] Prescriptive evolutionary ethics is possible in quite a different sense. We can learn more about the non-moral facts relevant for moral behaviour and judgement. This is at least one sense in which Greene and others claim that evolutionary considerations have normative implications Greene (2010).

evolutionary ethics on metaethical questions about the nature of moral facts, moral justification, and moral knowledge.

Though metaethics is concerned with normative questions too (i.e. what ought we, epistemically, believe about morality?), it does not suffer from the shortcomings of prescriptive evolutionary ethics, as long as we are very clear about the epistemic, normative principles that allow us to take the hypothesis about the evolution of norms to derive a metaethical conclusion. As we have seen in the main introduction, finding out which epistemic, normative principles these are is the main task of my thesis. With the empirical and historical background clarified, and with prescriptive evolutionary ethics set apart, we can now turn to the question of how evolution might influence how we think about the justificatory status of our moral beliefs.

## 1.4 Metaethical Implications

The metaethical debate about the evolutionary challenge is led primarily in reference to the works of three philosophers: Michael Ruse (1995a, 1998), Sharon Street (2006), and Richard Joyce (2001, 2006). All three claim that evolutionary explanations of morality have important metaethical consequences that are *debunking* in some aspect. Such *evolutionary debunking arguments* (EDAs) come in different kinds, emphasising different features of the evolutionary process, aiming at different conclusions, and relying on a variety of epistemic principles. Some authors have sought to reduce all such arguments to a single canonical form, but I think it best to leave their heterogeneity more open to view.[29]

I will roughly follow Joyce's (2013b) stipulative distinction between the different existing evolutionary challenges, according to which the EDAs of Ruse, Street, and Joyce aim to establish metaphysical, metatheoretical, and epistemological conclusions, respectively. Adding to Joyce's distinction, I distinguish EDAs based on whether they aim to establish (a) that all moral

[29] E.g. Kahane (2011); Mason (2010). For example, the evolutionary debunking arguments that I discuss below have been thought to affect the possibility of moral knowledge (e.g. Bogardus 2016; Brosnan 2011; Tropman 2014; Wielenberg 2010) or moral universalism (Talbott 2015), but such claims are hard to assess because proponents of evolutionary debunking arguments do not even mention these concepts to be their target in the first place.

judgements are false (what I call *evolutionary rebutting*), (b) that all moral judgments are unjustified, conditional on a particular metatheoretical view (*conditional evolutionary undercutting*), and (c) that all moral judgments are unjustified, unconditional on a particular metatheoretical view (*unconditional evolutionary undercutting*). This distinction is stipulative and primarily an attempt to impose order. Proponents of all three kinds of debunking arguments can, at times, be read as offering a different kind of debunking argument.[30] In what follows, I shall briefly recapitulate their arguments. Then I will show, in section 1.5, how addressing my main question plays a crucial role in assessing the strength of Joyce's and Street's EDAs and why I set aside arguments akin to Ruse's EDA.

### 1.4.1 Evolutionary Rebutting (Ruse)

Metaphysical debunking arguments aim at showing that all moral beliefs are false insofar as they presuppose that morality is objective. They aim at rebutting certain moral beliefs based on evolutionary explanations of morality. Take, for example, the metaphysical debunking argument of Michael Ruse.[31]

According to Ruse, the metaethical implication of evolutionary explanations of morality is that "morality is a collective illusion foisted upon us by our genes" (Ruse 1998: 253). Ruse does not mean to deny that we have moral beliefs and act in characteristically moral ways; rather, the illusion he refers to is that morality is *objective*:

> What is really important to the evolutionist's case is the claim that ethics is illusory *inasmuch as it persuades us that it has an objective reference.* (Ruse 1998: 235 emphasis added)

By 'objective reference' Ruse means the following:

> [Moral objectivists claim that], in some sense, the ultimate basis of ethics is objective. By this is meant that moral norms exist independently of humans

[30] For example, both Joyce (2006) and Street (2006), to whom I attribute 'justification-debunking' and 'theory-debunking' arguments, respectively, invoke Ockham's razor, suggesting that they also draw metaphysical conclusions (Joyce 2006: 209). Similarly, Ruse, whom I take to offer a paradigmatic metaphysical debunking argument, may also be read as offering theory-debunking, insofar as he is not always explicit that morality demands 'objectivity'.

[31] Olson (2014), and recently a paper by Sterelny and Fraser (2017), defends a similar argument.

> – at least, independent of human emotions – in some non-physical way. It is usually also claimed that we humans intuit or otherwise rationally grasp morality… In analogy with mathematics, the 'objectivists' tend to think of such norms as fixed and eternal. (Ruse 1998: 214)

According to Ruse, *thinking* that morality is objective is an integral part of the function of morality: "morality simply does not work (from a biological perspective) unless we believe that it is objective" (Ruse 1998: 253). However, he claims that "in light of what we know of evolutionary processes, the objective foundation [of morality] has to be judged redundant" (Ruse 1986: 242). Ruse argues:

> [The objectivist must agree] that his/her ultimate principles are (given Darwinism) redundant. You would believe what you do about right and wrong, irrespective of whether or not a "true" right and wrong existed! The Darwinian claims that his/her theory gives an entire analysis of our moral sentiments. Nothing more is needed. Given two worlds, identical except that one has an objective morality and the other does not, the humans therein would think and act in exactly the same ways. (Ruse 1998: 254)

This shows, argues Ruse, that "morality has no objective reference" and that "morality is subjective" instead (Ruse 1995a: 236).[32]

If Ruse is right, then all moral beliefs that purport to be objectively true are false. It is not entirely clear whether Ruse is committed to the semantic thesis, usually advanced by error theorists, that moral beliefs do purport to be objectively true.[33] Ruse's account is underdetermined in this respect. If he were correct in that moral judgements purport to be objectively true and that there are no objective moral facts, then all moral judgements would be false. Clearly, learning that a belief is false is a reason to give up that belief. In that sense, Ruse's metaphysical debunking argument is an instance of evolutionary rebutting and would, if

[32] Ruse also writes that "ethics is subjective, but its meaning is objective" (Ruse 2006: 22) and that "the meaning of morality is that it is objective" (Ruse 2009: 507). In the cited passages, Ruse appears to claim that reference to objective facts is an indispensable feature of moral concepts. If that is his view, then the rejection of objective moral facts would commit him to a moral error theory.

[33] For example, Joyce (2001: 137) claims that evolutionary explanations of morality have "error theoretical implications". Joyce does not defend this view in later publications, as we will see below (cf. Joyce 2013a: 354).

successful, provide any moral reasoner with reasons to abandon his or her objectivist moral beliefs.[34]

### 1.4.2 *Conditional* Evolutionary Undercutting (Street)

Evolutionary considerations have been claimed to give us reason to reject metaethical theories. This form of evolutionary metaethics is 'theory-debunking', and it aims at showing that moral beliefs are *undercut*, insofar as we assume that a particular metaethical view is true (Joyce 2013b). The most widely cited evolutionary debunking argument, that of Sharon Street (2006), is of this form.

Street intends to show that moral realism ought to be rejected in favour of moral anti-realism.[35] The way in which Street understands moral realism makes it a kind of moral objectivism as stipulated in this thesis. Street's "opening premise" is that "the forces of natural selection have had a tremendous influence on the content of human evaluative judgments" (Street 2006: 113). Many of the deepest evaluative tendencies that are common to the morality of all cultures qua content enhanced human fitness in the evolutionary sense. If we want to explain why these and other beliefs are common to all human societies, evolutionary explanations provide the answer.[36] From this follows a dilemma for realists:

> The basic problem for realism is that it needs to take a position on what relation there is, if any, between the selective forces that have influenced the content of our evaluative judgments, on the one hand, and the independent evaluative truths that realism posits, on the other. Realists have two options: they may either assert or deny a relation (Street 2006: 121)

Asserting that there is a relation commits realists to defend what Street calls a 'tracking account', according to which evolutionary forces have tended to make

---

[34] This is too easy, of course. Provided we have some justification to hold objectivist moral beliefs, to some degree D, we would have to be justified in believing the conclusion of the metaphysical debunking argument at least to a degree higher than D to give up our moral beliefs.

[35] Strictly speaking, Street's aim is wider, as she targets normative realism rather than moral realism. I take the moral to be a subset of the normative and thus will interpret her argument as applying to moral realism.

[36] Street also remarks that there is a striking continuity of human evaluative tendencies and the basic evaluative tendencies of other animals (Street 2006: 117). Joyce (2001: 135ff) also asks *why* humans developed moralising tendencies and then argues that "natural selection" is the answer. Here, Joyce partly relies on evolutionary theory to fulfil an explanatory deed.

our normative judgements track the attitude-independent normative truth *because* it promoted our ancestors' reproductive success to make true normative judgements (or to make proto versions of them) (Street 2006: 121ff, 134). However, Street argues that her non-realist account of

> why we tend to make some evaluative judgments rather than others … wins the competition hands down [against the realist's account], judged by all the usual criteria of scientific inquiry (Street 2006: 129)

Street argues that the tracking account is bad science; she insists that it is far more scientifically respectable, in terms of parsimony, clarity, and explanatory power, to explain at least some of our normative judgements in terms of the evolutionary advantageousness that came with making those judgements (Street 2006: 129ff). Denying that there is a relation lands realists on the second horn of Street's dilemma, which leads to

> the implausible sceptical conclusion that our evaluative judgments are in all likelihood off track, for our system of evaluative judgments is revealed to be utterly saturated and contaminated with illegitimate influence" (Street 2006: 122)

This shows, argues Street, that we would have to conclude that "most of our evaluative judgments have nothing to do with the truth" (Street 2006: 122). Realists cannot appeal to the power of reasoning to, in a sense, correct the "distorting" influence of evolutionary forces on our evaluative tendencies, because rational reflection will derive from the starting fund of basic evaluative tendencies that have been influenced by evolutionary forces (Street 2006: 124). Hence, Street concludes that "in the absence of an incredible coincidence, most of our evaluative judgments are likely to be false" (Street 2006: 125, 2011: 14).[37] In conclusion, Street notes that either option is unacceptable for normative realists and thus proposes to reject normative realism.[38]

---

[37] Sinnott-Armstrong (2006a) defends a related and very influential argument that purports to show that moral psychology shows moral beliefs to be probably false; see Ballantyne and Thurow (2013) for discussion. It will become clear in section 1.5. of this chapter why I think that such arguments fail.

[38] Ruse sometimes also claims that his debunking argument has a meta-theoretic conclusion (e.g. Ruse 1998: 254). Some philosophers mistakenly attribute to Street the claim "that our moral beliefs are probably false" (Brosnan 2011: 52). This is incorrect, as Street only claims that moral realism is probably false.

Street can be read as offering a conditional evolutionary undercutting argument on the second horn of her dilemma. In defence of the second horn of her dilemma, she argues that *if* moral realism is true, then all beliefs *thus construed* are likely to be false or it would be a coincidence if they were true.

### 1.4.3 Unconditional Evolutionary Undercutting (Joyce)

Finally, justification-debunking takes evolutionary explanations of morality to show that some or all moral beliefs are epistemically unjustified. Richard Joyce has offered the most influential argument to this effect (Joyce 2006, 2013a, 2016c, 2016d).[39] It aims at an undercutting conclusion, without presupposing that the argument applies only to particular metaethical views.

First, Joyce argues that we have "no reason to think in the case of the moral sense that natural selection is likely to have produced true beliefs" (Joyce 2006: 182, 2013a: 353). However, when we learn about the fact that our moral beliefs are generated without regard for their truth, we should be agnostic about the truth of our moral beliefs:

> Were it not for a certain social ancestry affecting our biology … we wouldn't have concepts like *obligation, virtue, property, desert,* and *fairness* at all … [T]herefore, it would appear that once we become aware of this genealogy of morality we should (epistemically) cultivate agnosticism regarding all positive beliefs involving these concepts until we find some solid evidence either for or against them (Joyce 2006: 181)

Second, like the moral domain, there are other domains of inquiry, such as mathematics, that are concerned with unobservable or even causally inert properties too, but these domains of inquiry do not face the same epistemic fate as morality inasmuch as it is likely that natural selection produced true beliefs in these domains:

> Can we make sense of its having been useful for our ancestors to form beliefs concerning *rightness* and *wrongness* independently of the existence of rightness and wrongness? Here I think the answer is a resounding 'Quite possibly.' ... Not so for the mathematical case. Were someone foolish enough

[39] In earlier publications, Joyce suggested that the Darwinian hypothesis has "error theoretic implications" (Joyce 2001: 158, 2006: 223), but he recants this claim in later publications. Some discussions of Joyce's argument proceed under this mistaken assumption, e.g. Mason (2010: 775) and James (2011: 181).

> to doubt that 1+1 = 2, the plausibility of the evolutionary story concerning how having this belief enhanced our ancestors' fitness would evaporate. (Joyce 2006: 183)

Third, moral judgements cannot be vindicated by a naturalist reduction of moral properties to non-moral properties, which is a reduction that would be required to prevent moral judgements from being undermined, or so Joyce argues:

> If the moral facts are reducible to the non-moral facts invoked in the genealogical explanation, then the former cannot be eliminated on grounds of parsimony, any more than cats should be eliminated from our ontology because we can explain them in terms of physics. […] [However], no such naturalism can accommodate the sense of inescapable practical authority with which moral claims appear to be imbued. (Joyce 2006: 189–90)

Joyce argues that a failure to explain the practical authority of moral claims is an all-things-considered reason to reject moral naturalism and goes on to argue that moral naturalism does fail for "independent arguments" (Joyce 2006: 210).

Across his multiple publications on the topic, Joyce has derived two different conclusions from this argument and apparently worked under two different assumptions.[40] The 'bold' Joyce (2006) concludes that the justificatory status of moral beliefs is "undermined" on the assumption that moral beliefs are defeasibly justified (Joyce 2006: 217). He affirms this when he writes that the evolutionary challenge has shifted the burden of proof to defenders of morality to explain how objectivist moral beliefs can be justified, "but until that is accomplished, they cannot be considered justified" (Joyce 2016c: 152–4). Joyce concludes that

> the fan of morality has some work to do if justification is to be established or reinstated…. The role of the [evolutionary defeat challenge] is to place the burden of proof on the shoulders of those who believe in justified moral belief. (Joyce 2016c: 154–5)[41]

[40] A minor remark: Joyce explicitly defended his argument in reference to a sensitivity principle (Joyce 2001), but no longer thinks that modal analyses are applicable at all Joyce (2016d: 132).

[41] Joyce (2006) draws a metaphysical conclusion. With moral naturalism refuted (according to Joyce's argument), "Ockham's Razor really can come in and do its thing, for non-naturalism and super-naturalism do posit extra ontology in the world, but the presence of the non-moral genealogy shows this ontology to be explanatorily superfluous" (Joyce 2006: 209–10).

Joyce is 'bold' here because he implies that his challenge works even on the generous assumption (from the point of view of the moral objectivist) that moral beliefs are prima facie justified (when he writes, for example, that fans of morality have to show how justification is to be "reinstantiated" (Joyce 2016c: 154), and that the evolutionary challenge defeats this justification).

The 'modest' Joyce (2016d), however, denies his opponents the right to assume the prima facie justification of moral beliefs in the first place (Joyce 2016d: 139). He is modest, insofar as he does not grant too much to the objectivist. Instead, he challenges "fans of morality" to explain truth-tracking, and he insists that the explanation has to be "plausible" and that the explanatory role of moral properties must not be "mysterious or yet-to-be-explained or hand-wavy" (2016d: 142). Indeed, Joyce (2016d, 2018) seems to take the evolutionary challenge as an argument against epistemic conservatism (roughly, the view that beliefs of a certain type are prima facie justified) in the moral domain.[42]

It is difficult to say what legitimate assumptions are in metaethical arguments (cf. Sinclair forthcoming). I will not venture an answer in this thesis.[43] Instead, I will focus on 'bold' Joyce who grants that moral beliefs are defeasibly justified and tries to show that the evolutionary challenge undercuts this justification. This is in line with my assumption, set out in the main introduction of this thesis, that there is 'a moral epistemology' for moral objectivism.[44]

This completes my review of the three predominant evolutionary debunking arguments. Though importantly different, I hope to have shown that there are two common features between them that allow me to sort them according to whether they, if successful, yield a rebutting defeater or an undercutting defeater or our objectivist moral beliefs. Thus far, I have clarified the empirical background of the evolutionary defeat challenge, set aside prescriptive evolutionary ethics, and, turning to metaethics, introduced the three predominant evolutionary debunking

[42] See Harman (1988) and Lycan (1988) for a defence of conservatism.

[43] Though see Klenk (2017d) for a review of Joyce's modest debunking argument, which suggests that it might be giving up too much.

[44] See Wielenberg (2016b) and Clarke-Doane (2016a), who also interpret Joyce in this light.

arguments. I will now consider how my research question is relevant for the three predominant metaethical debunking arguments.

## 1.5 Defeat and Evolutionary Debunking Arguments

My construal of the evolutionary defeat challenge raises a question: how is my research question relevant for these different metaethical debunking arguments? Am I not looking only at a very narrow aspect of the 'metaethical implications of evolutionary explanations of morality' by describing evolutionary defeat as I do?

I want to answer these questions by showing how the debunking arguments that I discussed depend on finding an answer to my research question. In answering this question, I switch from exposition to offensive and argue that evolutionary rebutting arguments, which would show that all moral beliefs are false, can be set aside and that the success of evolutionary undercutting arguments, of both conditional and unconditional form, depend on an answer to my research question.[45] I will raise two points that apply to both Ruse's rebutting argument and the first horn of Street's dilemma, where she argues that moral realism fails on scientific grounds.

### 1.5.1 Rebutting Defeat is Straightforward…

The epistemic consequences of rebutting arguments are relatively straightforward. If truth-debunking could successfully show that there are no moral facts, it would show our moral beliefs to be false (assuming a correspondence theory of truth), and it seems evident that we should then revise them.[46] This is because it is comparatively straightforward to understand why we should (epistemically) give up a belief upon learning that the belief is false. First, virtually all epistemologists agree that one ought to hold a belief only if the belief is true (cf. Williams 1973: 137). Second, if there are no facts for the content of our moral beliefs to correspond with, all (positive) beliefs would be systematically false. Hence, if it could be established that the truth conditions of our moral beliefs are such that no (positive)

[45] Note that most discussions in the literature also focus on undercutting arguments.

[46] I am oversimplifying a bit here: rebutting defeaters are not necessarily much better understood than undercutting defeaters, as we will see in chapter 2. But the consensus is that they are *less* problematic than undercutting defeaters.

moral belief is true, and there would be no possibility of construing differently the truth conditions of moral beliefs, then it seems clear that we ought to (epistemically) give up our moral beliefs.

I am not saying that there are no interesting issues with rebutting arguments in general. On the contrary, there is some debate about how to formally draw the distinction between information that shows that your belief is false and information that impugns the justification of your belief without showing that it is false.[47] Moreover, one might argue that it is psychologically impossible to give up our deeply held moral beliefs (about the fundamental equality of persons, for example) and, given that ought implies can, it would be false that we ought to give up all our moral beliefs, even if we find out that they are all false.

There is a clear intuitive distinction, however, between rebutters and undercutters, which I take sufficient for this thesis. Concerning the point about 'ought implies can', there is little evidence that there are moral beliefs that are psychologically impossible to give up. In experiments that assess how people perceive the objectivity of moral judgements, researchers find both intra- and inter-subjective variability about what people take to be a moral issue and whether they consider it to be an objective moral truth (Beebe and Sackris 2016; Fisher et al. 2017; Goodwin and Darley 2008, 2010, 2012; Wright et al. 2013). Since the variability people's judgments about morality is evident, we have little reason to expect that there are some moral judgements that are 'psychologically inescapable' for all or even most humans, and so many would have to give up those moral judgements upon learning that they are false.

Hence, I take the potential effects of successful truth-debunking arguments to be comparatively uncontroversial and thus less interesting to study. If rebutting arguments work, it is clear why moral beliefs ought to be given up. Of course, the crucial question is whether they do succeed, and there are reasons to be sceptical about this, as I argue in the next section.

[47] Cf. Kotzen (2010); Melis (2014); Pryor (2013).

### 1.5.2 … and Unlikely to Succeed

The rebutting arguments of Ruse and Street (that is, the arguments that support the first horn of Street's dilemma) must show that there are no moral facts. These arguments are unlikely to succeed.

Evolutionary rebutting arguments depend on the assumption that it is impossible to provide a reductive account of objective moral facts according to which all moral facts are natural facts (Joyce 2013a: 358, 2016b).[48] If such an account were possible, following Sturgeon (1988 [1995]), we could say what moral facts are without invoking sui generis non-natural properties.[49] In that case, truth-debunking could not succeed easily by appealing to the principle of parsimony insofar as it is an open question whether objectivism is less parsimonious than non-objectivism. Whether a reduction of moral properties to non-moral properties is possible (or whether all moral concepts can be derived from conceptual truths) is beyond the scope of this thesis because such arguments would be independent from considerations about the evolutionary origins or our moral beliefs.

Of course, some objectivists maintain that moral properties cannot be reduced to natural properties, but there are problems for rebutting arguments nonetheless, as the next point illustrates (Cuneo 2014; Enoch 2011b; FitzPatrick 2008; Shafer-Landau 2003; Wielenberg 2014).[50]

It is an open question whether we should call for Ockham's razor to do its work even if moral objectivism is less parsimonious, from the perspective of an evolutionary explanation of morality, than other metaethical views. Let's grant that objectivist moral facts are not needed to explain why certain moral beliefs are shared by most humans or why humans have the capacity to make moral judgements in the first place (Buchanan and Powell 2015; though see Huemer 2016), this does not show that there are no other reasons to postulate the existence of moral facts.

---

[48] Since the late 1990s, some have maintained that belief aims at knowledge, not just mere truth; see Williamson (2000). Insofar as truth is necessary for knowledge, this complication need not concern us here.

[49] Of course, pointing out that we have no reason to believe in moral facts does not imply that we have reason to disbelieve them; see Joyce (2006: 210).

[50] See Klenk (2016c) for a review.

On the one hand, moral objectivists have filled books with arguments that purport to show that moral facts play an important role in theorising about morality and in the practical aspects of morality. For instance, moral facts might be required in other explanatory projects, such as attempts to offer a moral semantics (Blackburn 1984: 192; Geach 1965; Schroeder 2008). On the other hand, non-explanatory projects such as deliberation might vindicate ontological commitment too (cf. Enoch 2011b).[51] There is also some empirical research that investigates whether there might be prudential reasons to posit the existence of objective moral facts. For example, framing behavioural options as objectively impermissible influences behaviour, at least in some cases, even when transgressions could not be punished (Rai and Holyoak 2013; Young and Durwin 2013). It would be unwarranted to assume that objective moral facts are explanatorily idle with regard to all relevant explanatory projects and that explanatory parsimoniousness is the sole criterion of ontological commitment.

These considerations about other explanatory or relevant non-explanatory 'jobs' for moral objectivism are particularly relevant for the assessment of the first horn of Street's dilemma. Street at times alludes to the fact that the *evolutionary* explanation of our moral beliefs is not what matters: what matters is that there is *some* causal explanation of our evaluative tendencies that does not presuppose their truth (Street 2006: 155). Hence, an account that shows, implausibly, how all evaluative tendencies are purely cultural products would be just as acceptable for Street's purposes, as long as such an account nowhere presupposes the truth of moral beliefs (more on this in chapter 3). However, in light of this admission, it seems that the first horn of Street's dilemma is really just the argument from parsimony again. The moral properties posited by moral objectivism are explanatorily redundant, and we might want Ockham's razor to come in and do its work. But, as we have seen, whether this should be done depends on the overall balance of explanatory 'jobs' that we want objectivist moral properties to do and on

[51] Shafer-Landau (2007: 323) argues that the 'job' of moral facts is not to "explain nonnormative phenomena but rather to specify ideals, or standards that in some way must be met". Sober (2009: 141) similarly suggests that normative ethical propositions have the job of telling us how we ought to act, not of explaining why we in fact act as we do". See Joyce (2016d: 134) for criticism.

whether a role in the explanatory project is the sole way in which a kind of property can earn its ontological keep.[52]

Of course, this is not to deny that the lack of an explanatory role for moral facts (in explaining our observations) is a cost to moral objectivism. It would be good for moral objectivists if moral facts would play an explanatory role in explaining the origins of our moral judgments and this route to justifying moral judgments is blocked (Tersman 2006: 46); we might not need to postulate objective moral facts to explain shared moral beliefs (cf. Mackie 1977). Insofar as moral objectivism gains any support from being a good explanatory account of the origins of our moral judgments in the first place, both evolutionary rebutting and the first horn of Street's Darwinian dilemma would take away support for moral objectivism.

However, my final point against evolutionary rebutting is that it is doubtful that moral objectivists presuppose that the justification of our moral judgments depends in an important sense, let alone exclusively, on the best explanation of our observations. Hence, it is doubtful whether parsimony-based arguments could defeat our non-empirically justified moral judgments in the first place.

Consider an analogy. Suppose I arrive at the office one day and see little black pellets sprinkled all over the place. I wonder what explains this and speculate that there might be mice in this old Dutch building (yuck!). But suppose my trustworthy colleagues tell me that they had a communal breakfast with 'hagelslag', a Dutch favourite of little chocolate pellets that are supposed to go on your toast but very often end up elsewhere. It would be odd to say that the best explanation of the

[52] Finally, Elliott Sober has argued that Ockham's Razor simply fails to apply in the moral case (Sober 2016: 264–8). The *lex parsimoniae* is first and foremost a methodological guideline rather than an epistemological or metaphysical principle. It tells scientists to develop theoretical models with as many explanantia as necessary, but as few as possible, to explain the particular explanandum in question. One motivation for following this methodological guideline is that fewer explanantia are easier to falsify empirically. Another motivation is that models that postulate two independent causes for a given event are, in general, less probable than models that postulate one cause Sober (2016: ch. 2). Sober argues that these considerations do not apply in the moral case. Moral facts supervene on non-moral facts and so they are not probabilistically independent. If Sober is right, which I unfortunately cannot address here, evolutionary rebutting, based on explanatory parsimony arguments, might not be applicable in the moral case at all. On that note, Joyce (2006: 187) writes that "there is no rationale for requiring that the moral facts be describable in the language of the natural sciences; a purely ontological relation will suffice".

black pellets is that my colleagues had a Dutch breakfast *and* that there are no mice in the building. Adding a negative existential claim does not add anything to the explanation (cf. Horn 2017: 367–70). One might say that if the only justification for believing that there are mice in the building would be an abductive inference from my observation of the pellets, then learning about the alternative explanation offered by my colleague might undercut my doxastic justification. But here the analogy breaks down, because the best explanation of our observations is *not* the only source of justification for our moral judgments, or so moral objectivists argue. Moral objectivists have put forward many arguments for their position, but few (if any) rely on the claim that moral facts are indispensable in causally explaining moral phenomena. Thus, pointing out that such facts are dispensable in causally explaining moral phenomena should, by itself, not trouble objectivists in the least. Objectivists can accept that moral facts do not earn their place in the best ontology in virtue of their explanatory power in genealogies of moral beliefs and continue to defend the many other arguments in favour of their view.

An exhaustive consideration of moral objectivism must take these considerations into account, counting as 'plausibility points', as Enoch (2011) suggests. As should be clear, however, my project is not an exhaustive assessment of moral objectivism, and so these considerations can safely be set aside. Evolutionary rebutting succeeds only if certain controversial views in these metaphysical and methodological debates are true.

I will therefore set aside evolutionary rebutting (thus ignoring Ruse-style arguments) and, with regard to Street's argument, my main research question will apply to the second horn of her Darwinian dilemma, according to which objectivist moral beliefs are either coincidentally true or likely to be false.[53] On this interpretation, Street's undercutting argument is congenial to the undercutting argument of Joyce, at least on the modest interpretation of Joyce's arguments discussed above. The question is therefore whether evolutionary explanations of morality give us sufficient epistemic reason to revise our defeasibly justified moral

[53] I do not thereby unduly restrict my assessment of the evolutionary challenge because Street's dilemma has force only if both defences of the two horns of her dilemma succeed. If one horn can be rejected, as many philosophers have argued, the strength or convincingness of the other horn is irrelevant for the assessment of the dilemma.

beliefs, without showing that they are false. Does evolution undercut all moral judgements? In the next section, I show that there are threats to the survival of (evolutionary) defeat.

## 1.6 Threats to the Survival of Evolutionary Defeat

This section will support my claim that the survival of (evolutionary) defeat is in jeopardy by reviewing existing criticism of the evolutionary defeat challenge.

Take as the starting point the arguments of Street and Joyce. Street's answer to my main research question on the second horn of her Darwinian dilemma is that we ought to give up our objectivist moral beliefs because they are shown to be in all likelihood false or only coincidentally true. According to Joyce, we ought to give up our moral beliefs because their origins can be shown to lack an appropriate explanatory relation to their truth. In assessing these claims, we can think back to figure 0.1 from the main introduction, which I reproduce here with more details filled in:

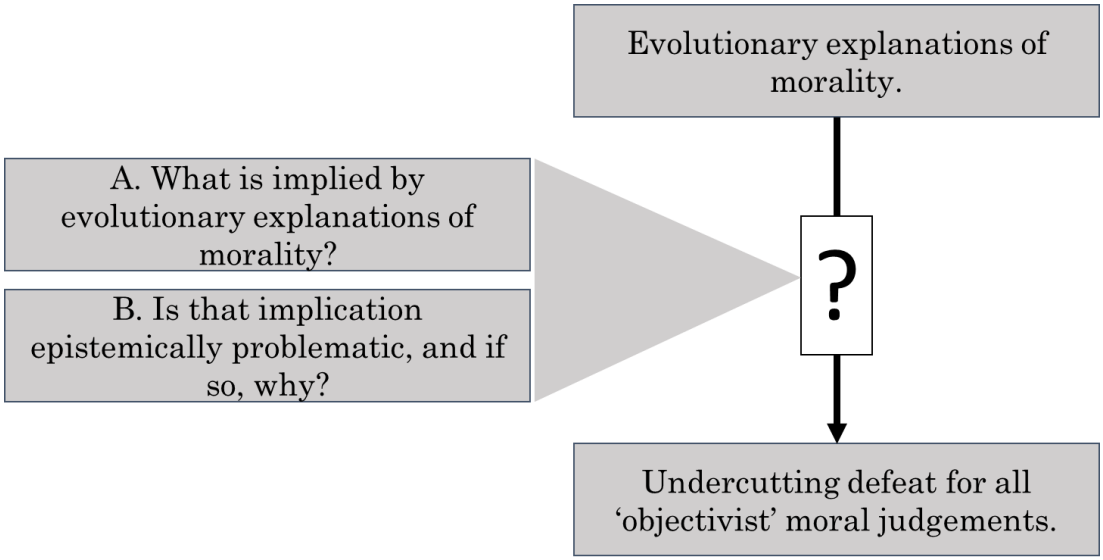


*Figure 1.1 How can evolution defeat objectivist moral judgements?*

We have to answer two questions to establish whether evolutionary explanations of morality lead to undercutting defeat for all objectivist moral judgements. First, what information do evolutionary explanations unearth (answering A)? Second, does that information constitute an undercutting defeater (answering B)? The recent discussion of these questions has been mainly negative, putting pressure on both the claim that the evolutionary challenge might

instantiate anything of epistemic relevance and the defeating power of whatever the challenge might instantiate.[54]

### 1.6.1 Defeat by Error or Coincidence

Katja Vavova has critically examined Street's suggestion that evolutionary explanations of morality show that our moral beliefs are probably false (Vavova 2014a). Vavova starts by rejecting the view that the evolutionary challenge might shift the burden to defenders of morality to provide independent (i.e. non-moral) reasons for thinking that our moral beliefs are *not* mistaken. Demanding *independent* reasons in favour of the reliability of a belief or type of belief is just the challenge of the radical sceptic. However, to distinguish themselves from the radical challenge, evolutionary explanations of morality must show us that our moral beliefs are probably mistaken: they must give us "evidence of error" (Vavova 2014a: 82, 2015: 112).[55] Vavova avows that evidence of error would be a reason to revise our beliefs (Vavova 2018) but denies that the evolutionary challenge can provide evidence of error (Vavova 2014a, 2015). The reason for this is that a substantial assumption about the moral truth would be required of proponents of the evolutionary challenge, an assumption that they cannot make. Several other philosophers have endorsed a thesis similar to Vavova's, suggesting that the evolutionary hypothesis cannot show us that our moral beliefs are probably false (Clarke-Doane 2017a; Lutz forthcoming; Sinclair forthcoming; Warren 2017).

[54] A number of responses that have been given to evolutionary debunking assume that evolutionary defeat arises but try to reject it via a *reductio ad absurdum,* suggesting that a successful evolutionary defeat challenge would also defeat lots of other beliefs that we would not want to give up; see, for example, Graber (2012), Severini and Sterpetti (2017), Tropman (2014), and Slater (2014). I do not discuss such arguments because they take for granted what I want to scrutinise: can an evolutionary defeater arise in the first place?

[55] In relation to the previous discussion about rebutting arguments, I should note that we may get evidence of error, according to Vavova, but no proof that our moral beliefs are wrong, as the proponent of a rebutting argument would have it. Vavova herself does not discuss this relation, because she excludes debunking arguments with metaphysical conclusions from the start.

Hence, the claim that our moral beliefs can be shown in all likelihood to be false cannot be maintained.[56]

What about Street's (2006, 2011, 2012) claim that evolutionary explanations of morality might imply that our moral beliefs are only coincidentally true and that we ought to give up beliefs upon learning that they are coincidentally true, if true at all? This proposal has been challenged on three accounts. Shafer-Landau and Cuneo appeal to the essence of moral concepts and object to the idea that the evolutionary challenge cannot instantiate a coincidence, because at least some moral truths, which they call the 'moral fixed points', are conceptual truths. It is therefore irrational to believe that they could be otherwise (Cuneo and Shafer-Landau 2014). Since we actually endorse many of the moral fixed points, and it would be incoherent to fail to endorse the morally fixed points, if we were to adopt moral beliefs at all, there is no coincidence, contrary to Street's claim.[57] In contrast, Mogensen (2014) assumes that the set of (conceptually) possible moral truths is infinite. However, starting off by considering the fact that there is no general ban on believing coincidences, he goes on to show that there is no helpful notion of 'coincidence' that seems epistemically problematic and that a probabilistic approach to determining whether a coincidence occurs ultimately falters. Thus, evolutionary explanations of morality do not show the truth of our moral beliefs to be coincidental in the first place (Mogensen 2016a, 2018). A third way to reach the same conclusion is to take seriously an important implication of evolutionary explanations of morality: if certain types of norms, such as norms about reciprocity, fulfil an evolutionary function in group animals like us, then we should expect these norms to be stable across slightly different evolutionary trajectories. That we endorse some such norms today is not coincidental but to be expected (Clarke-Doane 2016a: 29).

Hence, the claim that our moral beliefs might be coincidentally true or probably false cannot be substantiated on evolutionary grounds. Though it seems

[56] I assume here that we are *not* considering a truth-debunking argument, for the reasons elucidated above. If a truth-debunking argument worked, all moral beliefs would certainly be false, not just probably.

[57] Bedke (2009, 2014) raises an issue with coincidence by holding fixed only the causal order but not the metaphysical relations that hold between facts. It is not clear why objectivists should accept this assumption.

correct that we ought to give up beliefs if we have (sufficient) evidence of error, we do not gain evidence of error through evolutionary explanations of morality. And though coincidences are sometimes reasons to consider alternative explanatory hypotheses or to question one's background assumptions, we do not gain evidence of problematic coincidences through evolutionary explanations of morality.[58] Thus, the 'error' or 'coincidence' view of the evolutionary defeat challenge does not look promising and I will not consider these views in what follows.

### 1.6.2 Defeat by Lack of Sensitivity or Safety

Instead, many have questioned the idea that evolutionary explanations of morality might imply something about the epistemic *sensitivity* or epistemic *safety* of our moral beliefs. S's belief that p is sensitive if and only if it is true that had p been false, S would not have believed that p. S's belief that p is safe if and only if it is true that p is true in all nearby worlds where S holds the belief that p. Both conditions have been proposed as necessary for epistemic justification. The idea is, roughly, this: a belief's epistemic sensitivity and epistemic safety is a necessary conditions for the belief to be justified and evolutionary explanations of morality show us that one (or both) of these conditions are *not* satisfied in the case of moral beliefs (contrary to what we might have thought).[59] Criticism of this interpretation of the evolutionary defeat challenge takes two forms.

On the one hand, some philosophers point to epistemological reasons against sensitivity or safety, sometimes suggesting that neither sensitivity nor safety is plausibly seen as necessary for justification (Bogardus 2016; Srinivasan 2015). The thought is that the evolutionary challenge shows that our moral beliefs lack something that is not epistemically important in the first place. On such views, the evolutionary defeat challenge fails because whatever evolutionary explanations of morality show is not epistemically relevant.

On the other hand, some philosophers accept the epistemic relevance of safety and sensitivity but go on to deny that evolutionary explanations of morality show our moral beliefs to lack either quality. For example, Clarke-Doane interprets the

[58] Cf. Horwich (2016); Schlesinger (1991).

[59] I consider a neglected variant of sensitivity, called adherence, and its relevance for my main question in section 7.4.1.

evolutionary challenge as a problem about explaining the reliability of our moral beliefs (Clarke-Doane 2015, 2017a, 2017b). After considering several proposals, Clarke-Doane concludes that establishing the safety and sensitivity of a belief is sufficient to explain the reliability of a belief such that the reliability commitment of moral objectivism, which I outlined in the main introduction, is satisfied. Importantly, he denies that the evolutionary challenge provides reasons to doubt or deny the safety or sensitivity of moral beliefs. On this view, evolutionary explanations of morality fail to imply a justificatory loss for our moral judgement too. I will now consider two interpretations that seem more promising.

### 1.6.3 Defeat by Lack of Reliability

In the main introduction of the thesis, it was noted that proponents of the reliability view interpret the evolutionary hypothesis as raising an explanatory challenge about the need to explain how considered moral beliefs are by and large reliable. Supporters of this interpretation can be allocated to two different camps.

According to the first camp, the evolutionary hypothesis challenges moral objectivists to explain how moral beliefs are reliable, in the sense of being correct more often than not (Baras 2017; Crow 2016; Schechter 2010, 2013, 2018; Sinclair forthcoming; Talbott forthcoming; Tersman 2016, 2017). Since this is only a challenge, not a refutation, the adherents of this camp consider it a separate question whether objectivists *can* provide an adequate explanation. If no explanation is in principle possible, these philosophers think, *then* the justification of moral beliefs will be undercut. However, many believe that objectivists can offer adequate explanations. The most favoured response to the reliability challenge is a so-called third-factor explanation, which aims at showing how moral beliefs are reliable, despite their evolutionary origins (cf. Enoch 2010). I will be concerned with third-factor explanations in detail in chapters 4 and 6 and will thus not describe them in greater detail here. For the moment, the important point is that *if* the evolutionary hypothesis merely raises a *concern* about the reliability, then this concern can apparently be met, or so these philosophers argue.

Adherents of the second camp also think that the evolutionary hypothesis raises an *explanatory* worry, but claim that this is not only a challenge but also an undercutter in its own right. This is how Joyce presents his argument in recent

publications (Joyce 2016c). Following Joyce, Lutz (forthcoming) argues that the evolutionary hypothesis shows us that the best explanation of our moral observations does not imply their truth and that this instantiates undercutting defeat (without a detour about the need to explain the reliability of our moral beliefs).[60] Indeed, Lutz and Joyce claim that showing that the best explanation of some evidence does not entail the truth of a belief base is just what it *is* to undercut a belief.

However, as discussed in overview section of the main introduction, Clarke-Doane (2015, 2016a, 2017a) has raised a deep problem for such views.[61] Irrespective of whether there is an epistemic requirement to show how the best explanation of a belief's base implies its truth, such a requirement cannot be legitimate in an effort to undercut defeasibly and non-empirically justified moral beliefs. Objectivist accounts of morality would not assume that the truth of moral beliefs is implied by the best explanation of why we come to hold the moral beliefs – we thus cannot rely on an epistemic principle that would imply that such beliefs would not be justified to begin with to defeat these beliefs without begging the question against moral objectivists.

Looking at both camps, we find the following situation. The approach favoured by the second camp, deriving an undercutter directly from evolutionary explanations of morality, does not succeed in light of Clarke-Doane's objection, which I will address in chapter 7. The approach favoured by the first camp, deriving an undercutter from the alleged inability of objectivists to explain the reliability of moral beliefs, seems to fail in light of the availability of explanations of the reliability of moral beliefs.

Hence, without an answer to Clarke-Doane's rebuttal of the second camp of the reliability view or a demonstration that third-factor explanations fail, it is doubtful whether worries about explanatory connections can substantiate the evolutionary defeat challenge.

---

[60] Similar principles are suggested, though not always explicitly endorsed, by Setiya (2012), Braddock (2017), Locke (2014), and Schechter (2018).

[61] See, for example, Warren (2017) and Hill (2016) for recent endorsements of this view.

### 1.6.4 Defeat by Disagreement

In light of some of these problems, the disagreement view already discussed in the introduction gained currency. The disagreement view has it that the evolutionary challenge is intimately connected with the epistemic significance of disagreement (Bogardus 2016; Mogensen 2016a, 2017; Tersman 2016; White 2010). They claim that evolutionary explanations of morality might establish a kind of intractable counterfactual moral disagreement with our hypothetical peers (to wit, ourselves had we taken a different evolutionary path) and that that intractable disagreement provides an undercutting defeater of our moral beliefs. On this view, the evolutionary challenge succeeds, insofar as it shows that there is an epistemically significant disagreement with our hypothetical peers. Since many accept that disagreement is epistemically significant, at least under certain conditions, the disagreement view seems to offer a promising route to substantiate the evolutionary defeat challenge (Elga 2007; Kelly 2005).[62] The problems of alternative interpretations of the evolutionary defeat challenge, and the strong prima facie support for the epistemic relevance of disagreement, make the disagreement view a strong contender as an answer to my research question. I will come back to the account in chapters 5 and 6.

## 1.7 Concluding Remarks

In this chapter, I provided the empirical, historical, and metaethical background of my thesis. The best evolutionary account of morality maintains that a normative capacity evolved as a tool to enhance cooperation and that selective pressures acted both on the individual and the group. Though I did not specify the moral as a subset of the normative, I have argued that it is enough to show that a normative capacity, as well as the content of some normative beliefs, evolved to make a sufficient case for the evolutionary origins of morality. To set aside first-order normative questions, I identified the pitfalls of prescriptive evolutionary ethics of the early 20th century and then turned to metaethics. I introduced the three most prominent evolutionary debunking arguments in metaethics, which aim at both rebutting and undercutting conclusions. A successful rebutting argument would have

[62] Though see Enoch (2011a).

straightforward epistemic consequences (we would have to give up false beliefs), but such arguments rely on controversial philosophical assumptions that are beyond the scope of this thesis. This leaves us with evolutionary undercutting arguments and the question whether they can support an undercutter. In regard to this question, I discussed the most prominent interpretations of evolutionary defeat. The best contenders seem to be the reliability view and the disagreement view. The former seems to be the most prominent account of evolutionary defeat, but it faces a strong objection. The latter is less frequently defended, but it faces no obvious objections yet. Hence, the emerging picture of the current debate is that the evolutionary challenge succeeds, but only if it piggybacks on the epistemic significance of disagreement. I will come back to both views in chapters 5 to 7. For now, we will inspect more closely the conditions of epistemic defeat in general which will later help us see what to look for in an account of evolutionary defeat.

This page intentionally contains only this sentence.

# 2 What is Epistemic Defeat?

## Reader's Guide

Most things are what they seem to be. The venerable tables, coffee cups, trees, and cats on mats invoked in many philosophical thought experiments not only seem to be tables, coffee cups, trees, and cats on mats, but are. At least, this is so if we are to assume that scepticism of a Cartesian bent is misguided.[1] Cartesian scepticism presupposes that to know anything, one must be able to exclude on independent grounds a sceptical hypothesis according to which things are not what they seem to be.

However, even as non-sceptics we do not believe everything we see. Instead, we give most beliefs the benefit of the doubt. Perception provides us with seemings that we take at face value in ordinary circumstances (Pollock and Cruz 1999). Scientific methods and confirmation by peers allow us to rely on seemings in extraordinary circumstances, such as when the stakes are high or the conditions for perception bad. However, it seems prudent to keep open the possibility of error. We are fallible; our beliefs might still turn out to be false and some things might not be what they seemed to be after all.[2] The phenomenon of defeat shows that a person's belief can have a particular epistemic status such as 'knowledge' or 'justification' at one time but lose that status at another time.[3] An understanding of defeat is thus crucial to be able to say that we do have knowledge and justified

[1] Cf. Descartes (1998 [1637]) for the *locus classicus*. Stroud (1984) and Unger (1978) provide modern defences.

[2] Doughterty and Rysiew (2009: 123) note "the near-universal acceptance of fallibilism in epistemology". Examples of this view are legion. Williams (2001: 5) and Kitcher (2012: 168) both insist that "we are all fallibilists" now; cited in Climenhaga (2017). Some might also believe, as Rorty (1979) argued, that the death of foundationalism has left epistemology as an impossible and unnecessary discipline, given that epistemologists have traditionally attempted to discover some area of human belief that transcends the possibility of doubt. Adopting fallibilism will reinstate the significance of epistemology even on such a sceptical view.

[3] Though some philosophers doubt that defeat of knowledge is possible; see Lasonen-Aarnio (2010b) and Baker-Hytch and Benton (2015).

beliefs now, even though they are not certain, as opposed to holding our beliefs in a limbo of epistemic uncertainty.

In this chapter, I introduce the concept of epistemic defeat in greater detail and sketch out why current objective accounts of undercutting defeat, which specify perspective-independent rules regarding when a belief is undercut, fail to imply that evolutionary explanations of morality undercut all non-empirically justified moral beliefs. To underscore my point that an objective account of undercutting defeat is required to make the evolutionary defeat challenge suceed, I then argue that subjective accounts of undercutting defeat, which maintain that defeat is perspective-dependent, are doomed to fail. Since the nature of defeat plays a crucial role in my thesis, I take more time in this section to introduce the concept of defeat and its relevance for philosophy, but then pick up the pace when pointing out how and why current objectivist accounts of defeat fall short of supporting the evolutionary defeat challenge. Let's begin with the relevance of epistemic defeat.

Though epistemic defeat might not be how many people think about the updating and revising of their beliefs, it nonetheless seems to be a natural description of a significant aspect of our epistemic lives. In the main introduction of this thesis, I mentioned two typical forms of defeat, rebutting and undercutting. Let's consider a story about defeat to freshen up the distinction. Consider first a case of a *rebutting defeater*:

> After months of waiting, your heart jumps as you read that Umut University invited you for a job interview: 'Dear …. We are happy to inform you… The interview will begin at 2pm…We are looking forward to meeting you.'

After reading this email, you justifiably believe that the interview starts at 2pm (let this be the belief that p). However, you then receive the following email:

> Dear …. Unfortunately, there is a fire drill in the building from 1–4pm. Your interview has to be postponed to 5pm. Sorry for the inconvenience.

Reading the follow-up email rebuts your initial belief about the starting time by showing you that it is *false* that your interview will begin at 2pm. *Undercutting defeat* arises in the following case:

> After the postponement, you plan to be at Umut University at 5pm. But just after you received the postponement email, your trusted friend, who works in IT at Umut University, calls to chat about the latest news: 'You won't believe it', she says. 'Pranksters hacked our email server at Umut and it looks like they have sent out emails with false information earlier today!'

Practically speaking, you'd be well advised to check whether your interview really got postponed to 5pm after receiving your friend's call (let your friend's testimony be information D). Epistemically speaking, it seems that D at least *lowers* your justification for believing that you have an interview at 5pm even though D did *not* imply that your belief is false.[4] After all, you don't know whether the email you received was sent by the pranksters or by the university officials. Finally, consider a related case of undercutting defeat.

> Albert is an expert logician and has an excellent track record in completing logical proofs. His neurosurgeon friend Patricia, however, found a suspicious pattern in his brain that flares up on each of the rare occasions when he gets a proof wrong. Albert laboured through a proof (e) and arrived at a certain conclusion (p), while Patricia monitored his brain. After Albert finished the proof, confident he was correct, Patricia confided to him ($D_{Albert}$) that <the suspicious pattern that indicates that you get a proof wrong flared up while you worked on the proof>.

It seems as if D and $D_{Albert}$ give you and Albert, respectively, good epistemic reason to reduce confidence in your respective beliefs that p. Importantly, we are asking what you and Albert *ought* to do (epistemically) about your beliefs that p, not what you will in fact do. It seems as if you both ought to revise your beliefs even though your cases are importantly different. Your belief about the interview time is empirically justified, while Albert's belief is seen by some as a priori justified, that is, independently of experience or non-empirically. Nonetheless,

[4] The genealogy of a belief does not show that a belief is false in most cases. Exceptions are, for example, when you believe that your belief B has no cause and you are shown that B has a cause, or cases in which a belief stems from an obviously fraudulent source, as when you believe something out of the 'Book Full of False Claims'.

Albert's beliefs seem to be just as defeasible as your empirically justified beliefs. [5] This illustrates that non-empirically justified moral beliefs might also be undercut.

## Pollock's Objectivist Account of Defeat

The most sophisticated account of epistemic defeat is that of John Pollock, and my brief assessment of existing objectivist accounts of defeat naturally starts with his account. Pollock devised a system of rules that aim to unambiguously determine the conditions needed for defeat. Given any system of beliefs, and any new information, it can then be determined how the belief system should change once the new information is taken into account. Kvanvig (2007) compares Pollock's account to a 'front-door' approach to defeat: new information is assessed as it enters the cognitive system and whether it counts as a defeater is thus determined at the 'front door' of the cognitive system.

Pollock's first rule of epistemology is to trust the percepts. That is, 'p seems to be the case' is a good, albeit defeasible reason, for believing p. From the percepts, inductive arguments can be drawn to defeasible conclusions. Such arguments, however, are always accompanied by potential defeaters. On Pollock's front-door approach, a defeater is defined as follows:

> If P is a logical reason for S to believe that Q, then R is a defeater for this reason iff the conjunction (P&R) is not a logical reason for S to believe that Q. (Pollock 1974: 42; Pollock and Cruz 1999: 195)

Both non-belief mental states and beliefs can be defeaters, and defeaters are themselves reasons (ibid.). The following definitions make this clear:

> Rebutting Defeater: If M is a defeasible reason for S to believe Q, M* is a rebutting defeater for this reason iff
>
> a) M* is a defeater for M as a reason for S to believe Q
>
> b) M* is a reason for S to believe ~Q.

[5] Some doubt that a priori beliefs can be defeasible and take this as a reason to reject the a priori; see Kitcher (1980). See Casullo (2003) for a recent defence of the claim that a priori beliefs can be defeated.

> Undercutting Defeater: If M is a defeasible reason for S to believe Q, M* is an undermining defeater for this reason iff
>
> a) M* is a defeater for M as a reason for S to believe Q
>
> b) M* is a reason for S to doubt or deny that S would not be in state M unless Q were true. (Pollock and Cruz 1999: 196–7)

For a fuller appraisal of Pollock's front-door approach, we would need to understand more about what reasons are on his account. However, this question can be set aside, for the interesting question is how we should understand condition b) of Pollock's definition of undercutting defeat. What is it for S to doubt or deny that M would not occur unless Q were true? Pollock qualifies it as follows:

> S would not be in state M unless Q were true' can be read more simply as 'M does not guarantee Q. (Pollock and Cruz 1999: 197; Pollock and Gillies 2000: 75)

Pollock explicitly rejects reading 'Q guarantees M' as 'Q → M' where → is the material conditional. However, he also resists reading 'guaranteeing' as presupposing or requiring a causal relation between Q and M (ibid.). The latter qualification seems right, because most epistemologists today reject a causal constraint on justification and knowledge (Ichikawa and Steup 2017). It would, therefore, be unwise to make defeat dependent on such a constraint.

The problem is that Pollock's front-door approach requires crystal-clear rules about the mechanism of undercutting defeat, but whether or not undercutting defeat is instantiated is left on an intuitive level: it depends on how we cash out the notion of 'guaranteeing'. Undercutting defeat is, at its heart, not thoroughly understood. Some consequences of this problem beyond the debate about evolutionary challenges are the discussion about the higher-order requirement of defeat (that is, whether undercutting defeat requires a belief whose content is 'Q does not guarantee M'),[6] the relation of undercutting defeat and rebutting defeat,[7] and the computation of defeat in belief systems.[8]

[6] Cf. Sturgeon (2014); Melis (2014).
[7] E.g. Chandler (2013); Kotzen (2010); Pryor (2013); Thurow (2006).
[8] E.g. Lasonen-Aarnio (2010a).

There have been very recent attempts to precisify Pollock's account of undercutting defeat, and the crucial notion of 'guaranteeing' and considering them will show that none provides a good case for undercutting all objectivist moral beliefs. Pollock's account has been costed in *modal*, *probabilistic*, and *explanatory* terms. None of these terms implies that evolution undercuts non-empirically justified moral beliefs.

Lutz (forthcoming) discusses Pollock's account of defeat and suggests that the notion of 'guaranteeing' should be understood in modal terms. Indeed, Pollock (1987), writes that "[D] is an undercutting defeater for [E] as a prima facie reason for S to believe [P] iff [D] is a reason for denying that [E] would not be true unless [P] were true" (quoted in Lutz forthcoming; Pollock 1987: 485). Lutz takes this to support a counterfactual account of defeat, along the following lines: 'E would not be true unless p were true.' As Lutz and others have noted, however, if that interpretation were right, then necessary truths could not be undercut, because the counterfactual 'E would not be true unless p were true' would be trivially true if p is a necessary truth (Clarke-Doane 2015; Lutz forthcoming).[9]

Many philosophers argue that at least some moral truths are metaphysically necessary (an assumption that I will discuss in greater detail in chapter 7) and so evolutionary explanations of morality would not undercut moral beliefs. However, it should be noted that Lutz is too quick to reject Pollock's account of defeat because it is clear that Pollock is *not* committed to a counterfactual reading of the conditional (Pollock and Cruz 1999: 199; Pollock and Gillies 2000: 75). As we saw above, Pollock also suggests that evidence can support a belief only if the belief and the evidence are connected in the right kind of way, and he resisted the idea that this could be costed as a modal or causal relation. Although Pollock's more recent account does not give trivial results if applied to necessary truths, it is not very illuminating. *How* does evidence have to be connected to a belief and *what* must new information imply about the connection to be undercutting?

An alternative to the modal account proceeds in probabilistic terms (Kotzen 2010; Pryor 2013). For example, Kotzen suggests an account of undercutting defeat along the following lines: D undercuts the support that E provides for P iff Pr(P | E)

[9] This assumes a counterfactual interpretation of the conditional's truth value, and I discuss this in greater detail in the reader's guide of chapter 7.

> $Pr(P \mid E\&D) \geq Pr(P)$. A probabilistic account of defeat makes good sense in cases of empirically justified beliefs whose content is contingently true or false (though see Kotzen 2013). In the case of necessary truths, however, the account seems to fail. If probabilities are objective then the probabilistic account will also imply that necessary truths cannot be defeated, because they carry an objective probability of 1 (Baras 2017; Clarke-Doane 2015).

Of course, we could adopt a subjective account of probability, which would alleviate this problem. But my discussion below will show that we cannot accept a subjectivist account of defeat. Moreover, most Bayesians adopt a stance on probability that is between pure subjectivism and pure objectivism. But such a mixed account would tell us that a defeater is a defeater for all thinkers or only for those who assign a certain subjective probability to the belief in question. In the latter case, we have the problem of an implausible account of subjectivist defeat again (as I will show later). In the former case, we are thrown back to the problem that an objectivist account of defeat will make undercutting defeat of beliefs about necessary truths impossible.

Since modal and probabilistic proposals seem inadequate, Lutz (forthcoming) argues that the relevant sense of 'guaranteeing' should be understood as an *explanatory* connection. According to Lutz, evidence E 'guarantees' the truth of the belief that p insofar as p is entailed by the best explanation of E. So, an undercutting defeater of your belief that p based on evidence E is information that fully explains E *without* implying that p is true. This account seems correct in many cases. For example, we can cast the case about Umut University in these terms: you believed that your interview was shifted based on an email you assumed to be from the university. But when you received your friend's call, you had a full explanation of your evidence (that is, the receipt of an email) that did not imply the truth of your belief about the starting time of the interview. Lutz's explanatory connections view seems to give the right result: your belief about the starting time of your interview is undercut and you ought to give up your belief.

However, Lutz's account cannot explain how evolutionary explanations undercut non-empirically justified moral beliefs. Since I am interested in just these beliefs, Lutz's proposal does not help much to illuminate the case of the evolutionary defeat challenge. Moreover, since it is plausible that at least some

moral beliefs are non-empirically justified, we would be well advised to find an account of undercutting defeat for these beliefs too. Lutz relies on the idea that an explanation E might explain away some of the evidence for a belief B and thus *mitigate* support for B. In some cases, E might explain away *all* support for B. In other cases, E might explain away only *some* support for B, but leave intact other sources of support for B. Non-empirically justified moral beliefs will not, ex hypothesi, derive all evidential support from empirical facts. Their support can hardly be completely mitigated by offering an account of the causal origins of moral beliefs, apart from fantastical hypotheses (with little empirical support) about deceiving demons and the like. Of course, Lutz's account might explain why *some* justification of moral beliefs is lost, but it could not explain why all of it is lost. Even if this objection against Lutz's view fails (since there might be good reason to think that even direct perception (Audi 2013) or understanding (Hills 2010) of moral truths is mediated by psychological events), there is a deeper worry about his account.

The epistemic principle that motivates Lutz's account is very closely related to the empiricist principles that have most famously been defended by Harman (1975, 1977, 1986) and Quine (1980). Their accounts are 'empiricist' insofar as they aim to vindicate the idea that the best explanations of our observations determine the justification of our beliefs (and they are sceptical that there are any non-empirically justified beliefs whose content is not analytically true). This is not the place to discuss the merits of those views (though see Sturgeon 1986; 1992). The pertinent point is that *if* these views are correct, then, as Clarke-Doane (2015, 2016a) emphasised, there would be no reason to suppose that objectivist moral judgements *could* be justified in the first place. It is precisely the point of objectivists that *there are* non-empirically justified moral beliefs (to wit, justified beliefs whose truth is *not* implied by their best explanation). We can reject these views, as Harman and Quine urge us to, but that would require a discussion about whether objectivist moral beliefs *can be justified,* which should be asked prior to my question about whether objectivist moral beliefs *can be undercut, assuming that they are justified*. Thus, proponents of moral objectivism rely on the idea that some moral beliefs are non-empirically justified, and I assumed they were justified for the sake of argument. If Lutz is on the right track with this account of

undercutting defeat, then moral objectivist beliefs would not be defeated because these beliefs would not be justified to begin with. Adopting the view that we are only justified in believing what the best explanation of our observations imply should make us reject moral objectivism (as specified in the introduction) from the very start. We might end up rejecting moral objectivism, but this is a question beyond the scope of my thesis. Lutz's explanatory connections view does not show how evolution could undercut objectivist moral judgements.

In light of these problems for objectivist accounts of defeat, one might think that they can be avoided by adopting a subjectivist account of defeat, or what Kvanvig (2007) calls a 'back-door' approach to defeat. On a back-door approach, defeat is determined post hoc by whatever leaves the system of beliefs after new information entered the system. On this view, no particular view about the mechanisms of defeat is required. Instead, defeat is defined by the reaction of the belief system to the new information. Thus, a subject's belief is defeated if the subject *takes* a belief to be defeated. On a subjectivist view, defeat of non-empirically justified beliefs in necessary truths would be unproblematic: it would depend on whether the subject takes them to be justified. According to Kvanvig, Plantinga's account of defeat is such a back-door approach (Plantinga 2000). However, as I argue below, a back-door approach does not succeed. This finding puts pressure on the development of a front-door approach in sufficient detail. Before we come to the rejection of a subjectivist approach of defeat, however, let me point out in greater detail how the problem of defeat matters for the evolutionary challenge.

## Evolutionary Defeat: Often Suggested but Rarely Explained

In a number of recent discussions of the metaethical implications of evolutionary explanations of morality, philosophers either implicitly or explicitly appeal to the relevance of undercutting defeat. Many write that the evolutionary challenge 'undermines' moral beliefs or certain moral theories, thereby suggesting that the

problem has to do with undercutting defeat. Others argue explicitly that the evolutionary challenge instantiates a defeater of objectivist moral beliefs.[10]

However, these authors all assume that it is sufficient for evolutionary explanations of morality to defeat our moral beliefs to provide information that the moral beliefs are not, in one sense or another, 'reliably connected' to the truth. Variations of this theme abound. Let me mention just two examples. Silva (2016) argues that historical variability is evidence that the factors that influenced one's belief that p are "disconnected" from the truth about whether or not p (Silva 2016: 3). Braddock argues that the evolutionary challenge shows that moral judgements are not "likely to be true" and that the processes that produced moral beliefs are not "sufficiently reliability conferring" (Braddock 2016: 845–6).

Of course, I agree with these scholars that the evolutionary challenge is intimately connected to the phenomenon of undercutting defeat. In fact, my view is stronger in that I view the evolutionary challenge as an instance of undercutting defeat.

However, the currently available elaboration of the connection between undercutting defeat and the evolutionary challenge fall short precisely because it fails to explicate the *reason* that an undercutting defeater provides for giving up one's defeasibly justified belief. This is partly a problem of the best available account of undercutting defeat, as we have seen above, but also a problem in the discussion of the evolutionary challenge. The literature I reviewed in section 1.6 of chapter 1 purports to show that there is no epistemically viable way to make sense of an undercutting defeater: apparent problems with the coincidence or reliability of our moral beliefs, or the fact that they lack a connection with the truth, turn out to be chimaeras, or so critics of the evolutionary challenge argue.

Therefore, two questions about undercutting defeat must be addressed to make progress with the main question of my thesis.

First, what are the conditions for undercutting defeat? Is a front-door approach or a back-door approach to be preferred? If the former, then the current inability to spell out the precise conditions under which a given mental state fails to 'guarantee' the truth of a related belief is problematic. If the latter, then how

[10] Cf. Ballantyne (2013); DiPaolo and Simpson (2016); Leben (2013); Lutz (forthcoming); Nichols (2014).

does defeat relate to normative concepts in epistemology such as justification? I will address this question, and argue in favour of the back-door approach, in the next section.

Second, what is it about the connection between the causes of a belief and the truth of that belief that is supposed to be undercutting? This is the question that I will address in the subsequent chapters.

## Abstract[1]

I make the case for distinguishing clearly between subjective and objective accounts of undercutting defeat and for rejecting a hybrid view that takes both subjective and objective elements to be relevant for whether or not a belief is defeated. Subjectivists claim that taking a belief to be defeated is sufficient for the belief to be defeated; subjectivist idealists add that if an idealised agent takes a belief to be defeated, then the belief is defeated. A purely subjectivist view of defeat implausibly implies that justification comes cheap. Subjectivist idealism depends on conflicting intuitions and can be shown to yield inconsistent results in some cases. Both views should be rejected. We should be objectivists regarding undercutting defeat.

## 2.1 Introduction

Suppose you are at the bookshop and see Tom Grabit, whom you know to be a notorious thief, come flying out the door, rushing away with a stack of books barely hidden under his coat. As you walk off in astonishment, believing that Tom stole the books, you meet a trustworthy friend who tells you that Tom's identical twin brother is in town. What does your friend's testimony do to your belief that Tom stole the books? Arguably, your friend's testimony reduces or even nullifies your epistemic justification for believing that Tom stole the books. After all, taking your friend's testimony seriously means that you can no longer be sure whether Tom stole the books or whether his twin brother did.

The phenomenon that accounts for the loss of epistemic justification for your belief about Tom is known as undercutting defeat (Chisholm 1964; Hart 1948; Pollock 1970, 1995; Pollock and Cruz 1999). At the most general level, *defeat* describes a belief's ceasing to be epistemically appropriate (Bergmann 2006: 162). Pollock claims that undercutting defeaters are the "most important kinds of defeaters for understanding any complicated reasoning" (Pollock 1995: 85). Undercutting defeaters are usually distinguished from *rebutting defeaters*, which imply that a given belief is false and thereby give you reason to disbelieve the defeated belief (Pollock 1995: 85).[2] On the assumption that few, if any, of our beliefs

[1] This chapter is currently under review under the title 'A case for Objectivist Conditions for Defeat'.

[2] The focus of this chapter is *mental state defeaters*. Typically, mental state defeaters are beliefs, with propositional contents, but experiences can be defeaters too (Bergmann

are certain and most human reasoning is non-deductive rather than deductive, the concept of defeat, of both the rebutting and the undercutting kind, is essential in contemporary epistemology (Kvanvig 2007; Spohn 2012: 115).

However, there is a lacuna in the current understanding of undercutting defeat (and, by extension, our current understanding of fallibilism). The key question is when does new information undercut a belief (such that the believer lacks positive justification for maintaining the belief)?[3] We could adopt an objectivist or a subjectivist account of defeat.[4] Let's look at the subjectivist version. According to a subjectivist account of defeat, defeat is perspective-dependent, such that whenever you believe that new information E undercuts your belief B, your belief B *is undercut*. The subjectivist account of defeat is motivated by what I call the 'primacy of the subjective' – it takes seriously the subject's considerations about evidence, without regard for whether or not these considerations are correct. I will show that a pure subjectivist account is untenable in light of a recent discussion by Casullo (2016): on a pure subjectivist view, defeat and hence justification come too cheap. In response to this discussion, several philosophers have tried to take seriously the primacy of the subjective perspective and have also added to their account of undercutting defeat an idealised perspective-dependent condition for undercutting defeat. Having an idealised condition means that if an *idealised* version of you *would* believe that new information E undercuts your belief B, your belief B is undercut. I call such accounts 'subjectivist idealist' accounts of

2006; Pollock 1995). Mental-state-undercutting defeaters defeat the justification of beliefs or the power of reasons to confer justification on beliefs (Alston 1989: 238–9; Sudduth 2017). What follows will be independent of specifications about occurrent, aware, or accessible mental states. In contrast to mental state defeaters, *propositional defeaters* are true propositions that, if added to a subject's evidence base, would make some of the subject's beliefs unjustified (Klein 1971; Lehrer and Paxson 1969). The way that I use 'subjective' and 'objective' in this chapter is orthogonal to the distinction between internalist views of defeat, which take defeaters to be mental states, and externalist views, like those of Lehrer and Paxson, which take defeaters to be propositions.

[3] More precisely, when does new information learned by S undercut the justification of one of S's beliefs?

[4] *Objectivist* approaches to undercutting defeat argue that not every believed defeater is an actual defeater; see Casullo (2016), Melis (2016), Alston (2002), and Pollock (1995). I use 'objective' in the sense that E being a defeater for S, from the S's perspective, is not sufficient for E to be a defeater for S; compare Bergmann (2006: 112).

undercutting defeat.[5] Will a subjectivist idealist version work? No. My answer to the key question will be that a 'subjectivist idealist' account of undercutting defeat fails for two reasons. First, though one might be content with describing two different concepts of defeat (applicable in different contexts, perhaps), available subjectivist idealist accounts fail to distinguish both concepts. They attempt to elucidate *the* concept of undercutting defeat, but in doing so they depend on incompatible intuitions, as I will show. Second, in philosophical debate, we must rely on an objectivist notion of defeat to establish when new information defeats a position, not when a subject takes a defeater to be a defeater. The failure of the subjectivist idealist account means that we must be objectivist about undercutting defeat. Those who take the *objectivist approach to* undercutting defeat argue that not every believed defeater is an actual defeater (Alston 2002; Casullo 2016; Melis 2016; Pollock 1995). What the criteria for defeat are, however, is less frequently discussed.

For the sake of clarity, and because he defended the most elaborate version, I focus on Bergmann's (2006, 2009) account of subjectivist idealism. The problem that I address, however, applies to several other recent defenders of subjectivist idealism, such as Plantinga (1993, 1994, 2000) and Melis (2014). Section 2.2. introduces the subjectivist approach to undercutting defeat and Casullo's (2016) recent objections against it. I introduce Bergmann's representative account of the subjectivist idealist approach to undercutting defeat in section 2.3. and argue that it escapes Casullo's objections. Section 2.4. contains my argument against subjectivist idealism, and I conclude in section 2.5.

## 2.2 The Subjectivist Approach to Undercutting Defeat

### 2.2.1 Sturgeon and Melis on Defeat

Subjectivist accounts of defeat have recently been defended by Sturgeon (2014) and Melis (2014, 2016). Both are mostly interested in defending the view that undercutting defeat requires higher-order beliefs,[6] in the sense that you have to

[5] Cf. Bergmann (2006); Goldman (1986), Melis (2014, 2016), Plantinga (1993, 1994, 2000), Sturgeon (2014).

[6] For how to distinguish properly undercutting from rebutting defeaters, see Melis (2014: 436) and Sturgeon (2014: 117).

believe that information D defeats your belief that p for your belief that p to be undercut. In defending this view, however, they claim that having the higher-order belief that your belief that p is undercut is sufficient for your belief that p is undercut (Sturgeon 2014: 114). The defence of their view is exclusively based on examples such as the following:

> S is a normal person who, with her eyes closed, desires to know whether something red is before her. S has a firm presupposition [a higher-order belief] that it is not the case that it wouldn't look to her as if a red thing were before her unless a red thing were before her. She opens her eyes; it looks as if a red thing is before her, and she comes to believe that a red thing is before her. (Sturgeon 2014: 116 reworded for brevity)

Sturgeon writes that S "has [an] undercutting defeat[er] for her visual experience as a reason for her belief" (Sturgeon 2014: 116). Accordingly, merely *taking* a belief to be undercut, without justification, is sufficient for the belief to be undercut. Melis concludes from his view on higher-order requirements for undercutting defeat that "unreflective agents cannot suffer undermining defeat" (Melis 2014: 441). Insofar as unreflective agents cannot *take* a belief to be defeated, Melis's view lends support to a subjectivist approach to undercutting defeat because *taking* a belief to be defeated is a necessary requirement for defeat.

Sturgeon and Melis do not consider how their view of undercutting defeat would apply in philosophical practice, but the recent debate about evolutionary debunking arguments in metaethics provides a neat case study. Evolutionary explanations of morality are sometimes considered as instantianting undercutting defeat (e.g. Leben 2013; Lutz forthcoming). The goal of some debunking arguments is to show that people are not anymore epistemically justified in holding certain beliefs about objective moral facts (Joyce 2006, 2016c; Street 2006). If a subjectivist approach to undercutting defeat were correct, undercutting would most easily be achieved by getting people to *take* their beliefs about moral facts to be undermined. Do evolutionary considerations undercut some moral beliefs? That would depend on whether people *take* that to be the case. Clearly, however, whether people take those arguments to be convincing is not what's of interest. The interesting aspect is whether there are good reasons to give up one's beliefs (cf. Srinivasan 2015).

Apart from the practical point of view, are there good reasons against a subjectivist view of defeat?

### 2.2.2 Subjectivism Fails: Casullo's Objection

Casullo (2016) argues that a subjectivist account of undercutting defeat provides "implausible" results, given the assumption that undercutting defeat requires higher-order beliefs of the sort 'my belief that p is based on source S'. I will briefly recap Casullo's arguments. For present purposes, we can accept the higher-order view of defeat because, as we have seen above, subjectivists accept it too.

First, Casullo argues that subjectivism incurs three odd results in cases where subjects lack higher-order beliefs or where they have mistaken higher-order beliefs. In other words, Casullo's objection suggests that the subjectivist view has implausible consequences in cases of *ignorant* or *misinformed* subjects:

> C1a: The beliefs of unreflective subjects that do not form higher-order beliefs would be immune to undercutting. For example, if a child is unaware of the difference between the colour of a surface and lighting conditions then the child's belief that the wall is illuminated by a red light will not undercut his or her belief that the wall is red.
>
> C1b: The beliefs of misinformed subjects, those with mistaken higher-order beliefs, would be undercut by information E even if E has no bearing on their first-order beliefs. For example, if you have the higher-order belief that your philosophical beliefs are reliable because of the power of the Mountain Dew that you drink every morning, information E, according to which you drank fake Mountain Dew this morning, would undercut your philosophical beliefs.
>
> C1c: The beliefs of misinformed subjects, those with mistaken higher-order beliefs, could fail to be undercut if their mistaken higher-order beliefs make them ignore information that seems relevant to their first-order beliefs. For example, if you have the higher-order belief that your perceptual beliefs are reliable because they are based on drinking Mountain Dew in the morning, information E, which says that you are myopic, would *not* undercut your perceptual beliefs.

Second, subjectivism opens the door for epistemically *irresponsible* subjects to immunise themselves from defeat. Justification comes too cheap if the subjectivist view of defeat is true (Casullo 2016: 6 emphasis added):

> C2: If one has a justified belief that some source S is unreliable, one can reject any other belief B merely by forming the higher-order belief that B is based on source S. Conversely, one can insulate one's beliefs from undercutting defeaters to the effect that some source S is unreliable merely *by refraining* from believing that any of one's beliefs are based on source S. Given the implausibility of C1 and C2, we should reject the subjectivist view of undercutting defeat, or so Casullo argues.

Since Casullo puts much stress on the alleged counterexamples C1a–c, one might object to Casullo's criticism of subjectivism, which is based on the view that non-paradigmatic cases do not hurt subjectivism, insofar as subjectivism is doing a fine job of illuminating the concept of defeat in a wide range of more typical cases. Behind this objection is the thought that there is more than one concept of defeat, where different concepts of defeat might be applicable in different contexts. For instance, a subjectivist notion of defeat might be applicable in a court of law when the question is why the defendant, who is suffering from a psychosis, did not take seriously his victims' claims that he is *not* the devil.

However, it can be shown that these considerations against Casullo's view fail to be convincing. First, even though a purely subjective notion of defeat might sometimes be applicable, such as in a court of law, many contexts that allow for a subjective notion of defeat also deal in more *objective* terms. For example, it seems relevant and common to judge that a defendant should or should not (epistemically) have taken certain information into account, and to account for this normative appraisal we need more than a purely subjective account of defeat. Moreover, it seems that Casullo needs C1a–c only for illustrative purposes, while C2 would be sufficient for his criticism. A pure subjectivist notion of defeat has illegitimate consequences for what subjects are *justified in believing*. As Casullo has shown, a subjectivist notion of defeat opens the door to epistemically legislated ignorance of new information. This results in cases where thinkers gain new information that seems to affect their justification for holding certain beliefs, but

the subjective account is incapable of explaining why this is the case. This observation is sufficient to make a strong case against subjectivism.

In defending this point, I follow Casullo in assuming that there is a fact of the matter whether a thinker is epistemically justified in holding a certain belief (given the thinker's evidence, cognitive processes, context, aims, etc.). This is an uncontroversial assumption in epistemology, insofar as an important goal of contemporary epistemology is to establish just what this objective notion of justification is (Goldman 1986; Pollock and Cruz 1999; Sinnott-Armstrong 2006b).

So, Casullo's reliance on non-paradigmatic cases does not weaken his objection against subjectivism. There are good reasons, then, to reject an approach to undercutting defeat that is a pure subjectivist one.

Casullo's objections, however, fall short of indicating the need for an objectivist account of undercutting defeat. He attacks Sturgeon (2014), but asserts explicitly that his objections apply to Bergmann (2006) too (Casullo 2016: 4). In the next section, I take up this claim and show that it is incorrect, because Casullo's objections are efficacious against a *pure* subjectivist view only. Bergmann does not defend a pure subjectivist view. Let's consider whether Bergmann's subjectivist idealism saves the day.

## 2.3 The Subjectivist Idealist Approach to Undercutting Defeat

### 2.3.1 Bergmann's Account of Subjectivist Idealism

Bergmann defines defeat as follows: for any subject S, and any belief B of S, every believed defeater D of S of B is an actual defeater of B. For S to believe that D is a defeater of B is for S to take B to be epistemically inappropriate (Bergmann 2006: 163).

Second, there is a 'no-believed-defeater' condition for justification (Bergmann 2006: 163): S's belief that B is justified only if S does not take B to be epistemically inappropriate. Since Bergmann defines defeaters in terms of their power to make beliefs unjustified, and justification (partly) in terms of the absence of believed defeat, any believed defeater is an actual defeater (ibid.). To be precise, "it is not only *justified* believed defeaters that count as actual defeaters. *All believed defeaters* count as actual defeaters" (Bergmann 2006: 175 emphasis in original).

Bergmann's defence of his account of undercutting defeat relies on intuitions about cases. For example:

> [W]hat happens to the justificational status of your belief that you have hands once you become convinced that you are a brain-in-a-vat? Is that hand belief defeated? It was justified. Does it lose its justification? I think the answer is "yes" [because this would be] an appropriate response to the rest of your evidence, which includes not only your perceptual experience but also your belief that your perceptual experience cannot be relied on to indicate the truth about whether you have hands. And it seems intuitively that the belief that you have hands is *not* an appropriate response to that combined evidence. (Bergmann 2006: 165)

The important distinction from pure subjectivism is that Bergmann incorporates an idealist commitment that links justification to an external criterion, which, in Bergmann's case, is proper function.[7] Whether B is justified depends on the proper function of the cognitive system that generated B. So, assuming that a properly functioning human would *take* his belief that Tom stole the books to be undermined upon learning that Tom's twin brother is in town, any human who fails to make the connection between seeing what appears to be Tom and Tom's twin brother *should* take his or her belief about Tom stealing the books to be undercut (Bergmann 2006: 174).[8] I will call the defeater that an idealised subject in circumstances C would believe in a *called-for defeater*. For example:

> Jill agrees on a bet with her brother: she will win if her parents are not home. A few moments later Jill sees her parents coming home, but nonetheless continues to believe that she will win the bet. This is a case where we think Jill has an actual defeater for her belief that she will soon be receiving $300. And this is so even if Jill fails to put two and two together. For Jill *should* have a believed defeater for that belief. (Bergmann 2006: 170–1 reworded for brevity)

Cases like Jill's illustrate that "it is not only *believed* defeaters which are actual defeaters; there are also times when a person doesn't have a believed defeater but she *should* have one. In that case, she too has an actual defeater" (Bergmann 2006: 174). This completes Bergmann's account of subjectivist

[7] Bergmann (2006: 134) writes that "when we say cognitive faculties are functioning properly, the basic idea is that their functioning results in cognitively healthy doxastic response to the circumstances in which they are operating".

[8] See also Bergmann (2006: 118).

idealism. Setting aside differences about the concept of proper function, Plantinga (1993, 1994, 2000, 2002) shares Bergmann's commitment to the view that taking a belief to be undercut or having an epistemic reason to take a belief to be undercut is sufficient for the belief to be defeated. There are thus two kinds of undercutting defeaters for subjectivist idealists:

- Believed defeaters: information that the subject takes to be defeating.
- Called-for defeaters: information that the idealised [fill in externalist criteria of choice, e.g. properly functioning] subject takes to be defeating.

Both believed defeaters and called-for defeaters are actual defeaters.[9] That is, they reduce or nullify the subject's positive, adequate epistemic justification for holding the defeated belief. Let's turn to a criticism of subjectivism next.

### 2.3.2 A Reply to Casullo on Behalf of Subjectivist Idealists

We are now in a position to see that Casullo's objections do not extend to Bergmann's subjectivist idealism, contrary to Casullo's claim. C1a and C1c, the counterexamples that suggest that ignorant or misinformed subjects can fail to have a defeater when they should have one, are rebutted by Bergmann's idealist commitment. Ignorant or misinformed subjects may fail to have a *believed* defeater but still have a *called-for* defeater, depending on what their proper function demands. Since called-for defeaters are actual defeaters, however, subjectivist idealists avoid the implausible results that Casullo emphasises regarding pure subjectivism. One might retort that if an idealised subject still has a false higher-order belief, then counterexamples along the lines of C1c still apply. Whether this is possible depends on what the proper function is of the subject in question. Bergmann nowhere suggests that having false higher-order beliefs could be part of a proper function. I will assume that they are not; if that's the case, then C1c can be rejected. If not, then subjectivist idealism is still open to such a

[9] Though Bergmann does not assert this explicitly, it is evident in his discussion of the Jill case and in his reply to an objection by Fumerton (1988), which confronts reliabilists with the problem of accounting for defeat of justification in cases of misleading information (i.e. where D implies that the belief that B is mistaken, although B is in fact based on a perfectly reliable process).

counterexample. Pursuing the issue would require a fuller discussion of proper function, which I cannot provide here.

On the assumption that proper function does not allow systematically mistaken higher-order beliefs, we are left with C1b, the claim that it is implausible to think that defeat can come through misleading meta-beliefs. In essence, however, this is just a denial of the central thesis of subjectivist idealism according to which all believed defeaters are actual defeaters. Casullo cannot simply reject this claim without begging the question against subjectivist idealism, and he has not offered further reasons against the view other than appeals to intuitions about its alleged implausibility. As mentioned above, there is a clear sense of a subjectivist concept of defeat. Bergmann places much weight on it (Bergmann 2006: 163–8). Thus, Casullo's objections against subjectivism-idealism are not conclusive.

Moreover, C2 seems to gain significant support from a plausibly mistaken assumption about the nature of believing. The objection's strength is based on taking believing to be an action. It would seem bad indeed if subjects could *voluntarily* immunise themselves from defeat. But believing or refraining from believing are very probably not (voluntary) actions. But if believing and refraining from believing are not actions, then C2 collapses into a variant of C1: in some cases, a subject might seem to have a defeater but lack one, and vice versa. As we have seen, however, subjectivist idealists can deal with this objection and explain why a subject that seems to have a defeater does have one, according to their theory.

As Casullo's objections are mute against subjectivism enriched by an *idealist* component, it might seem as if subjectivist idealism is a good way of viewing undercutting defeat. However, this is not so, as the next section shows.

## 2.4 Against the Subjectivist Idealist Approach

### 2.4.1 The Subjectivist Idealist Approach is Unmotivated

Subjectivist idealism would vindicate the epistemic power of *believed defeaters* in reference to the primacy of subjectivity:

THE PRIMACY OF SUBJECTIVITY: the defeat status of S's mental states depends on what S takes their defeat status to be.[10]

The defeating power of *called-for defeaters* is defended in reference to an objectivist criterion (which is proper function, in Bergmann's case). Talking about two different kinds of defeaters is not a problem per se. As we have seen above, we can acknowledge that there might be a purely subjective kind of defeat, applicable in some contexts, such as in a court of law. However, insofar as we think that justification is not entirely perspective-dependent, as Bergmann does, we must give a non-subjectivist account of defeat in cases where it seems that the subject *should* have a defeater. Subjectivist idealism does not do this, as I will show in this section.

Consider that subject S will have a certain degree of epistemic justification, which might be zero, for any of her beliefs. Bergmann has to say that called-for defeaters affect S's epistemic justification, since that is what allows him to escape Casullo's objection. But believed defeaters also affect S's epistemic justification, since this follows from Bergmann's account of justification. What is the relation between both kinds of defeaters? Bergmann maintains that if *either* S has a believed undercutter for the belief B *or* S has a called-for undercutter for B then B is undercut for S.

I object that the explanation of why a believed defeater is epistemically relevant is different from why a called-for defeater is epistemically relevant. Both explanations cannot be reconciled because the primacy of subjectivity (which explains the defeating power of believed defeaters) implies that called-for defeaters are not actual defeaters, and the idealist criterion (which explains the defeating power of called-for defeaters) implies that believed defeaters are not actual defeaters.

To elaborate, consider that the primacy of subjectivity captures something intuitively appealing about the epistemic effects of holding some belief vividly. Try imagining that you very deeply believe that you are a brain-in-a-vat (BIV). Could

[10] One might object that accepting the primacy of subjectivity commits one to treating *justification* as a subjectivist phenomenon too; see Alston (2002). No such thing follows, however, as the primacy of subjectivity is properly restricted to defeaters and defeaters are importantly different to justifiers; see Bergmann (2006: 161ff).

you still rationally maintain that you have hands? Bergmann and other subjectivist idealists think that the answer is no (Bergmann 2005: 425).

You might feel a Peircean worry that it is quite impossible for humans to *really* believe a sceptical scenario as radical as the BIV scenario, and thus discount the relevance of the BIV example. But more mundane scenarios lead to similar results:

> Steve enthusiastically starts a new office job. All goes well in the first weeks and, on the basis of his positive self-image, Steve believes that chances are good that he will make it through the three-month probationary period (call this belief PP). However, four weeks into the job, Steve's partner of just two months ends their relationship. Steve sinks into a deep depression. No one, Steve believes, can bear his company for longer than two months (D). At the same time, nothing changes at work. His colleagues continue to be open and positive; his boss seems satisfied. In light of D, Steve wonders how he could ever believe PP. After all, who would want to hire such an awful person like himself? Steve, therefore, withholds belief in PP and, to be on the safe side, he starts applying for a new job.

Clearly, Steve's break-up does not tell us anything about his chances of keeping the job. Suppose that a properly functioning human (where proper function is, of course, a normative, not a descriptive term) would have kept both issues apart. Still, taking seriously Steve's perspective on the issue, there is a clear sense in which Steve's world view after his break-up makes his revoking of PP reasonable. The intuition is thus that given what Steve takes the world to be, he should withhold belief in PP. This intuition seems triggered by the primacy of subjectivity. It seems that Steve must feel a kind of internal pressure, an urge to disbelieve that he will make it through the probationary period.

The distinction between proper function and the internal perspective could be drawn using a distinction between internal and external rationality. According to Plantinga (2000: 110–2), internal rationality has to do with what goes on in belief formation "downstream from experience", whereas external rationality is broader in that it depends on what goes on in belief formation prior to experience. Bergmann (2009: 337) uses this idea to define a belief as "internally rational iff it

is an epistemically appropriate response to the subject's mental states" and "externally rational iff the believer's cognitive processing mechanisms are working as they epistemically should be in producing the belief". Steve's belief might be internally rational, but not externally rational.[11]

However, if we prize the primacy of subjectivity in deciding about believed defeat cases, then why should we think that subjects without a believed defeater, with a perfectly harmonious inner economy, nonetheless have a defeater? According to the subjectivist idealists' own grounds for defending the relevance of believed defeaters, the perfectly harmonious inner epistemic economy of a subject who fails to believe a called-for defeater (but does not notice that one exists), should take precedence. It would seem ad hoc for subjectivist idealists to uphold their commitment to idealism. Subjectivist idealism defers to the *actual* agent (such as Steve) when it comes to assessing the effects of held beliefs, but the component containing idealist views defers to the *properly functioning* agent (certainly not Steve) when it comes to assessing the effects of beliefs that are not held. But we want to know *why* some information is defeating, and subjectivist idealism gives differing explanations that make clear the defeating power of believed defeaters and called-for defeaters.

Bergmann might reply that the explanation for the defeating power of believed defeaters and called-for defeaters is the same: in each case, there is a violation of the requirement that beliefs be produced in accord with the *right kind of proper function* (Bergmann, personal communication). Bergmann's reply might seem promising. A believed defeater is an actual defeater *because* any design plan that might confer epistemic justification must require the believer not to hold B when the believer takes B to be defeated. A called-for defeater is an actual defeater for S's beliefs *because* it might be the proper function of S to take such information to be a defeater.

Bergmann's reply does not resolve the problem. Why believe that proper function can produce justified belief *only if* a design plan specifies proper function

[11] In a later publication, amending his account in Bergmann (2006), Bergmann writes that "justification is equivalent to internal rationality" (Bergmann 2009: fn 9). If that were so, however, then Bergmann's account would succumb to Casullo's objections, discussed in section 2.2.2.

such that S does not hold B if S takes B to be inappropriate? Bergmann's answer is that we should take seriously the primacy of the subjective. Here we look to internal rationality to know what defeat is. But the reason why called-for defeaters have defeating power on Bergmann's account is that this is what proper function might require, and here we end up citing *external rationality* as that which determines defeat. We still end up with fundamentally different explanations of the defeating power of believed defeaters and called-for defeaters.

### 2.4.2 The Subjectivist Idealist Approach is Unsubstantiated

The differing explanations of undercutting defeat offered by subjectivist idealism lead to a deeper problem, which causes Bergmann's subjectivist idealism either to yield conflicting verdicts about defeat or to succumb to Casullo's objections after all.

To begin with, note that defeaters can themselves be defeated by so-called defeater-defeaters (there are also defeater-defeater-defeaters, and so on). Every case of undercutting defeat by a believed defeater can, then, also be described as a case of undercutting defeat by way of a called-for defeater. Consider a briefer version of Steve's case (the time markers are for illustrative purposes only):

- ACTUAL: At t1, Steve believes PP. At t2, Steve believes, without evidence, that nobody can cope with him for longer than two months (D). The belief that D, B(D), defeats B(PP).

Now consider the same case where the actual Steve is compared with a hypothetical, properly functioning Steve:

- HYPOTHETICAL: At t1, Steve believes PP. At t2, Steve believes, without evidence, that nobody can cope with him for longer than two months, D. At t2, Steve*, Steve's properly functioning counterfactual self, believes D*, that the fact that nothing changed at work gives him no reason to believe that he will not make it through the probationary period. B(D*) defeats B(D), leaving the epistemic status of B(PP) unchanged.

Bergmann has to say that in HYPOTHETICAL, we take Steve's* beliefs to be decisive in deciding whether Steve has a defeater. Otherwise, we end up with the

objections raised by Casullo. This suggests that Steve's* beliefs are decisive after all. But this does not fit well with Bergmann's commitment to the primacy of the subjective. This seems to suggest that we are committed to an objectivist view of defeat after all.

To recap, subjectivist idealism faces a problem due to its bifurcation of the concept of undercutting defeat because, first, the differing explanation of the epistemic significance of defeat in the case of believed defeaters and called-for defeaters marks only an arbitrary distinction. Second, that arbitrary distinction cannot be substantiated, because it leaves subjectivist idealists unable to tell whether a belief is undermined or not. Subjectivist idealism does not explain how to overcome the bifurcation it introduces regarding undercutting defeat.

## 2.5 Concluding Remarks

Undercutting defeat plays an important role in epistemology: most if not all of our beliefs are only defeasibly justified and updating our beliefs often proceeds via undercutting and rebutting defeat. According to subjectivist idealism, both believed defeaters and called-for defeaters are actual defeaters. This chapter asked whether subjectivist idealism is a good way of viewing undercutting defeat.

The answer is no. Subjectivist idealism offers an attempt to reconcile a powerful intuition about the requirements of 'internal' rationality with an idealist component that allows it to answer some objections that affect purely subjectivist views. However, taking on board the idealist component leads to a bifurcation in the subjectivist idealist's concept of defeat and no clear way to reconcile the two concepts. At the very least, subjectivist idealists are really talking about two different concepts of undercutting defeat.

Therefore, subjectivist idealism's view of undercutting defeat should be rejected. When our beliefs tumble, there'd better be good epistemic reason for it.

# 3 Is Evolution Special?

## Reader's Guide

> Origin of man now proved.— Metaphysic must flourish.— He who understands baboon ~~will~~ would do more towards metaphysics than Locke.
> Darwin 1838: 84, 16 Aug 1838

Darwin's note evinces excitement about the power of evolutionary theory to make progress with philosophical questions. We have already encountered a similar sense of excitement in the quote of the Cambridge student in the main introduction, who enthused about the "boundless possibilities" offered by applying natural selection to ethics.[1] Today, in countless discussions of evolutionary explanations of morality, empirical input is taken as an invitation to roll up one's sleeves and get one's hands dirty, using a posteriori data to solve perennial problems. Within the genus of sceptical challenges in moral philosophy, the challenge posed by evolutionary explanations of morality is thought to be a special beast (Joyce 2006: 155; Locke 2014: 228; Schafer 2010: 475; Street 2006: 155).[2]

In this chapter, I will offer a nuanced reaction to this view. Empirical data plays a supporting role in the evolutionary defeat challenge, but it is ultimately inessential. On the other hand, I will underscore my earlier point that the challenge is still importantly, and interestingly, different from the challenge of radical scepticism. Coming from a discussion of defeat simpliciter in the previous chapter, this chapter turns to an an assessment of the sources of defeat in the case

[1] As Tersman writes in a related context, "philosophical debates seldom can be adjudicated with reference to hard- and well-established empirical facts. So when an opportunity appears to arise, it is difficult to resist the temptation" (Tersman 2006: 21). One strong factor might be the general attraction of a naturalistic methodology, heeding the sound advice of Darwall, Gibbard, and Railton, who lament that "too many moral philosophers … have been content to invent their psychology or anthropology from scratch" (Darwall et al. 1992: 34–5).

[2] And many discussions of the challenge seem to suppose that the challenge is interesting only if it can be shown to be special and different from other sceptical challenges. See Vavova (2014a: 85), who writes that "an empirical claim of some sort is essential – this is the distinctive feature of such arguments".

of the evolutionary defeat challenge. The chapter contributes to my main research question by showing what *kind of information* would give rise to a defeater when we learn about evolutionary explanations of morality. If consensus is right, then the type of information that gives rise to a defeater is empirical and, given the robustness of evolutionary theory, the defeater would be a difficult one to resist. Moreover, insofar as *a posteriori evidence* is taken to imply something about the justification of moral beliefs, some might be tempted to think that we can circumvent the need to look for an objectivist account of defeat and simply settle the case against moral objectivism on empirical grounds. As we will see, this hope is unfounded.

That is because there is a clear sense in which evolutionary information as a kind of a posteriori information is irrelevant to the evolutionary defeat challenge. Proponents of the evolutionary challenge are under no illusions about the contingent relevance of evolutionary claims for the kind of challenge that they want to raise. Joyce, for example, writes:

> The evolutionary perspective is, strictly, dispensable. Were we to explain our moral beliefs by reference to, say, developmental and socialisation processes, then, so long as these processes similarly nowhere imply or presuppose that our or anyone else's moral judgements are true, the same epistemological conclusion could be drawn. (Joyce 2006: 185–6; compare Joyce 2016d: 125)

Similarly, Street notes that the evolutionary claim is dispensable from her anti-objectivist argument:

> In principle, […] an analogous dilemma could be construed using any kind of causal influence on the content of our evaluative judgments. For the argument to work, two conditions must hold. First, the causal influence in question must be extensive enough to yield a sceptical conclusion if the realist goes the route of viewing those causes as distorting. Second, it must be possible to defeat whatever version of the tracking account is put forward with a scientifically better explanation. The fact that there are *any* good scientific explanations of our evaluative judgments is a problem for the realist about value. (Street 2006: 155)

As a matter of fact, the evolutionary challenge relies on evolutionary explanations of our moral beliefs because they provide, at least this is the assumption, the most explanatorily powerful genealogy of our moral beliefs. In particular, evolutionary explanations of morality play a special role amongst the

many possible causal explanations of our moral beliefs because they provide an *ultimate*[3] explanation of our moral judgements that we would otherwise lack. Without an evolutionary explanation, we would thus not have a *full* explanation of our moral beliefs.

In principle, however, many different scientific disciplines and sources of information about the genealogy of morals might be relevant. For example, John Doris suggests the following:

> A [moral] value is a child with many parents, and establishing a value's pedigree may involve, inter alia, biological considerations (such as those drawn from evolutionary theory), social considerations (such as those drawn from economic thought), and psychological considerations (such as those drawn from cognitive science), as much as historical considerations. (Doris 2009: 705–6)

Thus, in principle, it does not matter much whether the information we gain about the causal origins of our beliefs stems from descriptive evolutionary ethics or any other field of inquiry. Any equally explanatorily powerful thesis about the origins of our moral beliefs could be used to raise an analogous challenge. For example, if the best explanation of the content of our moral beliefs would be in terms of cultural influences, then proponents of the evolutionary challenge might instead raise an analogous 'cultural' challenge that also aims at undercutting the justification of all objectivist moral beliefs. Recall that some philosophers doubt that the content of moral rules, or the evolution of a moral as opposed to a more general normative capacity, can be seen as an adaptation (e.g. Prinz 2007: ch. 12, 2009). If a non-nativist story turns out to be right, like that suggested by Prinz, who takes "every moral value as a *cultural* artefact with a history just waiting to be discovered" (Prinz 2007: 217), then a challenge that is analogous to the evolutionary challenge might work nonetheless, although it would be based on different empirical claims. The requirement would just be that the causal mechanisms that produced our moral beliefs do not imply that our moral beliefs are true.

[3] Recall that an ultimate explanation shows a trait's biological function in reference to natural selection (or related evolutionary forces), whereas a proximate explanation shows a trait's biological function in reference to immediate psychological or environmental factors.

At the same time, proponents of the evolutionary challenge seem to think that *some* sort of a posteriori information about our moral beliefs is essential for their challenges to create sceptical conclusions. As I will show later, this is not the case. The problem of defeat of non-empirically justified beliefs can arise on purely a priori grounds. Before turning to the argument, however, I want to show why the evolutionary defeat challenge, though a priori at heart, is nonetheless different from the challenge of radical sceptics.

Proponents of the evolutionary challenge are keen to distinguish their challenge from what they call 'general scepticism' or 'standard philosophical scepticism'. Joyce argues that the evolutionary challenge is *not* "unimpressively analogous to standard challenges from the philosophical sceptic" because the sceptical hypothesis in the moral case is "empirically confirmed" (Joyce 2006: 187). He writes:

> It is not just that [with the evolutionary challenge] we can *make up* a consistent hypothesis according to which a bunch of our ordinary beliefs are false; rather it is that we might have empirical evidence supporting the hypothesis that explains how [our moral beliefs] came about but does not require that they be true. (Joyce 2006: 187)

Similarly, Street maintains that her "Darwinian Dilemma is not a routine, general sceptical worry deployed selectively" (Street 2006: 116). She suggests that this is because an empirically well-established claim is at the heart of her evolutionary challenge (Street 2006: 155). When Street and Joyce mention 'general philosophical scepticism' they seem to think of arguments that (a) target the justification of *all* perceptual beliefs, as opposed to only all moral beliefs, and (b) are based on sceptical hypotheses that are not empirically confirmed.

In the main introduction, I already addressed briefly the distinction between radical scepticism and the evolutionary defeater challenge. Drawing that distinction is crucial for two reasons. First, it shows that the prospects of the evolutionary challenge are independent of the prospects of radical scepticism. Even if radical scepticism could be rejected, the evolutionary challenge might yet succeed.[4] Second, raising the evolutionary challenge would be a particular problem for objectivist moral beliefs, but not for other types of beliefs, such as beliefs based

[4] Cf. Shafer-Landau (2003: 239-4); Huemer (2005: 12).

on perception. This is a desideratum for proponents of the evolutionary defeat challenge to moral objectivity who are not sceptical about other domains.[5]

Here is a rough-and-ready formulation of the radical scepticism for epistemic justification. (1) for subject S to be justified in believing that p, where p is a non-tautologous proposition about the external world, S must be epistemically justified in believing that S is not in a sceptical scenario. (2) nobody can justifiably believe that he or she is not in a sceptical scenario.[6] Sceptical scenarios appear to be exactly like our world and yet it is conceivable that everything we currently believe about contingent facts is false in such sceptical scenarios.

It seems safe to say that resistance against radical scepticism banks on the claim that there are good epistemic reasons that somehow rule out the epistemic relevance of sceptical scenarios and thus (2) should be rejected (cf. Pritchard 2002).[7] The evolutionary challenge avoids this alleged weakness of the radical sceptical argument because the evolutionary challenge is supposed to provide us with a kind of empirically confirmed 'sceptical scenario'. Whatever reason we have to reject the radical sceptical scenario, if there are any, proponents of the evolutionary challenge claim that they do not apply in the case of the evolutionary challenge.

---

[5] A second line of comparison with greedy scepticism that is suggested by Street (2006: 155), and emphasised by several commentators (Shafer-Landau 2012; Vavova 2015), is that the evolutionary challenge is a *targeted* sceptical challenge that is supposed to apply only to the moral domain, as opposed to a *general* sceptical challenge that applies across many different domains.

[6] Cartesian scepticism, or 'indiscernibility-based' scepticism in more general terms, is but one subcategory of greedy arguments. The other is regress-based scepticism. Regress-based scepticism aims at the same conclusion as indiscernibility-based scepticism and relies on the premises, which are, roughly, that to justifiably believe any proposition p, one must be able to infer p from another of one's justified beliefs and that no belief, or relation between a number of beliefs, can eventually, ultimately justify belief in p. See Sinnott-Armstrong (2006b) for a fuller treatment of regress-based scepticism about moral beliefs.

[7] For example, some have tried to reject (2) by pointing out that we are justified in holding beliefs about the external world that are based on perception, such as the belief that we have hands, that we therefore have knowledge of the external world, and that we can therefore know that we are not in a sceptical scenario (Moore 1939; Pryor 2000). Whether and how this line of response succeeds is a controversial issue; it succeeds by denying that a Cartesian sceptical scenario counts as a legitimate epistemic worry because we have a default entitlement to believe that we are not in a Cartesian sceptical scenario. The core question in this context is whether we are allowed to assume the truth of beliefs of type T in establishing the truth of other beliefs of type T. Some such entitlement is required for all responses to the Cartesian challenge (as well as the regress challenge).

Thus, reliance on empirical evidence seems to play a special role in distinguishing the evolutionary challenge from radical sceptical challenges. As we have seen, that distinction is important, because there are good reasons to make the prospects of the evolutionary challenge independent of the prospects of the radical sceptical challenge.

The appearance of the crucial relevance of empirical input in the evolutionary defeat challenge will turn out to be unfounded. Nonetheless, as I have shown in the main introduction, the evolutionary defeat challenge does not thereby reduce to the radical sceptical challenge because we start by assuming the defeasible justification of moral beliefs. It is now time to show why a posteriori information will play a negligible role in the evolutionary defeat challenge. In what follows, I discuss Street's conditional undercutting argument. I focus on Street for ease of exposition and because she puts forward the most discussed evolutionary undercutting argument. I will focus on the conditions under which a posteriori information can undercut moral beliefs, and since this is a general question about undercutting, my points apply in equal force to Joyce's account of evolutionary undercutting, though I do not discuss Joyce's account explicitly in the next section. Since Street explicitly sets up her argument against robust moral realism, I follow her and tailor my discussion, in the next section, to robust moral realism too. Robust moral realism is a form of moral objectivism that contains a particular metaphysical view about the existence and nature of moral facts. As I pointed out in the main introduction, moral objectivism is more encompassing than robust moral realism in that it allows for 'relaxed' views about moral metaphysics, like that of Scanlon (2014) or Parfit (2011a). Hence, it should be kept in mind that my argument in the next section will apply, by extension, to moral objectivism more generally.

## Abstract[1]

This chapter aims at showing that no evolutionary, causal explanations play an essential role in reaching the epistemological conclusion that arises on one horn of Sharon Street's evolutionary argument against moral realism, the metaethical view that there are non-natural and mind-independent moral properties and facts that we can know about. I aim to show that Street's argument depends on the Benacerraf-Field challenge, which is the challenge that explains the reliability of our moral beliefs about causally inert moral properties or entities. The Benacerraf-Field challenge applied to metaethics relies on metaphysically necessary facts about realist moral properties rather than on contingent Darwinian facts about the origin of our moral beliefs. Attempting to include an essential causal empirical premise yet avoiding recourse to the Benacerraf-Field problem yields an argument that is either self-defeating or of limited scope. Ultimately, evolutionary, causal explanations of our moral beliefs and their consequences do not present the strongest case against robust moral realism. Rather, the question is whether knowledge of causally inert, mind-independent properties is plausible at all.

## 3.1 Introduction

Much current metaethical research focuses on the consequences of the assumption that evolutionary forces shaped our moral beliefs at least to some extent. The human capacity to produce moral judgements and, to some extent, the content of our moral judgements and beliefs may be the products of natural selection (Joyce 2006: ch. 4; Street 2006: 115–21).[2] While there are considerable difficulties in establishing this claim, there are also persuasive arguments in its favour (Buchanan and Powell 2015; Fraser 2014; Kitcher 2011).

[1] This chapter is based on a paper published as 'Old Wine in New Bottles. Evolutionary Debunking Arguments and the Benacerraf-Field Challenge' in *Ethical Theory and Moral Practice* (2017) 20: 781-795.

[2] The distinction between the *capacity* to make moral judgements and their *content* is crucial for evolutionary debunking arguments in metaethics, and it is controversial whether the empirical claim about evolutionary influences on the content of our moral beliefs is well supported (Buchanan and Powell 2016; FitzPatrick 2014a; Mogensen 2016b). I assume it here for the sake of the present discussion.

Some proponents of evolutionary debunking arguments (EDAs)[3] purport to show that robust moral realism[4] (the view that there are non-natural and mind-independent moral properties and facts about which we can have knowledge) is implausible, given the evolutionary influence on our moral cognition, and so should be rejected (Ruse and Wilson 2006; Street 2006).[5] These debunkers reach epistemological conclusions by taking into account *Darwinian considerations* about the origins of our moral beliefs. The grand ambition of these debunkers is to use a well-established empirical account of the origins of our moral beliefs to discredit moral realism. The argumentative strategy is intriguing: evolutionary theory is well supported, and it would be an immense cost to any metaethical theory if it could not incorporate it. Sharon Street puts forward the most influential evolution-based critique.

Street reaches the sceptical epistemological conclusion that we have sufficient reason to doubt the truth of *all* of our moral beliefs, on the assumption that moral realism is true, and suggests that Darwinian considerations do important work in the argument (Street 2006: 109) and, eventually, "settle the [realism vs. anti-realism] debate in favour of the anti-realist view" (Street 2008a: 214). She argues that if robust moral realism were true, then moral knowledge is unlikely because mind-independent moral properties were evolutionarily irrelevant and thus beliefs about these properties were not selected for.

An ever-increasing number of philosophers are concerned with the metaethical implications of evolutionary theory, suggesting that much in metaethics depends on one or the other explanation of the causal origins of our moral beliefs (Artiga 2015; Deem 2016; Fraser 2014; Ruse and Richards 2017).

---

[3] I focus on Street's *global* EDA that intends to lower the status of *all* moral judgements Kahane (2011).

[4] Henceforth, 'realism' designates robust moral realism. 'Moral properties' entail 'moral facts'. Exemplary proponents of this are Enoch (2011b), Shafer-Landau (2003), and Wielenberg (2014). Street means to include moral naturalism as a target of the EDA too, and while most see moral naturalism as being well placed to answer the evolutionary challenge (Enoch 2010: 422), others have recently called this into question (Barkhausen 2016; Bogardus 2016).

[5] Not all areas of discourse are threatened by evolutionary considerations. For example, the truth of our beliefs about ordinary objects is consistent with their purported evolutionary background; cf. Boudry and Vlerick (2014); Wilkins and Griffiths (2013).

Street's argument against robust moral realism depends on a version of the Benacerraf-Field challenge, according to which moral realists must show that it is in principle possible to explain how we can have reliable beliefs about the moral properties postulated by moral realism. The Benacerraf-Field challenge does *not* rely on a premise about human evolution, or any alternative empirical explanation of the origins of our moral beliefs. Hence, contrary to the received view, no evolutionary, causal explanation plays an essential role in reaching the debunking argument's epistemological conclusion.

But attempts to include an essential empirical premise about the origins of our moral beliefs, while not relying on a version of the Benacerraf-Field challenge, do not succeed: the EDA would be self-defeating if applied to *all* moral beliefs because it would require a substantive moral claim which would itself be called into question by the EDA, or it would debunk only some of our moral beliefs, but not all of them. Hence, Street's *evolutionary, causal* debunking argument against robust moral realism fails.

This should alter the locus of the metaethical debunking debate. The crucial point is not what follows from evolutionary theory or alternative causal explanations of our moral beliefs, but whether knowledge of causally inert, mind-independent, and irreducibly normative properties is possible at all.

I proceed as follows. Section 3.2 introduces Street's EDA and presents a possible escape route for robust realists, which highlights a crucial premise of the EDA. Section 3.3 shows that this crucial premise is conceptual rather than empirical and that the argument depends on the Benacerraf-Field challenge. Section 3.4. argues that attempts to reintroduce an essential empirical premise fail. I conclude in section 3.5.

## 3.2 Street's Evolutionary Debunking Argument

### 3.2.1 Reconstructing Street's Argument

The first premise of Street's argument is the supposition, for the sake of argument, that robust moral realism is true (Street 2006: 109, 121):

REALISM: Robust moral realism is true.

Evolutionary debunking arguments generally rest on claims about our moral beliefs, typically about their origins. The factual premises of the argument – claims about our moral beliefs – provide the grounds for the 'debunking premise', which states the reason(s) for viewing with suspicion the facts about our moral beliefs that were established by the factual premises (Kahane 2011). Thus, the factual premises need to be combined with a debunking premise to yield a conclusion.

What are Street's 'factual' premises? Almost all discussions of her argument (Artiga 2015; Copp 2008; Vavova 2015) follow Street in interpreting her first factual premise as stating that natural selection, and other evolutionary factors, have had a 'tremendous', albeit indirect, influence on the content of human evaluative attitudes (Street 2006: 113). However, a careful reading shows that Street is, somewhat in tension with the tenor of her piece, not committed exclusively to evolutionary factors: she also mentions that 'other causal influences can shape our evaluative judgements' (Street 2006: 120), and as long as there is 'some sort of causal explanation' (Street 2006: 153) of our evaluative beliefs, 'whether Darwinian or otherwise' (Street 2006: 155), this fact seems sufficient 'for the purposes of the argument' (Street 2006: 158). Hence, Street's challenge is best understood if her first factual premise is construed as making a broad claim about the *causal* origins of our moral beliefs. Let's call this premise INFLUENCE:[6]

INFLUENCE: All our moral beliefs were influenced by causal forces.

Street argues that the causal forces that shaped our moral beliefs are not systematically connected to moral truths and argues that the realist must 'take a position on what relation there is, if any, between the selective forces that have influenced the content of our evaluative judgements […] and the independent evaluative truths that realism posits' (Street 2006: 121). One option is to assume that there is no connection between the causal forces that influenced our basic evaluative dispositions and the moral truth. The other option is to affirm that there is such a connection. The latter option is rejected by most realists (FitzPatrick 2014b: 241).[7] No realist that I know of endorses the view that there is a relation

[6] Vavova (2015) uses the same term but refers to evolutionary forces exclusively.

[7] Assuming that there is a connection between evolutionary forces and moral truth confronts realists with an inference for the best explanation (IBE): the content of our moral

between evolutionary forces and the moral truths. Hence, the second factual, empirical premise in Street's argument is about the independence between the causal influences on our moral beliefs and the moral truth. Let's call this premise MISLEAD:

> MISLEAD: The causal forces that influenced our moral beliefs have no systematic relation to the moral truth.

Street argues that REALISM, INFLUENCE, and MISLEAD lead to a sceptical conclusion (Street 2006: 122). We can fill in the required DEBUNKING CONDITIONAL:

> DEBUNKING CONDITIONAL:[8] If REALISM, INFLUENCE, and MISLEAD are true, then we have sufficient reason to doubt the truth of all our moral beliefs.
>
> DOUBT (conclusion):[9] So, we have sufficient reason to doubt the truth of all our moral beliefs (REALISM, INFLUENCE, MISLEAD, and DEBUNKING CONDITIONAL).

The independence of evaluative truths and the causal origins of our moral beliefs bode trouble for realism. Intriguingly, the argument seems to have sweeping metaethical implications that are based on empirical evidence. If debunkers could thereby challenge robust moral realism, they would change our

---

beliefs can be explained without invoking *moral* properties. Qua parsimony, the evolutionary explanation is better than the realist's explanation. I do not consider the 'EDA as IBE' interpretation here since it would leave much room for the realists to reply. They might claim, for instance, that there are further, prudential reasons to stick with realism; see Enoch (2011b); Copp (2008: 190).

[8] The epistemic principle behind the DEBUNKING CONDITIONAL is controversial and a crucial point of the debunking debate Bedke (2014); Bogardus (2016); Vavova (2015). I assess two interpretations in section 3.3 to support my point about the irrelevance of genealogical claims, but I do not address the debate about the correct *epistemic* principle in this chapter.

[9] Street's argument extends beyond DOUBT: she concludes that robust realism ought to be rejected Street (2006: 135). I am sympathetic to Vavova's interpretation of the argument as the "beginning of a *reductio* of realism" Vavova (2015: 108). This makes sense if we make explicit that REALISM entails the possibility of moral knowledge, and an auxiliary premise that states that a metaethical theory ought to be rejected if it entails that moral knowledge is possible and at the same time gives us reason to doubt the truth of all our moral beliefs. My argument focuses on the steps that lead to DOUBT, so I need not make the additional steps towards the rejection of realism explicit.

concept of "man's position in the universe" through reference to solid empirical facts – truly in "Darwin's spirit" (Mayr 2003: xxi).

### 3.2.2 An Escape Route for Realists

However, thus far the argument does not secure a sceptical conclusion as it ignores an easily overlooked complication. Since we assume REALISM, moral properties do exist, and *at least some* of our moral beliefs could, in principle, reliably track them (FitzPatrick 2015; Huemer 2016; Wielenberg 2014). This would be enough to reject the claim that empirical evidence about the origins of our beliefs gives us sufficient reason to doubt the truth of *all* of our moral beliefs. Thus, realists might grant REALISM, INFLUENCE, and MISLEAD but reject the DEBUNKING CONDITIONAL. Realists might argue as follows:

> CORRECT: At least some of our moral beliefs are likely to track realist moral properties.[10]
>
> ANTI-DEBUNKING CONDITIONAL: If REALISM and CORRECT are true then we do *not* have sufficient reason to doubt the truth of all our moral beliefs.
>
> TRUST (conclusion): So, we do *not* have sufficient reason to doubt the truth of all our moral beliefs (REALISM, CORRECT, and ANTI-DEBUNKING CONDITIONAL).

We can infer TRUST from CORRECT and the ANTI-DEBUNKING CONDITIONAL since CORRECT gives us reason to believe that at least some of our moral beliefs are non-accidentally connected to the moral properties. The DEBUNKING CONDITIONAL and the ANTI-DEBUNKING CONDITIONAL cannot both be true, but which premise realists have to accept depends on whether INFLUENCE and MISLEAD rule out CORRECT.

As mentioned above, the realists this chapter is concerned with are committed to a non-naturalistic conception of mind-independent, causally inert moral properties (Enoch 2011b: 7, 159; Shafer-Landau 2003: 107).[11] This suggests that

[10] The notion of 'tracking' as I use it in this chapter should be broadly understood as capturing any non-accidental, systematic connection between our moral beliefs and realist moral properties.

[11] There are realists who claim that moral properties are causally efficacious, for example Oddie (2009) and the Cornell realists.

their claim to CORRECT is unaffected by INFLUENCE and MISLEAD because realists' accounts of how moral believers track moral properties do not rely on causal relations between moral properties and moral beliefs in the first place. Whether we believe that stealing is wrong, for instance, because of some Darwinian force, or because of our upbringing, or because of some other causal factor, realists can hold on to the claim that the belief is non-accidentally true because debunkers have not ruled out realists' non-causal accounts of reliable access to its truth.[12] How do realists defend the claim that our moral beliefs are likely to track realist moral properties? They might claim, for instance, that moral beliefs are reliably formed through rational intuition or direct perception, or that there is a constitutive relation between moral properties and moral beliefs, or that divine revelation plays a role in shaping our moral beliefs (Bengson 2015, 2015; Bogardus 2016: 642f; Cuneo and Shafer-Landau 2014; Huemer 2005: 4–6).[13]

Hence, debunkers have to consider CORRECT as a live option, and realists would probably insist that, in an argument that begins with REALISM, debunkers first have to show that the realist's 'escape route' via CORRECT fails.

Debunkers have two options. They can reject CORRECT, which requires showing that none of our moral beliefs is likely to track moral properties. If successful, they could infer DOUBT. Alternatively, as we will see, they may argue that the DEBUNKING CONDITIONAL, and the inference to DOUBT, are valid despite CORRECT.

## 3.3 The A Priori Base of Evolutionary Defeat

### 3.3.1 The Benacerraf-Field Challenge

Note that debunkers cannot just assume without argument that CORRECT is false because that would beg the question in the argument against realism. As indicated

[12] Vavova (2014a) suggests that the empirical premises alone suffice to rule out CORRECT, if they provide evidence of error. I show that this route fails in section 3.4.

[13] I do not address the merits of possible realist replies in this chapter. My sole concern here is to show that the debunking argument depends on the Benacerraf-Field challenge. I am sceptical about the ultimate viability of the mentioned realist replies, but my point here is simply that they are live options and not ruled out by the empirical premises of the argument, which is why debunkers rely on the Benacerraf-Field challenge to counter these claims.

above, their empirical premises do not directly refute CORRECT either. Therefore, debunkers need a different argument to show that CORRECT fails; such an argument depends on the Benacerraf-Field challenge.

The Benacerraf-Field challenge is a problem for knowledge of mind-independent, causally inert entities. The worry originates in Benacerraf's work on the possibility of mathematical knowledge (Benacerraf 1973). He writes:

> I think, that something must be said to bridge the chasm, created by … [a] realistic and platonistic interpretation of mathematical propositions, between the entities that form the subject matter of mathematics and the human knower. (Benacerraf 1973: 675)

The abstract entities postulated by mathematical Platonism share two important features with moral properties: mind-independence and causal inertness. Therefore, as Peacocke recognises, Benacerraf's problem concerning mathematical Platonism seems to be a problem for moral realism too:

> What Benacerraf ... asserts about mathematical truth applies to any subject matter. The concept of truth, as it is explicated for any given subject matter, must fit into an overall account of knowledge in a way that makes it intelligible how we have the knowledge in that domain that we do have. (Peacocke 1999: 1–2)

Benacerraf's worry presupposes a causal theory of knowledge, but Hartry Field's development of the challenge makes it independent of this precondition. Thus, it also applies to robust moral realists, who commonly reject a causal theory of knowledge. Field challenges realists to explain

> how our beliefs about [abstract] entities can so well reflect the facts about them ... [I]f it appears in principle impossible to explain this, then that tends to undermine the belief in mathematical entities, despite whatever reason we might have for believing in them. (Field 1989: 26)

Field's adapted challenge "depends on the idea that we should view with suspicion any claim to know facts about a certain domain if we believe it impossible in principle to explain the reliability of our beliefs about that domain" (Field 1989: 232–3). The Benacerraf-Field challenge raises suspicion about the reliability of beliefs about causally inert and mind-independent properties. This is problematic because it is, other things being equal, to a theory's costs if it treats knowledge about properties presupposed by the theory as merely accidental or altogether

inexplicable. We can now see that the Benacerraf-Field challenge arises because of the metaphysical properties of abstract entities and that moral properties are relevantly similar because they are also mind-independent and causally inert. If there is no way in which robust realists can explain how our moral beliefs are likely to track realist moral properties (i.e. give an argument for CORRECT), then moral knowledge would indeed be a startling fact.

### 3.3.2 Benacerraf-Field 2, Darwin 0

The Benacerraf-Field challenge affords a way of arguing that it is likely that none of our moral beliefs can reliably track mind-independent and causally inert moral properties (Clarke-Doane 2017c). In other words, Benacerraf and Field allege that the nature of moral properties, as conceived of by robust moral realism, makes it likely that CORRECT is false. If their allegation is correct, then we have reason to accept the DEBUNKING CONDITIONAL and not the ANTI-DEBUNKING CONDITIONAL.[14]

Debunkers rely on the Benacerraf-Field challenge to reject CORRECT and to force realists to accept the DEBUNKING CONDITIONAL. This is because the empirical claims of the debunking argument leave open the possibility of non-causal, truth-tracking determinants of our moral beliefs and so CORRECT is not falsified (cf. Bogardus 2016). Conversely, if realists could vindicate CORRECT, we would have reason to accept the claim that we do *not* have sufficient reason to doubt the truth of all our moral beliefs. Debunkers have to challenge realists, in the spirit of Benacerraf and Field, to explain the reliability of our moral beliefs to get their sceptical argument off the ground.

Importantly, the Benacerraf-Field challenge is a conceptual, normative challenge. It is conceptual, as opposed to empirical, because it relies on the epistemological suspicion that knowledge of mind-independent, causally inert moral properties is inexplicable,[15] and it is normative because it demands that proponents of such properties explain how our beliefs about these properties could be reliable. Debunkers rely on it in their argument against robust moral realism.

[14] This chapter does not assess how the Benacerraf-Field challenge fares against CORRECT. I turn to this question in chapter 7.

[15] The precise epistemic principles behind the Benacerraf-Field challenge are controversial; see Clarke-Doane (2017c) and chapter 7.

Therefore, empirical, causal considerations, as in INFLUENCE or MISLEAD, are not sufficient to reach a sceptical conclusion against robust moral realism. The empirical debunking argument turns out to rely on a conceptual, normative claim that is based on the metaphysical nature of moral properties.

In addition, it seems that the reliance on the Benacerraf-Field challenge makes the empirical premises of the debunking argument redundant too. Once the Benacerraf-Field challenge establishes that none of our moral beliefs are likely to track moral properties (hence, that CORRECT is false), the empirical premises INFLUENCE and MISLEAD do not provide any additional sceptical oomph: blocking the realist's escape route via CORRECT would already secure the sceptical conclusion that we have sufficient reason to doubt the truth of all our moral beliefs on a realist account.

Debunkers might object: even though the empirical premises, INFLUENCE and MISLEAD, are not sufficient to reach the sceptical conclusion, they seem to provide *additional reason* to doubt realism. Let us consider an example to make the objection vivid and then see how it can be answered:

> Case 1: Suppose that Alf ingests an anti-maths drug which makes his mathematical beliefs unreliable; this gives him sufficient reason to doubt the truth of his mathematical beliefs.
>
> Case 2: Mathematical objects are causally inert; they cannot influence our mathematical beliefs; this also gives Alf sufficient reason to doubt the truth of his mathematical beliefs.
>
> Although case 1) and 2) lead to the same conclusion, they cannot thereby be reduced to one another. Case 1 gives Alf *additional* reason to doubt the truth of his mathematical beliefs.

The anti-maths drug case resembles the empirical, causal debunking argument: the evidence about Alf ingesting the drug seems roughly analogous to the empirical claim about the causal origins of our beliefs. Clearly, this gives us additional reason to be sceptical about Alf's maths beliefs.

However, the anti-maths drug case is different from the debunking of moral realism. In the case of the anti-maths drug, we might legitimately stipulate that there is no way for Alf to have at least some reliably true maths beliefs: we just

assume that the anti-maths drug rules out this possibility. However, this stipulation cannot be made in the causal debunking argument, since debunkers cannot presuppose that it is impossible for realists to fine-tune their moral beliefs without begging the question. Instead, debunkers need to argue for it, and they rely on the causal inertness and mind-independence of moral properties to fashion an argument to that effect. This, we have seen, is a version of the Benacerraf-Field challenge. Hence, even though the sceptical conclusions of the maths-belief cases above are independent, and the sceptical oomph that they provide is additive, the empirical debunking argument against realism is not independent of the conceptual argument. On the contrary, reaching the conclusion of the empirical argument *depends* on the conceptual argument against CORRECT. So, in contrast to the maths-case, posing the empirical debunking argument requires a successful conceptual debunking argument – and it is doubtful whether the empirical debunking argument would give us any *additional* reason in any interesting sense since realists would have had to concede defeat already.

More specifically, the sense in which the sceptical conclusion of the empirical causal debunking argument is *additive,* that is, gives us *additional* reason to be sceptical about realism, is only in the weak sense in which the following italicised variants of INFLUENCE give us additional reason to doubt the truth of specific realist moral beliefs:

> Case 3: Anton believes that donating to charities is good, not because the belief is reliable and true, but because *he heard his neighbour say it.*
>
> Case 4: Bob believes that eating animals is wrong, not because the belief is reliable and true, but because *he was moved by the cover of Peter Singer's book Animal Liberation.*
>
> Case 5: Cliff believes that gender equality is just, not because the belief is reliable and true, but because *he has an evolved sense of fairness*.

These particular instances of INFLUENCE surely seem spurious, and the impression created by Singer's book cover, for example, is in itself not a good indicator of the truth about animal ethics. But if we knew that Bob studied the book's content and, as a robust realist sympathetic to Singer's view would claim,

thereby formed reliable beliefs about the truth about animal ethics, we should conclude that doubting Bob's belief is unwarranted. Given the realist's commitment to non-causal explanations of moral knowledge, the evidence about the actual causal origins of Anton's, Bob's, and Cliff's beliefs are not, in themselves, troubling. It may raise our suspicion, but it cannot justify our suspicion if we have not ruled out their truth-tracking ability first. What debunkers need is the claim that Anton, Bob, and Cliff are unlikely to believe what they believe *because* it is true. However, once debunkers secure that point, we need not worry about the actual causal origins of our beliefs any more. Realists would already be in deep trouble if their explanation of non-causal, but nonetheless reliable, belief-forming methods failed; additional genealogical considerations would not worsen the problem for robust realists.

Hence, there are good reasons to think that the empirical premises, and the DEBUNKING CONDITIONAL, do not suffice to debunk moral realism, and also that *adding* the empirical premises does not *add* problems for robust realism.

### 3.3.3 Empirical Premises are not Required

Perhaps, however, debunkers might think that their empirical premises are necessary to reject CORRECT. Two prominent interpretations of the DEBUNKING CONDITIONAL, suggested by Street (2006) and taken up in the literature, suggest that they are not. The mind-independence and causal inertness of moral properties are doing all the work.

First, many have suggested that the debunking premise rests on probabilistic considerations. In the words of Shafer-Landau, the odds of adopting a true moral belief are low because "our actual moral beliefs represent only a small portion of all possible moral beliefs" (Shafer-Landau 2012: 11). It makes sense to interpret Street's evolutionary debunking argument as an argument that invokes probability, particularly when she writes:

> Of course it's possible that as a matter of sheer chance, some large portion of our evaluative judgements ended up true, due to a happy coincidence between the realist's independent evaluative truths and the evaluative directions in which natural selection tended to push us, but this would require a fluke of luck that's not only extremely unlikely, in view of the huge

> universe of logically possible evaluative judgements and truths, but also astoundingly convenient to the realist. (Street 2006: 122)[16]

The assumed fact that illegitimate forces had an impact on our moral beliefs is taken to imply that our actual moral belief system has only a slight chance of matching the correct moral belief system. Street concludes that many of our actual moral beliefs are likely to be "off track" (Street 2008b: 208). On this interpretation, the DEBUNKING CONDITIONAL is motivated by alluding to the conceptual possibility of there being many different possible moral belief systems but no way of discriminating the *correct* moral belief system. Thus, the low probability of adopting just the right moral belief system amongst countless false moral belief systems is sufficient reason to doubt the truth of our moral beliefs. However, regarding the claim that we are unable to identify the correct moral belief system to be convincing, we have to assume that CORRECT has faltered. Hence:

> LOW PROBABILITY (LP): If REALISM is true and CORRECT is false, then there are many conceptually possible moral belief systems and no way of telling whether the one we have currently adopted is the correct one.
>
> LP-DEBUNKING COND.: If LOW PROBABILITY is true, then we have sufficient reason to doubt the truth of all our moral beliefs.

Others interpret the argument's debunking claim as depending on considerations about counterfactuals. This interpretation is formulated most clearly by Clarke-Doane (2012). He argues that the debunking premise of the argument is substantiated by the following conditional:

> [I]f our moral beliefs were the products of evolutionary forces, then those forces would be "non-truth-tracking"—that is, […] if we were selected to have certain moral beliefs at all, then we would not be selected to have true moral beliefs. (Clarke-Doane 2012: 325)

This would create a problem for robust moral realism because "if we were not selected to have true moral beliefs, then had the moral truths been very different, our moral beliefs would have been the same" (Clarke-Doane 2012: 319). The essence of Clarke-Doane's understanding of the debunking premise is the claim

[16] See also Street (2008b: 208, 2011: 14).

that our moral beliefs are the products of 'non-truth-tracking' forces. However, Clarke-Doane's counterfactual also depends on the claim that non-causal determinants of our moral beliefs cannot play a role; hence, Clarke-Doane's claim depends on the assumption that CORRECT is false:

> COUNTERFACTUAL (C): If REALISM is true and CORRECT is false, then we would still have the same moral beliefs even if the moral truths were different.
>
> C-DEBUNKING COND.: IF COUNTERFACTUAL is true, then we have sufficient reason to doubt the truth of all our moral beliefs.

Both the probabilistic and the counterfactual interpretation of Street's DEBUNKING CONDITIONAL seem plausible. Both could be used to infer a version of DOUBT. However, these interpretations do not require an empirical, a posteriori claim about the origins of our moral beliefs to be convincing. Rather, they depend on the a priori claim that moral properties are similar to the properties of abstract entities in that they are causally inert and mind-independent. If it were not for this fact, neither interpretation would have much bite. Consider LOW PROBABILITY. Our inability to tell whether our current moral belief system is the one that matches the facts depends on (the conceptual possibility of) our beliefs being mismatched with the moral properties. That, in turn, depends on the moral properties being mind-independent and causally inert. Consider COUNTERFACTUAL. Our moral beliefs would be insensitive to changes in the moral truth only if the moral properties, which constitute the moral truth, are mind-independent and causally inert.

Therefore, both interpretations of the DEBUNKING CONDITIONAL require the same observation about the nature of moral properties to be convincing, which is, as we have seen, not an observation of any contingent empirical fact, but of necessary metaphysical facts.

Recent discussions of Street's debunking argument frequently miss this point. Bogardus, for instance, looks into the epistemic principle that underlies the argument, which indeed is a key question, but then expects, mistakenly, that it somehow needs to combine with "the facts of evolution" to create a sceptical conclusion (Bogardus 2016: 636). Vavova claims that the argument rests on a claim about evolutionary influences, but does not recognise that this is the case only if

CORRECT is rejected (Vavova 2015: 108). Referring to Street (2006), Deem (2016) and Artiga (2015) defend elaborate empirical hypotheses about the compatibility of particular causal explanations of our moral beliefs and moral realism, ignoring that the issue with CORRECT cannot be settled by any causal explanation.

When the crucial nature of moral properties, as conceived of by moral realism, is noted, for instance by Bedke (2014), who refers to moral beliefs being 'oblivious' to the moral properties, or by Crow (2016), then it is too often overlooked that no causal claim of any kind is required to make the argument work. Similarly, Clarke-Doane (2012) observes, in line with Street (2006), that the particular details of the causal history are irrelevant, but he does not argue that no causal claim is needed for the argument (Clarke-Doane 2012: 326). Enoch suggests that the core issue is to explain the correlation between the moral beliefs that we take to be true and realist moral properties, which resembles the Benacerraf-Field challenge (Enoch 2010: 421). But still, his answer to the challenge involves a causal explanation of our moral beliefs, which wrongly suggests that the causal history of our moral beliefs, and competing elucidations of it, are indeed the problem that underpins the debunking argument. Framing the issue like this is misleading as it puts too much emphasis on one or the other interpretation of the causal origins of our moral beliefs, on the soundness of such explanations, or on the seemingly extraordinary fact that we can illuminate the causal background of our moral beliefs. The crucial fact is, as we have seen, that the nature of moral properties puts the realist's claim to CORRECT in jeopardy.

The interpretations of Street's DEBUNKING CONDITIONAL considered here support the conclusion that the debunking argument relies on the Benacerraf-Field challenge and, conversely, on conceptual claims about the nature of realist moral properties. This does not show that *any* debunking argument that leads to DOUBT relies on a conceptual argument against CORRECT rather than on INFLUENCE and MISLEAD. However, the above already gives us strong reasons to suppose that the causal inertness and mind-independence of moral properties (and the resulting conceptual, epistemic problem) lie at the heart of other supposed causal debunking arguments against robust moral realism too. The alleged Darwinian debunking argument contains the spirit of Benacerraf and Field but not that of Darwin.

### 3.3.4 Intermediate Conclusion

Street's evolutionary debunking argument, an apparently empirically informed argument against realism, turns out to rest on the conceptual claim that moral properties are mind-independent and causally inert and that this it unlikely that our moral beliefs are true (more often than chance would predict). The latter observation 'crowds out' empirical considerations from the factual premises of the argument.

Debunkers should be wary of accepting this conclusion. Although the Benacerraf-Field challenge appears[17] to be a powerful epistemological challenge to robust moral realism, it fails to constitute an argument against realism that is based on an empirical premise about the causal origins of our moral beliefs. Debunking arguments would be misleading without an empirical premise because the arguments would render empirical premises redundant to making the argument sound.

## 3.4 Evolutionary Defeat Requires A Priori Claims

Debunkers might concede that one strategy against CORRECT relies on the Benacerraf-Field challenge but argue that another route shows how variants of the empirical premises INFLUENCE and MISLEAD can force realists to accept the DEBUNKING CONDITIONAL without a conceptual argument against CORRECT. This is how the debunking argument is often understood, and this section will attempt to dispel this myth.

This is the myth: that debunking arguments provide us with evidence that most or even all of our moral beliefs are influenced by such-and-such causal processes; and we can see that some of these beliefs are false.[18] Hence, since we know about the pervasive influence of causal forces on our beliefs and the

[17] I question this assumption in chapter 7.

[18] Debunkers must claim that some of the moral beliefs are *false*, as opposed to *lacking in truth-conduciveness*, because the latter would be an instance of the Benacerraf-Field challenge again. Street, of course, does not say that evolutionary influences show that our moral beliefs are false, but I take it that the previous section established that any claims to the effect that our moral beliefs are not truth-conducive are based on the Benacerraf-Field challenge.

occasional distorting effect of this, we can infer that we should be sceptical about the truth of *all* our moral beliefs; they do not seem to track moral properties, despite the possible corrective influence of the non-causal relations postulated by robust realists.[19] So, the myth goes, realism is debunked without relying on a version of the Benacerraf-Field challenge.

Why do debunkers need a substantive moral claim? Because to show that causal influences are distorting, without relying on the claim[20] that causal influences are the *only* possible determinants of our moral beliefs', requires evidence that our beliefs are distorted despite possible alternative, non-causal determinants of our beliefs (like those the realists claim exist). The required evidence can only be provided by showing that the causal influences in question lead us to have false moral beliefs, and to determine that we have false moral beliefs, debunkers need to commit to substantive moral claims about which moral beliefs are false.

However, assuming the truth of substantive moral claims is problematic in an argument that raises doubt about the truth of *all* our moral beliefs. Such a 'global' debunking argument would call these very assumptions into question.[21] Put another way, an inductive argument supporting the DEBUNKING CONDITIONAL would cast considerable doubt on the truth of all of our moral beliefs. However, in that case, we would also have reason to doubt the truth of the substantive moral claim that debunkers rely on to provide us with evidence of distortion: the argument would be self-defeating.

---

[19] The terminology Street uses when she describes '*distorting*' causal influences on our moral beliefs may be partly responsible for this myth Street (2006: 155). Street uses 'distorting' in a weak sense, in which case there is only a non-truth-tracking but no truth-tracking force present. This is suggested by her Bermuda analogy: setting out by boat for Bermuda and letting the wind and tides determine one's course is not clever because, without using the sail and the rudder, the wind and waves do not push you to your target Street (2006: 121–2). However, at this point in the dialectic, debunkers have to assume that correction is possible and thus understand distortion in a *strong sense*: a non-truth-tracking force distorts even though a truth-tracking force might be present too. Examples of this kind of distortion are the forces of waves that erode a cliff or the winds at Cape Horn that push a capable sailor off his intended course.

[20] Which would be an instance of the Benacerraf-Field challenge.

[21] 'Global' is also used by Kahane (2011) and Shafer-Landau (2012).

Debunkers might object that there are cases in which we can legitimately rely on substantive truths about the beliefs that we seek to debunk. Consider the following case:

> Suppose I provide you with evidence that your car's thermometer is distorted by the heat of the car's engine: "Your car thermometer displayed +7 degrees, but it was -5!" After receiving this information, you ought to doubt all future readings of your car thermometer because it might be distorted at any given moment.

In cases like the thermometer case, where a particular belief-producing faculty is being debunked, we can rely on our knowledge about the truth in question: we have no reason to assume that *our* beliefs about the temperature are unreliable. However, the debunking case is not analogous. The debunkers' 'fact-checking' of the moral truth is called into question by their own argument. They attempt to debunk moral cognition entirely, and thus cannot rely on beliefs that are safeguarded from the potentially debunking information. Hence, debunkers cannot rely on their own moral beliefs in the debunking argument; to avoid self-defeat they must stay agnostic about the moral truth.[22] However, in that case, their global debunking argument does not work because without evidence of distorted, false moral beliefs (via a substantive moral claim) or evidence of the absence of a possible corrective influence (via the Benacerraf-Field challenge) we do not have sufficient reason to doubt the truth of *all* of our moral beliefs.

Second, while this particular problem can be avoided by attempting to debunk only particular moral beliefs, that strategy fails too, because the inference to DOUBT would fail: we cannot infer sufficient reason to doubt *all* our moral beliefs from the falsity of *one* particular belief; if we did, the scope of the debunking argument would be drastically reduced.

Take the example of a xenophobe who believes that all non-group members must be killed. A plausible causal explanation of this belief might be the xenophobe's evolved tendency to distinguish between members of his in-group and members of the out-group (cf. Wielenberg 2014):

[22] See Shafer-Landau (2012), who distinguishes between agnostic- and knowledge-based EDAs.

XEN-INFLUENCE: The belief that all non-group members must be killed was influenced by evolutionary forces.

To show that the causal influence reported in XEN-INFLUENCE is distorting, despite the possibility that moral beliefs non-causally track moral properties, we need to point out that it is not, in fact, the case that all non-group members must be killed.

Now we have evidence of a causal influence on the xenophobic belief and evidence that the belief is false. In the case of the xenophobe, debunkers might claim, CORRECT evidently faltered. But this is no victory for debunkers: they cannot infer DOUBT from the falsity of one particular moral belief. The same holds if debunkers substitute the particular xenophobic belief with *subsets* of other beliefs, say all deontological beliefs, to increase the impact of their argument (cf. Greene 2008). They could argue as follows:

DEONTOLOGY INFLUENCE: All deontological moral beliefs are influenced by such-and-such causal forces.

But, again, the argument is limited in scope: it targets only a specific subset of moral beliefs, distinguished by its content or by its origin, but not *all* moral beliefs as in the desired 'global' debunking argument. The best that can be hoped for is an inference to doubt the truth of those beliefs that are part of particular subsets. So, 'going local' does not afford the desired inference to DOUBT and 'going global' puts the required substantive moral assumption into jeopardy.

Moreover, there is an issue with substantive moral assumptions even on the local level. It appears that causal considerations are, yet again, superfluous. We saw, for instance, that premises with substantive moral content, like the belief that belonging to the out-group is not a reason to treat one differently, are crucial for inferring sufficient reason to doubt the truth of particular moral beliefs, or subsets of moral beliefs. However, in that case, the substantive moral claim seems to do the debunking, not the factual claim about the genealogy of the beliefs in question.

For instance, suppose that I try to debunk moral beliefs that are moderated by disgust, which would involve an empirical claim about the subset of all moral beliefs moderated by disgust (cf. Sauer 2012). However, the sole fact that these beliefs are moderated by disgust should not bother us in the least. We should be

bothered only when we are convinced that moral beliefs moderated by disgust are often false. However, disgust-moderated moral beliefs are not false *because* they are influenced by disgust; they are false because disgust in itself seems morally irrelevant.

Hence, what counts in an attempt to debunk all disgust-moderated beliefs is the substantive moral judgement *about* the causal influence in question, not evidence of the causal factor itself. But if we already know that all disgust-related moral beliefs are false then it seems that *how* we came to have these beliefs is irrelevant for our verdict about them.

Therefore, the evolutionary, causal debunking argument against robust moral realism fails. An evolutionary argument on the global level, which would affect *all* moral beliefs, relies on substantive moral assumptions that are called into question by the debunkers' very own argument. The argument would be self-defeating. A 'local' argument that affects only particular moral beliefs, or subsets of moral beliefs, does not warrant an inference to DOUBT and is thus of limited scope. Moreover, since the local argument relies on substantive moral assumptions too, it is, yet again, doubtful whether the empirical premises play an important role after all.

## 3.5 Concluding Remarks

Street's EDA is 'old wine in new bottles': it lacks an essential Darwinian causal premise and relies on the Benacerraf-Field challenge, so it does not pose a novel, Darwinian challenge for robust moral realism. Construing the argument with an essential empirical premise is self-defeating or leads to a limited scope. Therefore, there is no *evolutionary* debunking argument against robust moral realism.

Old wine need not be foul; the debunker might well be content with stating a brushed-up version of the Benacerraf-Field challenge. In that case, however, the reference to empirical, Darwinian considerations is reduced to an illustrative veneer that is ultimately redundant to reaching the argument's conclusion. The implication is that we should readjust focus in the metaethical debunking debate: considering the causal history of our moral beliefs, in defence of realism or in criticism against it, is simply beside the point when considering the possibility of knowledge of objective moral facts.

# 4 Resisting Defeat

## Reader's Guide

Suppose that there is some way in which evolutionary explanations undercut all moral beliefs (though I have not described one yet); would that mean that moral beliefs are permanently unjustified? Or could we resist defeat, perhaps by regaining lost epistemic justification of our moral judgments?

In this chapter, I will address this question about the severity of the evolutionary defeat challenge. I will first draw out some general observations about defeating defeaters and then argue that *if* the challenge succeeds, we could not reinstate the justification of moral beliefs (objectively construed) by *defeating* or *deflecting* the defeater. The difference between a defeater-defeater and a defeater-deflector is that the former defeats a defeater whereas the latter prevents information from being defeating. Before explaining this distinction in more detail, I should emphasise that this chapter proceeds on the assumption that there is some way in which the evolutionary defeat challenge works. This is of course contrary to what we have seen so far. We have seen that existing accounts of undercutting defeat imply that evolution does not undercut moral beliefs (chapter 1) and that evolution cannot provide a defeater without the support of an a priori reason to doubt that moral judgements are justified (chapter 2). Nonetheless, it will enhance our understanding of the evolutionary defeat challenge to see how devastating it would be, should it succeed.

In principle, any defeater can itself be defeated by an undefeated mental state. For example, recall the example from section 2.1 about Tom Grabit, the book thief. Suppose you initially formed the belief that Tom stole the books as he rushed away with them. Your justification for holding this belief got defeated by your friend's testimony that Tom's twin brother is in town. However, the defeater you got from your friend, let's call her Kelly, can itself be defeated if you learn, from Clarke, another reliable source, that Kelly was lying to you about Tom's twin brother. As you might imagine, this chain of defeaters may go on and on (Pollock

1987, 1995).[1] As long as we know how to correctly determine relationships between defeaters, understanding this process is easy enough. For example, it seems plausible that we should trust Kelly's testimony and ignore Clarke's if we have a higher credence in Kelly's testimony than in Clarke's (Casullo 1988; Pollock and Cruz 1999). Thus, defeat of particular beliefs can be resisted by defeating the defeater itself.

Resisting global defeaters is a different matter. Let a global defeater be a defeater that defeats all of a subject's (non-analytic) beliefs. For example, consider the following case (inspired by Palmer 2018):

> Fascinated by the potential of newly available Brain-Computer interfaces, but short on money, you seek out a dodgy but economical practitioner who promises that implanting an electrode in the "reading-area" of your brain will at least double your reading speed. Unfortunately, the practitioner was a fraudster. During the operation, he irreparably severed your temporal lobe. Plagued by terrible headache post-operation, you visit a reliable and trustworthy doctor. The doctor diagnoses your headache as a syndrome of a severed temporal lobe and informs you that you will have irreparable problems with thinking and memory retention as a result of your injury.

If you found yourself in this predicament, then the justification of many if not all of your non-trivial beliefs would be defeated. Your doctor's testimony gives you a global defeater. After all, you will know that any belief you would form could be ill-formed due to your malfunctioning brain, and you would not be able to tell, in your own mind, whether you are thinking straight and whether your memories are truthful representations of the past. Whatever new non-trivial belief you form, you could not justifiably rely on it without checking with others whether you are right. Such a global defeater is *robust* in that it has implications for most other mental states and that it does not allow for any defeater-defeaters if it affects sufficiently many beliefs. Because of the latter implication, the case of resisting global defeat is similar to the case of resisting the radical sceptical challenge. A justified belief

[1] On Pollock's account, there is no clear 'justification in the limit', that is, there are only presently undefeated beliefs, but no guarantee that there are permanently undefeated beliefs.

seems required to rebut the challenge, but none is available (either because all justified beliefs have been defeated or because there has never been a justified belief in the first place, as argued by proponents of the radical sceptical challenge).[2]

The evolutionary challenge aims at defeating all beliefs of a certain type. Let's call defeaters that defeat all beliefs of a particular type 'type-defeaters'. Resisting type-defeaters can, in principle, be done in the same way as defeaters of specific beliefs are resisted: by defeating them. However, since defeater-defeaters must themselves be undefeated, type-defeaters raise interesting challenges if they can be defeated only by beliefs of the type that they are targeting. That is, there might be cases where the defeater D defeats beliefs of type T and where we also seem to need T-beliefs to resist D. For example, if your introspection is defeated, in the absence of any external information that discredits the defeater, you would need to rely on your introspection to defeat the defeater. With your introspection defeated, however, no belief would be available to defeat the defeater.

Of course, it is difficult to say how to categorise types of beliefs, and I won't venture into developing an account here. For present purposes, it suffices that we often intuitively distinguish between different types of belief, such as perceptual beliefs, moral beliefs, or beliefs based on introspection. These types might overlap (that is, a belief could be both a perceptual and a moral belief), but there seem to be relevant differences between them such that beliefs of some type are required to say anything of epistemic relevance about beliefs of another type. For the purposes of this reader's guide, we can set aside these difficulties.

My point here is that *defeat* of all moral judgements (objectively construed) would be definite: their justification could not be recovered. To see why, suppose that there is some way in which the evolutionary defeat challenge undercuts all moral beliefs and let this be defeater D. So, suppose that we have found an answer to my main research question and we know how to legitimately get from evolutionary explanations of morality to a defeater, D, of all moral judgements. How could D itself be defeated? Let's consider three 'types' of beliefs that could serve as defeaters for D: beliefs about science, epistemic beliefs, and moral beliefs.

[2] In the case of global defeat, the more interesting question seems to be whether it can be instantiated in the first place. How should a system of beliefs react towards the information that *all* its beliefs are distorted or will be distorted?

One obvious route would be to rebut the empirical, evolutionary claim about the causal origins of our moral beliefs. Given the strong support for evolutionary explanations of (a subset of) our moral norms, as outlined in chapter 1, this would be an outrageous claim, and except for Thomas Nagel, nobody has defended this view in print (Nagel 2012). Perhaps a weaker scientific claim would suffice to defeat D. Perhaps it suffices to rebut specific claims about, say, the extent of evolutionary influence on our moral judgements to defeat D. However, in light of the previous chapter, this would not help at all. We have already seen that specific claims about the causal origins of our moral beliefs will play no vital role in raising a defeater for objectivist moral beliefs. So, D cannot be defeated by a true scientific belief about the extent to which evolution influenced our moral judgements. Consider beliefs about epistemology next. Recall from chapter 2 that there have to be perspective-independent, objective conditions for defeat and thus there must be a fact of the matter about whether or not D undercuts our moral beliefs. It is therefore possible that D undercuts our moral beliefs, even if we think it does not, and D might *not* undercut our moral beliefs even if we (mistakenly) think it does. *If* D undercuts our moral beliefs, then there is a valid epistemic principle that gives rise to D.[3] In that case, true beliefs about epistemology cannot defeat D because they would have to imply that D is based on an invalid epistemic principle (and thus have *no* defeating power).[4] Of course, false beliefs about epistemology might defeat D, but I assume that this would not be a cause of celebration for defenders of morality. Hence, appealing to epistemic considerations to defeat D will not work either. This leaves moral considerations concerning defeating D. Clearly, however, since all moral beliefs are already defeated, none are available to defeat D. This is why there are good reasons to think that *if* the evolutionary defeat challenge

[3] I am assuming here that D is indeed a defeater and that it objectively reduces the justification of our moral beliefs. Again, we might not acknowledge this, or might fail to grasp that D depends on a correct principle and thus *mistakenly believe* that D will be undercut through, for example, switching epistemic theories. We will only think that D is a defeater, but it won't be.

[4] This is a good place to address a possible objection about the alleged *cumulative force* of different versions of the evolutionary defeat challenge. We might of course believe that D undercuts our moral beliefs and justify that impression by citing a number of plausible explanations. But if these explanations turn out to be flawed, then our impression is mistaken: D does not undercut our moral beliefs and several mistaken explanations of why D would do so do not change this picture.

succeeds (that is, if it creates a defeater), defeat could not be resisted. The damage to the justification of our objectivist moral beliefs would be irreparable.

In light of these considerations, Andrew Moon's (2017) recent suggestion of a novel route to resisting defeat is highly interesting: rather than defeating a defeater, he suggests that moral objectivists might have the chance to *deflect* defeat and thus to resist the evolutionary defeat challenge even if it the challenge would, all else being equal, produce a defeater. At the heart of Moon's proposal is the idea that some information might *deflect* the defeating power of information that would have resulted in defeat, had one not had the relevant deflecting information. Considering Moon's proposal is thus relevant to my main question because it sets out one way in which a defeater might be resisted. Since Moon discusses robust moral realism, I do so too. However, as discussed in the reader's guide to chapter 2, my argument applies to moral objectivism too. Let's take a thorough look at Moon's claims in the next section.

## Abstract[1]

I address Andrew Moon's recent discussion (2017) of the question whether third-factor accounts are valid responses to debunking arguments against moral realism. Moon argues that third-factor responses are valid under certain conditions but leaves open whether moral realists can use his interpretation of the third-factor response to defuse the evolutionary debunking challenge. I rebut Moon's claim and answer his question. Moon's third-factor reply is valid only if we accept an 'externalist view' of epistemic defeaters. However, even if we do, I argue, the conditions Moon identifies for a valid third-factor response are not met for moral realism to refute the evolutionary debunking argument.

## 4.1 Introduction

Most moral realists believe that we can have objective moral knowledge: moral properties and facts exist stance-independently, and we can have justified true beliefs about them (Enoch 2011b; Shafer-Landau 2003; Wielenberg 2014). The reliability challenge to objective moral knowledge is intended to show that all of our moral beliefs are unjustified – at least insofar as the moral beliefs are about stance-independent properties and facts of the sort defended by moral realists. The challenge is often based on evolutionary explanations of morality (Joyce 2006; Street 2006, 2016).[2] If the challenge succeeds, and if justification is required for knowledge, then objective moral knowledge seems impossible. Moon (2017) understands the reliability challenge to be a probabilistic challenge. Accordingly, the reliability challenge provides an epistemic defeater for $R_m$, where $R_m$ stands

[1] This chapter is based on a paper published as "Can Moral Realists Deflect Defeat Due to Evolutionary Explanations of Morality?" *Pacific Philosophical Quarterly* (2017) 98 (S1): 227–48.

[2] In this chapter, I adopt Moon's understanding of moral realism, according to which moral realism entails three claims: there are stance-independent moral properties and facts; it is possible to have knowledge about these properties and facts; and moral properties are irreducible to non-natural properties. This conception of moral realism excludes moral naturalists, such as Brink (1989), who does not think that moral properties are irreducible, and realists who do not think that moral properties are stance-independent, such as Railton (1986). This restriction makes sense because Moon's probabilistic interpretation of the reliability challenge entails the claim that realists have reason to believe that the probability is low that we have reliable moral beliefs, given their conception of morality and evolutionary explanations of morality. It is less clear, and worth a separate discussion, whether non-robust realists have equally strong reasons to believe that that probability is low Moon (2017: 216–7).

for the belief that our moral beliefs are generally reliable for anyone who comes to believe that the probability is low that our moral beliefs are reliable, given our evolutionary past and the truth of moral realism (Moon 2017: 216–7).[3]

Realists seem to have a straightforward answer to this challenge: so-called third-factor accounts assume the truth of some particular, substantive "Morality Claim" M (Moon 2017: 215) and then use M to demonstrate that our moral beliefs are by and large reliable (Brosnan 2011; Enoch 2010; Skarsaune 2011; Wielenberg 2010). For instance, the morality claim favoured by realist David Enoch is that "survival or reproductive success is at least somewhat good" (Enoch 2010: 430). Faced with a probabilistic version of the reliability challenge, realists might argue that, given that M is true, there is a moderately high probability that $R_m$ is true too (Moon 2017: 215). Hence, our moral beliefs seem justified after all and therefore the reliability challenge is thwarted.

However, whether realists are entitled to assume that M is true is the key question, if not the "heart of the debate between realists and the debunker" (Vavova 2015: 111). There are two aspects to this debate. On the one hand, the third-factor response against the reliability challenge appears problematic because the reliability of our moral beliefs is called into question by a defeater – it seems circular or question-begging to rely on a defeated moral belief to fend off the challenge. Many take this to be a reason to dismiss swiftly the realist reply (Fraser 2014: 471; Street 2008b; Vavova 2015: 111). On the other hand, circularity seems, to a certain extent, inherent in explanations of the reliability of any class of beliefs. Arguably, we can explain the reliability of any particular class of beliefs only if we assume the truth of some of the beliefs in question (Berker 2014; White 2010). For instance, if you wonder about the reliability of your maths beliefs, you might work out the 42nd decimal of pi and check your result against a reliable source (the answer is 9). Relatedly, some argue that we can trust our cognitive capacities without prior evidence and without begging the question (cf. Foley 2001). Third-

[3] Moon's overt project in his paper is the comparison between the evolutionary argument against naturalism and the evolutionary argument against moral realism. I do not discuss this part of his paper in this chapter. Instead, I focus on the implications that Moon so helpfully draws from his comparison.

factor responses could fall into either category, so they are not quite so easy to dismiss.[4]

The way forward in this debate suggested by Moon (2017) is to identify a response to epistemic defeaters that does not beg the question and to argue that if moral realists can adopt this response, then they are on their way to answering the reliability challenge.

In this chapter, I address the question that Moon leaves unanswered: is Moon's proposed interpretation of the third-factor response available to moral realists? Can they 'deflect' the possible defeat due to evolutionary explanations of morality? I argue for two points. First, the success of Moon's generic response to defeaters depends on the truth of externalism about epistemic defeaters.[5] However, Moon does not address the relevance of externalism for his interpretation of the third-factor response and what he writes suggests that he assumes that internalism is true. So, Moon's interpretation of a valid third-factor response is either false, if internalism is true, or misleading, insofar as Moon fails to address the relevance of externalism for his interpretation of the third-factor response.[6] Second, and more directly pertinent to the reliability challenge in the moral case, even if we assume the truth of externalism to make Moon's general strategy work, the particular case of moral realism does not satisfy the conditions for employing validly Moon's strategy.

---

[4] Provided that proponents of the reliability challenge want to be sceptical about morality only and avoid scepticism regarding logic, mathematics, science, and the external world, they have two options. Either they avoid the problem of reliability in certain domains, for instance by adopting constructivism and/or a deflationary theory of truth in these domains, or they show why circularity in explaining reliability is a problem particularly in the moral case, but not in any of the other cases. Lest we are interested in trite burden-shifting arguments (i.e. do moral realists have to show that the moral case is realist-friendly or do moral sceptics have to show that it is realist-averse?) It is worth identifying 'realist-friendly domains' in the sense that non-question-begging defences of our beliefs in that domain, realistically construed, are possible.

[5] Externalist views of epistemic defeaters imply that the defeating information need not be accessible (or conscious) to the subject, whereas internalism requires defeaters to be somehow accessible to the subject.

[6] Moon does not provide an explicit definition of epistemic defeaters in his (2017) at all. However, as I show below, Moon's discussion *implies* that an internalist conception of defeaters is correct, which is all I need for the claim that it is misleading to suggest a view that works only if externalism is true, while presenting it in accordance with internalist principles.

Although I ultimately reject the applicability of Moon's answer to the moral case, his proposal, and my discussion of it, should be of interest to those who are working on the metaethical debunking debate. It should help to locate the problem by answering the reliability challenge in the moral case and thus clarifying the constraints faced by moral realists.

I have a caveat: I assume for the sake of the present discussion that the reliability challenge (and evolutionary explanations in particular) provide us with a defeater of all our moral beliefs, irrespective of whether the defeater can be dealt with, quite like the defeaters in the examples that Moon discusses. This is a controversial assumption, and if realists can show that it is mistaken, then they would not need to rely on MOON'S WAY OUT in the moral case in the first place.[7]

Section 4.2 introduces Moon's argument. Section 4.3 contains my criticism of Moon's argument for a valid version of the third-factor response. Section 4.4 identifies disanalogies between Moon's artificial cases and the metaethical debunking challenge, and in section 4.5 I discuss and reject one further twist that might make the analogy between Moon's cases and the reliability challenge work.

## 4.2 Moon on Deflecting Defeat

Moon's probabilistic formulation of the reliability challenge relies on the notion of an epistemic defeater.[8] Short of an explicit definition, he uses the following case, adopted from Plantinga (2000), to illustrate what an epistemic defeater is:

> XX Pill Case: You learn that a pill, called "XX", destroys the cognitive reliability of 95% of those who ingest it. You take the pill and come to believe both that I've ingested XX and P(R/I've ingested XX) is low. (Moon 2017: 211)

[7] Some claim that evolutionary considerations do not provide a defeater in the first place; see Mogensen (2016b) and Hanson (2017). However, it stands to reason that the defeating power of the reliability challenge does not depend on evolutionary causal considerations per se, as we have seen in chapter 3, in which case the defeating power of the reliability challenge seems relevantly similar to the defeating power of the cognition-disrupting pills discussed by Moon. For reasons of space, I cannot engage further in this debate in this chapter.

[8] Defeaters are often distinguished as rebutting or undermining defeaters (Pollock 1970). Moon is concerned with the latter: they affect the justification of a belief, but do not show that the belief is false.

According to Moon, realists are confronted with an undermining defeater for $R_m$ – the belief that their moral cognition is generally reliable – if they believe that the probability is low that our moral cognition is reliable if moral realism is true and moral cognition is an adaptation. Moon assumes that realists do accept that the probability is low that $R_m$ is true, conditional on the claim that moral realism and the evolutionary claim are true (Moon 2017: 217). Adherents of the third-factor response want to use the morality claim to conclude that the probability that $R_m$ is true is at least moderately high. However, Moon notes that in most cases "it is illicit to use the Morality Claim to prevent the defeat of $R_m$ *when belief in the Morality Claim is a deliverance of the very faculties that are in question*" (Moon 2017: 217 emphasis added). This is the problem in the moral case.

However, Moon argues that it is false to assume that it is generally impossible to use beliefs produced by faculty F to avoid defeat of the faculty F. He uses the following case to illustrate this point:

> XX Deflector Case: All is as in the original XX pill case, except that before I took the pill, a scientist I know to be trustworthy had informed me that I am one of the 5% who is immune to the drug. I then take XX while knowing that I am one of the immune 5% and P(R/I've ingested XX and I am one of the immune 5%) is high. (Moon 2017: 221 italics in original)

In the 'XX-deflector case', Moon argues, we can legitimately use a belief (which is obviously a 'deliverance of' our cognitive faculty) to *deflect* a potential defeater of our cognitive faculty. Moon calls this way of using the morality claim a *defeater-deflector* and argues that the "charge of question-begging" against third-factor replies fails *if* the Morality Claim is understood as a defeater-deflector (Moon 2017: 221–2).

Why does the defeater-deflector account succeed according to Moon? He argues that the defeater-deflector account works because the subject in the XX-deflector case

> never gain[s] a reason to doubt R[9]: the subject starts with an undefeated belief in R, gains evidence that he is immune to the pill *prior* to ingesting the pill, takes the pill with the scientist's testimony still vividly in mind, and

[9] R stands for the belief that our cognitive faculties are generally reliable, see Moon (2017: 208).

> consequently is aware that he took a drug that he is immune to. (Moon 2017: 221–2)

Moon contends that this scenario would make it "very odd" to think that the XX pill could defeat R. So, Moon argues, *conscious memory* of the scientist's testimony received *before* the ingestion of the pill suffices to deflect the potential defeater gained by the XX pill (Moon 2017: 221–2). Moon emphasises the importance that the subject be "conscious" of the scientist's testimony after ingesting the pill, which is clear when he considers a turn of events that would cause the subject to lose its defeater-deflector:

> lost the belief (due to cognitive decline) that I am one of the immune 5%, but I did continue to believe that I took a drug that 95% of the population is vulnerable to. Then I would clearly get a defeater because I would no longer have the deflector. But so long as I continue to consciously believe that I am one of the immune 5%, it seems that the belief continues to have its deflecting powers. (Moon 2017: 222)

The facts about mental states highlighted by Moon, such as whether the subject consciously believes that he is immune to the defeater, are relevant only on internalist views on epistemic defeaters (cf. Grundmann 2011: 157). Moon's emphasis on these introspective facts implies that he is operating on the assumption that an internalist perspective about epistemic defeaters is correct. However, Moon leaves open whether what he calls a "strong internalist view" or a "moderate internalist" one is correct (Moon 2017: 225).[10] This distinction between both internalist views will be relevant for assessing Moon's account, so let me briefly explain it.

Strong internalism counts only "present, conscious states" as relevant (hence the emphasis that the subject be aware of the scientist's testimony), while the moderate internalists takes "unconscious states" to be "justificationally relevant" too (Moon 2017: 225). A moderate internalist will say that the subject

[10] The fact that Moon draws a contrast between two forms of internalism rather than between internalism and externalism lends further support to the impression that he operates under the assumption that internalism is correct. As I stated in the introduction, however, Moon does not explicitly endorse either position, and my claim is only that he does not seem to acknowledge the relevance of externalism for his interpretation of the third-factor response.

has a defeater-deflector even if the belief is not conscious but is nevertheless accessible (Moon 2017: 225).[11] The following figure depicts the sequence of events that are crucial in Moon's XX-deflector case:[12]

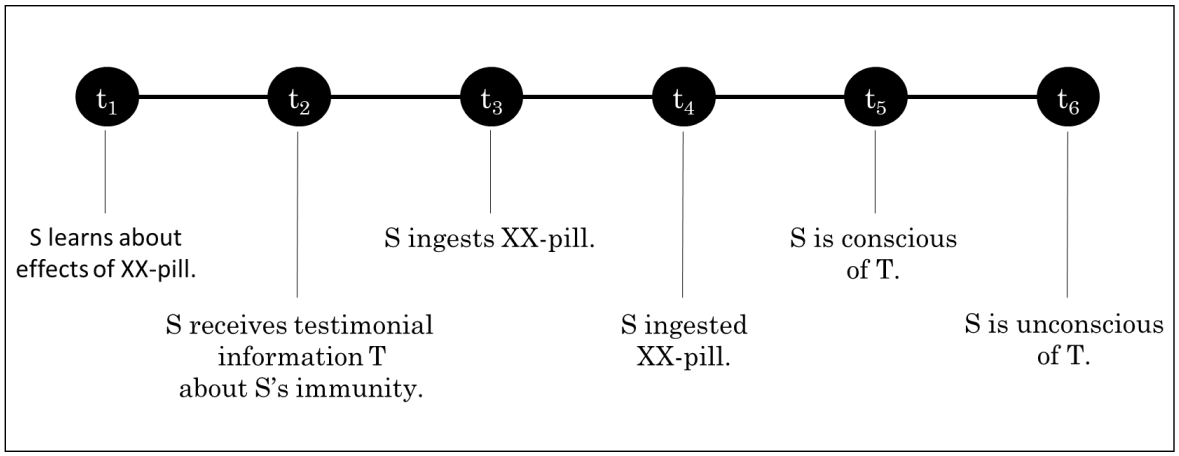


*Figure 4.1 Moon's case of defeat deflection*

S does not have a defeater prior to $t_3$. The decisive point is $t_4$. At $t_4$, S has already learned about the devastating effects of the pill and ingested the pill. According to Moon, S possesses a defeater-deflector at $t_4$. Moon suggests that, depending on whether strong or moderate internalism about defeaters is true, S continues to possess a defeater-deflector at $t_5$ but not at $t_6$ (strong internalism) or both at $t_5$ and $t_6$ (moderate internalism). Since Moon leaves both interpretations of internalism open, we can summarise Moon's principle about the sufficient conditions for responding to epistemic defeaters as follows:

> MOON'S WAY OUT: A subject S's belief f processed by faculty F can deflect the potential defeater PD of faculty F if f is formed prior to the reception of PD and, if strong internalism is true, S is conscious of f, or, if moderate internalism is true, S can access f.

Moon does not claim that the defeater-deflector interpretation expressed by MOON'S WAY OUT is the *only* escape route from the reliability challenge for realists. However, since he argues that two popular existing alternative interpretations of

[11] Like your belief that you live in house number so-and-so is 'conscious' now but was unconscious but nonetheless accessible before I prompted you to think of it.

[12] The depicted distance between events is not representative, nor does it matter for present purpose. The order of events does matter, however. $t_4$ can be arbitrarily close to $t_3$ but it must occur after $t_3$.

the third-factor account fail, he does regard the defeater-deflector interpretation as "the best way to understand the third-factor response" (Moon 2017: 219).[13]

But can moral realists use a claim in accordance with MOON'S WAY OUT in the moral case? Moon records a conditional answer: *if* realists are in a situation similar to the XX-deflector case, then they can use his interpretation of the third-factor reply. However, Moon is unsure about the antecedent of the conditional (Moon 2017: 222). Call this MOON'S QUESTION:

> MOON'S QUESTION: Can moral realists use a defeater-deflector as in the XX-deflector case?

Moon records that he does "not know" the answer, but he surmises that realists might be unable to use the defeater-deflector (Moon 2017: 222). He also offers a contrasting case that, if it resembles the realist's situation, bodes trouble for a successful third-factor response because it does not contain any beliefs that "have the power to deflect" defeaters:

> YY Colour Vision Deflector Case: [A pill, called 'YY' renders unreliable the colour vision of 95% of those who ingest it.] You do not know about YY, and you find yourself in [a room with objects that have no standard colour. For example, there are no bananas or blue jays, but there are plastic bowls, walls, and a chameleon]. You have already formed beliefs that the wall is red, that a bowl is white, and so on. Then a friend who you know to be a very reliable testifier tells you about YY and that YY was mixed into the dinner you had enjoyed earlier that evening. You come to believe that 'I ate YY' and that 'P(RC/I ate YY) is low'. (Moon 2017: 220)

I will come back to the YY-deflector case in section 4.4.3; the important features of the case are that, in contrast to the XX-deflector case, it does not as such contain a defeater-deflector and the defeater affects only one particular belief-forming faculty, as opposed to all belief-forming processes.

[13] I do not assess the claim that MOON'S WAY OUT is the *best* response to epistemic defeaters in this chapter. However, given that several authors reject the third-factor response as hopelessly question-begging (Fraser 2014; e.g. Street 2008b; Vavova 2015), and also reject Moon's critical analysis of alternative interpretations (Moon 2017: 218–21), it seems safe to say that a failure of MOON'S WAY OUT in the moral case would indeed be bad news for realists. Conversely, if MOON'S WAY OUT *were* true and applicable in the moral case, then realists should be able to provide a valid third-factor response that avoids the common criticisms of circularity or begging the question.

This completes Moon's picture as it is relevant to my discussion. The two crucial points are MOON'S WAY OUT (realists can respond to epistemic defeaters if they have a defeater-deflector), and MOON'S QUESTION (do moral realists have a defeater-deflector?). If MOON'S WAY OUT were true, it would provide a blueprint for responding to the most menacing epistemic defeaters. In what follows, I will challenge MOON'S WAY OUT and provide a negative answer to MOON'S QUESTION.

## 4.3 Moon's Way Out is False or Misleading

In this section, I argue that the truth of MOON'S WAY OUT depends on the truth of externalism about epistemic defeaters.[14] However, as we have seen, Moon emphasises that a subject's introspectively available information is important for assessing whether the subject has a defeater, which implicitly suggests that he operates on the assumption that internalism is correct. Hence, if internalism is true, then MOON'S WAY OUT is false. But if internalism is false, then MOON'S WAY OUT might work, but Moon's emphasis on internalist aspects of his argument for MOON'S WAY OUT is misleading.[15]

The problem with MOON'S WAY OUT is that the XX pill debilitates the reliability of S's cognitive faculties entirely. Cognitive faculties include memory. So, how could the subject be sure, at $t_4$, that his memory about his immunity to the pill is itself not just an effect of the pill?

Suppose we are dealing only with users of the pill who are *immune*. Moon allows them to carry on as usual if they keep in mind that they are immune to the pill, a concession that seems to rely on an odd picture of human psychology,

[14] I am concerned with MOON'S WAY OUT only and my argument touches only on the implications of either externalism or internalism for MOON'S WAY OUT, without claiming to suggest anything about the truth of either position.

[15] I discuss the comparability of the XX-deflector case and the reliability challenge for moral realists in section 4.4. The present section is nonetheless relevant for the metaethical debate: if MOON'S CLAIM is false even in the generic XX-deflector case, then realists would be hard pressed to find another valid way of responding to the reliability challenge, provided that Moon is right that the defeater-deflector interpretation is the best interpretation of the third-factor response.

because why is it important that users of the pill keep in mind[16] that they are immune to the belief pill?[17]

Moon does not explain why the mental states of the subjects matter when thinking about the defeater-deflector, but it seems that the *content* of IMMUNE alone cannot be relevant. Imagine a subject that is immune to the pill who, in contrast to the case outlined above, hears about his immunity only after ingesting the pill. From the subject's perspective, the belief in the reliability of his cognitive faculties would already be defeated, so he should doubt subsequent testimony, or memory about testimony, about his own immunity too. Hence, the content of IMMUNE alone does not suffice to deflect defeat.

Therefore, the important thing about IMMUNE is not its content per se, but that it is acquired *before* the ingestion of the pill. We can call this property of the belief TIME STAMPED: any belief that is TIME STAMPED was acquired *before* the ingestion of the pill.

Now, if subjects in the XX-deflector case could identify all TIME-STAMPED beliefs, then they could use the TIME-STAMPED belief IMMUNE to deflect the potential defeater of the XX pill, as suggested by Moon. Users of the pill would have to be capable of running through their beliefs, putting the 'good', i.e. TIME-STAMPED ones into the pot of reliable beliefs to act on, and discarding the potentially corrupted ones, just like the pigeons in the fairy tale *Cinderella* can pick good from bad lentils.[18] In that case, we could assume that immune users of the pill can introspectively and reliably detect the TIME-STAMPED beliefs.

However, we are not pigeons and, more importantly, our beliefs are not lentils. Lentils can be looked at to identify markers of quality carried by them. The

[16] That is, at least have the belief accessible.

[17] Distinguish four states of beliefs: 'conscious', 'accessible', 'inaccessible', 'absent'. Following Moon's choice of terminology, a belief is conscious at the time a subject is considering it. It is accessible if the subject could bring it to consciousness through wilful effort. It is inaccessible if there are traces of information present that cannot be wilfully accessed by the subject. Others call this 'implicit' memory. It is 'absent' if there is no information present. Moon does not discuss the latter two cases.

[18] The version of *Cinderella* recorded by the Brothers Grimm has Cinderella, who lives with her evil stepmother, faced with the task of cleaning lentils while her two evil stepsisters enjoy the king's festival. Cinderella, unexpectedly, gets helped by pigeons in sorting the lentils, allowing her to finish early, attend the festival, and meet her prince.

same is not true of beliefs. Introspectively, the property TIME STAMPED as it applies to the belief IMMUNE can only be identified by memorising a meta-belief about the belief IMMUNE. Undoubtedly, we do form relevantly similar meta-beliefs in lots of cases.[19] But the important question is whether such meta-beliefs are also introspectively reliable: can immune users rely on the meta-belief that <My belief IMMUNE is TIME STAMPED>? It is far from obvious that they can. First, coming to believe in IMMUNE would have to be a sufficiently salient event to stimulate the formation of a related meta-belief about TIME STAMPED. This might seem likely, given the potentially disastrous effects of the XX pill, but the salience also depends on the subject's estimate of the likelihood that he will later actually ingest the XX pill, about which we lack information in the XX-deflector case. Second, once he has ingested the pill, we should expect him to believe that IMMUNE is TIME STAMPED because that would allow him to continue believing the functioning of his cognition. It is thus likely that his meta-beliefs would be liable to confirmation bias (cf. Nickerson 1998). Third, it is unclear whether subjects could continuously keep IMMUNE and the meta-belief about TIME STAMPED conscious, as required by the 'strong internalist' scheme that Moon considers as a candidate position of the epistemology of epistemic defeaters. Rather, it seems likely that subjects will soon let both beliefs slip from their consciousness and thus, on strong internalism, their possession of a defeater-deflector seems at best short-lived. Fourth, on a 'moderate internalist' interpretation, we could accept that subjects are temporarily unconscious about TIME STAMPED, but the retrieval and storage of information in memory makes meta-beliefs particularly liable to corruption (cf. Schacter et al. 2011); hence, subjects' merely accessible beliefs about TIME STAMPED tend to become less reliable over continued recollection.

These are good reasons to suspect that users of the XX pill should not rely on their meta-beliefs about TIME STAMPED. But this intuition might not be shared, and I suspect that one reason for resisting this intuition is because in judging the case we note that users are, in fact, immune to the pill. In light of this fact there

[19] For instance, you believe that 'my boss told me that I have to finish the project by December 1 in our last Monday morning meeting' only after you talked to your boss during this meeting, and you will have had the meta-belief that you believe this since the last Monday morning meeting. Likewise, you would probably associate the stupefying belief that you or a relative is gravely ill with a specific moment in time.

might seem to be, from *our perspective*, no reason for the users of the pill to give up their belief in the reliability of their cognitive faculties.

However, Moon's discussion only leaves us with a choice between strong or moderate internalism. So, we have to imagine the situation as it looks from the perspective of a user of the pill. From the internalist perspective, we cannot incorporate the fact that some users are justified in relying on beliefs about their immunity while others are not. Suppose Anton is immune to the pill while Bert is not. Anton believes that IMMUNE is TIME STAMPED. But Bert does too. Bert, contrary to fact, vividly remembers someone telling him about his immunity to the pill and he recalls that he was told this just one week before he ingested the pill. The belief that <my belief IMMUNE is TIME STAMPED> will seem perfectly veridical for both Anton and Bert. Judging from their introspective capacities alone, we have no reason to suppose that their meta-beliefs are phenomenologically different and – importantly – no reason to suppose that Anton is in a better position to rely on his meta-belief than Bert is. Bert might very well believe that he is immune; he might very well remember that he spoke to a scientist who told him so. Although this is a false belief, his subjective experience will be exactly as that of the person who is, in fact, immune to the pill.

Therefore, users of the XX pill should not rely on beliefs that seem reliable *to them*. With a drug as pernicious as the XX pill, decisions about which beliefs to doubt and which to trust should be left to those who did not take the drug. This suggests how we could resolve the stalemate between Anton and Bert: by pointing out that Anton is, in fact, immune to the pill, while Bert is not. But this is certainly impossible according to strong internalism, one of Moon's proposed forms of internalism about defeaters, where we should evaluate how mental properties of the subject influence our judgement about the presence of a legitimate defeater-deflector.

Does adopting 'moderate internalism', which Moon leaves open as a possible alternative, rescue his analysis? Recall that moderate internalists hold that unconscious mental states are justificationally relevant too. In Figure 4.1, the subject is conscious of his immunity to the pill right up until $t_5$ but unconscious of it at $t_6$. Still, according to moderate internalism the subject would have a defeater at $t_6$ if the information is still present and accessible in the subject's memory. In

that case, Moon writes, moderate internalism implies that the "unconscious belief still has deflecting powers, despite its being unconscious", since it is still *accessible* (Moon 2017: 225). In a related paper, Moon suggests that *accessibility*, and not consciousness, is indeed the hallmark of an *internal* property, since he defines an internal property as internal if and only if it is "introspectively accessible to [a subject]" (Moon 2012: 347). However, the problem seems to be entirely independent of the debate between strong and moderate internalists. They might quibble over what happens at $t_6$ in Figure 4.1 above, but the real concern is with whatever happens from $t_4$ onwards.

The only way I can see of rescuing MOON'S WAY OUT is by adopting an externalist perspective on defeaters. The externalist view entails that defeaters are true propositions which need not be known to the agent in question. In that case, we can square MOON'S WAY OUT with the intuition that the subject's memory is called into question by the XX pill. The externalist perspective maintains that, as a matter of fact, he is immune to the pill throughout; hence when we who are not affected by the pill judge the case, we can rely on this information to conclude that the information about S's immunity deflects the potentially defeating information about the effects of the pill.[20]

Therefore, MOON'S WAY OUT is either false or incoherent. It is false if we accept internalism about defeaters as true because it seems that internal properties do not afford an introspective difference between subjects that are immune and subjects that are not immune in the XX-deflector case. We can make sense of MOON'S WAY OUT by adopting externalism about defeaters, but then Moon's repeated emphasis on aspects of the XX-deflector case that are relevant only from an internalist perspective seems incoherent.

We can already draw an important conclusion for moral realists. Based on the assumption that MOON'S WAY OUT is indeed the best hope for moral realists to respond to the reliability challenge, it seems that the legitimacy of third-factor accounts depend on externalism being correct.

[20] Adopting the externalist perspective on epistemic defeaters does not settle the question whether S has a defeater-deflector from his very own perspective: it merely allows us to uphold the view, in accordance with Moon, that a defeater-deflector is present from $t_3$ onwards, from some perspective, although it is still doubtful whether S, from his own perspective, may draw the same conclusion.

It gets worse for realists: in what follows I address MOON'S QUESTION and assess whether moral realists can use MOON'S WAY OUT. To do so, I assess whether moral realists are in a situation similar to the XX-deflector case. In my view, no matter how we understand the crucial feature that makes the XX-deflector case contain a deflector, that feature is not found in the case of the reliability challenge to moral realism. It appears that not even a cognition-destroying drug as in the generic XX-deflector case is a proper match and analogy for the dire epistemic situation of moral realism.

## 4.4 Realists Cannot Deflect Evolutionary Defeat

Suppose we agree with Moon that S's meta-belief about the scientist's testimony is reliable (either because we believe that externalism is correct or because we accept Moon's argument about the relevance of some information being *conscious* or *accessible*).

Even so, to take MOON'S WAY OUT in the moral case, we would have to accept another extremely controversial assumption: namely, that the scientist's testimony is reliable in the first place. In the context of the XX-deflector case, the assumption might be benign. But it gets difficult when we regard the XX-deflector case as an analogy for the reliability challenge in metaethics, where the big question is precisely *whether* there is a reliable source of information at all.

Moon seems to suggest three reasons to assume that there is a relevant source of reliable information available in the XX-deflector case. But none applies to the metaethical reliability challenge.

### 4.4.1 Token & Type

First, Moon suggests that we can distinguish between defeated and reliable *tokens* of belief-forming faculty *types*. That is, we should worry about the cognitive faculties of those who ingest the XX pill but not about the cognitive faculty of the scientist who testifies about the subject's immunity to the pill. However, the scientist uses his own cognitive capacities to form, process, and utter this judgement. In most cases, we have no reason to suspect that the scientist's belief-forming faculty is unreliable. This is because the scientist's *token* of the faculty-*type* 'cognition' is reliable (at least that is what we assume), while the tokens of the

cognitive faculty of all XX-pill users are in jeopardy. To illustrate, suppose you and I both taste sugar and agree about its sweetness. Then I give you what is commonly known as the 'miracle berry' (*Synsepalum dulcificum*), which causes sour food to taste sweet, while sweet food still tastes sweet. We then taste a white substance that could be either fine-grained sugar or lemon juice powder. In this case, where we both know that your taste faculty is disturbed, I can still rely on my taste judgement although it is based on a token of the faculty type that is disturbed in your case. There is no problem because the potentially disturbing effect is not *global*. It affects some, but not all tokens of a faculty type.

The reliability challenge, however, is very different. We are dealing with a defeater of a certain *type* of faculty, namely our moral cognition. There is no reason to suspect that only selected individuals, or tokens of certain types of belief-forming faculties, are unreliable. Instead, we must assume that all of our faculties are unreliable. In other words, the reliability challenge does not apply to tokens of faculties individuated on a personal basis, but rather to a specific type of faculty and the respective objects of the belief in general. In that case, the XX-deflector case is not analogous to the moral reliability challenge; one condition that might allow for the existence of a defeater-deflector in the XX-deflector case is not given in the moral reliability challenge.

Realists might respond that there are moral experts whose moral cognition is unperturbed by potential defeaters. However, while there might or might not be such moral experts, we should then ask *why* those moral experts have reliable moral cognition. Unless we accept the reliability of moral experts as beyond doubt, as rock bottom in our attempts to justify moral beliefs about objective moral facts, we cannot merely take their supposed reliability for granted.[21]

Therefore, individuating reliable and unreliable faculties on a token basis does not succeed in the case of the moral reliability challenge. We should not assume that certain individuals will have reliable information, as opposed to others, and so we should not expect this feature of the XX-deflector case to be analogous to the moral reliability challenge.

[21] The problem of moral disagreement might also be used to block this reply; see Enoch (2009).

### 4.4.2 Before & After

Second, Moon suggests that the deflecting information is unimpaired by the defeater because it *predates* the potentially defeating information. This suggests a distinction between information received (and memorised) *before* the reception of a potential defeater and *after*. Whether or not this criterion is ultimately convincing in determining the presence of a defeater-deflector in the XX-deflector case, it is not met in the case of the reliability challenge.

As Figure 4.1 illustrated, the subject hears about his immunity *before* ingesting the pill that has known consequences. We might say that the potentially deflecting information is in some sense stored away safely within the subject's cognition before the potentially defeating information does any damage. Again, this temporal distinction makes sense in a large number of cases that involve impaired information-processing and decision-making. For instance, when I sign a contract, it matters whether I got awfully drunk shortly before I signed or soon after. The latter should not worry us at all, legally speaking.

However, if MOON'S WAY OUT is supposed to work in the moral case, then the temporal characterisation of defeater-deflectors is unhelpful. MOON'S WAY OUT suggests that we should be on the lookout for defeater-deflectors that predate potential defeaters. But there is no prima facie reason to suppose that there are some domains in which prior-to-defeat deflection is possible and others in which it is not. While it is entirely conceivable that some situations involve appropriately timed deflectors (for example, those in which we talked to reliable informants earlier in the day) and situations in which we do not, this distinction breaks down in the case of potential defeaters that have no determinate point of reception (or ingestion). If there is a problem with our moral cognition of objective moral facts, then the issue with the purported illegitimate influence on our beliefs (for which the XX pill is an analogy) did not arise at a particular moment in time – human moral cognition would have been distorted from the very start. Many construe the reliability challenge as an evolutionary challenge, and it seems clear that we are not affected by evolutionary forces in the same way that we are by the ingestion of some drug. Moreover, there are good reasons to conceive of the reliability challenge

as an a priori challenge, and in this case, it is even clearer that it does not arise at a particular moment in time (cf. Benacerraf 1973; Field 1989).

Thus, there is no reason to suppose that we can deflect the moral reliability challenge by relying on information that predates this challenge, because the reliability challenge, in contrast to the XX pill, does not create a problem at some determinate point in time.

### 4.4.3 Local & Global

Third, Moon suggests that there is an epistemically relevant difference between the XX-deflector case and the YY-deflector case, which can be interpreted as a difference between *global* defeat of all cognitive faculties and *local* defeat of only some faculties.[22]

The YY pill, remember, destroys the reliability of the colour perception of most people who ingest it. According to Moon, the YY-deflector case does not contain a defeater-deflector, so third-factor replies fail if realists are in a situation like the YY-deflector case. The straightforward difference in the XX-deflector case is, as Moon suggests, that the XX-deflector case contains a defeater-deflector. As I argued above, this is mistaken, but even if we accept it for the sake of argument, it is not very informative as a contrast to the YY-deflector case. We have to ask *why* there is a defeater-deflector in one case, but not in the other.

The second distinguishing criterion is that the XX-deflector case contains a source of relevant, reliable information that provides the subject with (arguably) deflecting information before receiving a defeater; the YY-deflector case lacks both features.

Suppose we heed Moon's advice and take both cases as blueprints to check whether moral realists are in a situation similar to the XX-deflector case (in which they could hope, in line with MOON'S WAY OUT, to fashion a valid third-factor response) or are in a situation like the YY-deflector case (Moon 2017: 222). If the YY-deflector case were that weak, that exercise seems futile because the trite

[22] In his tentative answer to MOON'S QUESTION, Moon suggests that the XX-deflector case is the better case for realists to be in. On my interpretation, this is false: a local defeater, as in the YY-deflector case, is easier to deal with. But, as I argue in this section, we cannot interpret the reliability challenge in metaethics on the model of a local defeater.

question would then be whether the realist is in a situation that affords a defeater-deflector (as in the XX-deflector case) or in a situation where there is just plain defeat. That is the question that many scholars are currently asking, but since Moon suggests a way forward in the debate, we should construe the intended contrast as a bit more nuanced so that we understand *why* there is a defeater-deflector in one but not in the other case.

Fortunately, as Moon recognises himself, the YY-deflector case can be amended by supposing that relevant and appropriately timed reliable information about one's immunity to the YY pill exists (Moon 2017: 221). If the YY-deflector case is amended accordingly, we find that the affected subjects get testimonial evidence that they are immune to the effects of the YY pill and that their colour vision will not be impaired by it. Moon writes that his YY-deflector case is similar to cases discussed by Street (2008b: 218) and Locke (2014: 231). While the specifics of these cases do not matter here, their structures are instructive. In both Street's and Locke's case, someone has beliefs about a particular subject matter and learns that these beliefs are caused by an unreliable source; hence it seems that the subject receives a defeater for these particular beliefs. Street's and Locke's cases further entail the stipulation that there is no source of relevant information available to influence the erroneous beliefs that stem from the corrupted source.

This, however, is a crucial difference to the position in the XX-deflector case, which entails that *all* beliefs might be defeated, while the cases of Street and Locke entail that only beliefs formed through a *particular* process, source, or faculty about a *particular* topic, object, or event are defeated. To illustrate, imagine that you got all your beliefs about witches and sorcerers from the *Malleus Maleficarum*, a superstitious book written by a Dominican monk in 1486 about how to identify and deal with witches and wizards. Some believe that it influenced witch-hunting in Europe and later the US; be that as it may, we can be quite confident that it contains no truth whatsoever about witches and sorcerers. Similarly to Moon's original YY-deflector case, and the related cases found in the literature, we have a case where all beliefs about particular objects or events are formed based on a single source: in this case the *Malleus*. But we can amend all cases by imagining that there is another source of *relevant* information present. It is relevant in that it provides reliable information about the objects or events that your *corrupted*

beliefs are about. For instance, you might consult a historian, or indeed just about any sane person living in the 21st century to update your beliefs about witches and sorcerers so that they are reliable.

The XX-deflector case involves a potential defeater that affects, by stipulation, all of your belief-forming capacities, while the cases of Street, Locke, and my *Malleus* case involve potential defeaters that affect only a particular source of your beliefs.[23] We can call the former a *global* defeater and the latter a *local* defeater. The local defeater in Moon's YY-deflector case defeats one's colour vision, but not one's remaining cognitive skills. Hence, someone who ingests the pill can make use of his non-defeated faculties to marshal a defeater-deflector: he can use his memory of his friend's testimony about his immunity to the YY pill as a defeater-deflector. Once he ingests the pill, he knows that his colour perception might not function properly any more, but his memory of the scientist's testimony is clearly unimpaired: he knows that his colour vision is fine.

This turns Moon's tentative appraisal on its head: contrary to what he suggests, the XX-deflector case appears to be the difficult case for the realist, and the (amended) YY-deflector case seems to be the realist-friendly case. In this case, Moon's discussion indeed points to a crucial difference between domains where we can use defeater-deflectors and domains where we cannot: if there is *global* defeat, as in the XX-deflector case, then we cannot, but if the defeater is only *local*, then defeater-deflectors may be within reach of realists, even if the temporal- and agent-based distinctions discussed in section 4.4.1 and 4.4.2 fail.[24] The relevance of the distinction between local and global defeaters depends on three assumptions. First, that we can individuate belief-forming processes (such as a belief-forming method for colour beliefs, moral beliefs, etc.). Second, that some defeaters affect only particular belief-forming methods or sources of beliefs, such as beliefs formed by cognitive processes or beliefs formed based on the contents of a particular book. Third, that beliefs produced by one faculty or based on a particular source may provide information about the reliability of beliefs formed by another belief-

[23] Which presupposes a view according to which all beliefs are the result of cognitive processes, which means that there is no direct, i.e. non-cognitive, perception to form a belief.

[24] Global: all belief-forming faculties. Local: selected *sources* (e.g. information from this-or-that book) or selected belief-forming faculties (e.g. perceptual beliefs).

forming faculty or source. To illustrate: taking the pill that destroys your colour vision defeats your perceptual beliefs about colour, but not your memory that you are immune.

Recall that we assumed that evolutionary considerations would provide a defeater for all beliefs produced by moral cognition. Hence, realists could use the defeater-deflector for their third-factor response only if they can tap into a source of information that is both distinct from (i.e. not a product of) the deliverances of moral cognition and yet indicative of the reliability of moral cognition.[25]

## 4.5 Non-Moral Beliefs Cannot Vindicate Moral Reliability

In the remainder of this chapter, I consider whether evolutionary considerations, which are distinct from the deliverances of moral cognition, might provide information that could vindicate the reliability of moral cognition, although I ultimately reject this proposal.

The thought is as follows: even if all moral beliefs would potentially be defeated by evolutionary explanations of our moral beliefs, we might be able to defend the reliability of our moral beliefs in reference to *non-moral* beliefs that we have no reason to regard as unreliable. The blueprint for such an approach can be found in evolutionary vindications of the reliability of our beliefs in other domains of inquiry, such as epistemic beliefs (Cruz et al. 2011), beliefs about logic (Schechter 2013), or our perceptual beliefs (Boudry and Vlerick 2014). These evolutionary vindications make, roughly, the following points: evolutionary considerations show that the beliefs under scrutiny had to be *true* to be adaptive. On a representational model, this means that the relevant beliefs had to correctly represent the world. If our ancestors held true beliefs for these evolutionary reasons, then we currently hold sufficiently many beliefs, or derivatives of beliefs, whose truth is plausibly seen as adaptive. So, our belief-forming methods in the relevant domain are plausibly regarded as reliable.

Now, I do not wish to discuss the merits of these accounts here. What I want to point out is that if we regard considerations about the adaptiveness and truth

[25] Even though the idea that moral beliefs might be justified through non-moral beliefs appears to be quite a common one. Shafer-Landau, for instance, discusses and rejects it in relation to Hume's open question argument; see Shafer-Landau (2004: 121ff).

conditions of moral beliefs as non-moral considerations, then realists might gain independent evidence, similar to the scientist's testimony in the XX-deflector case, about the reliability of our moral beliefs. Realists would be on their way out of the reliability challenge.

However, such an evolutionary vindication is problematic and unlikely to succeed in the moral case, for at least two reasons. First, moral realists typically do not aim for an empirical vindication of their accounts, nor do they accept the relevance of an empirical vindication of their claims (e.g. Enoch 2011b). On the contrary, the moral realists with whom I am concerned in this chapter argue that moral properties are of an altogether different sort than natural properties. Hence, while realists might hold that our beliefs *about* moral properties might have played an evolutionary role, they would not require that moral properties themselves played any relevant role in our evolutionary history. It is a strange development (and one which somewhat betrays their theoretical commitments) that moral realists should nonetheless try to come up with reasons for thinking that moral properties played a causal role in human evolutionary history.

Second, and more importantly, realists would have to show that evolutionary considerations vindicate the evolutionary relevance of the *truth* of our moral beliefs. But this seems unlikely to be successful because the truth of our moral beliefs was unlikely to have played an evolutionary role. To see this, consider that claims to the effect that *true beliefs produced by faculty F were adaptive* imply three points. First, that beliefs formed by faculty F generally represented the world to be in a certain way; second, that the probability of certain actions increased due to these beliefs; and, third, that the combination of the representation of the world and the ensuing reaction was such that it conferred an evolutionary advantage to our ancestors. Thus, actions, not beliefs, are the primary factor in considerations about adaptiveness. The general point is that the content of beliefs matter in evolutionary explanations insofar as we can show that certain world-to-belief relations are more likely to lead to adaptive actions than others. The implications for moral realists is that beliefs matter in evolutionary explanations insofar as they might be able to show that, more often than not, *true* representations of the world were more likely to lead to adaptive actions.

However, actions purportedly guided by moral beliefs, such as speaking the truth or taking care of one's offspring, appear to be adaptive irrespective of whether their related moral beliefs correctly represent the world (cf. Gibbard 2003: ch. 13; Street 2006; Joyce 2006).[26] In other words, there is no need to claim that our moral beliefs were true, i.e. that they correctly represented the moral facts, to explain their potential evolutionary adaptiveness. While this does not imply that our moral beliefs are *false,* neither does it imply that evolutionary explanations show that our moral beliefs are *true*.

What's worse for realists, as long as it is possible to account for the adaptiveness of a moral belief without referring to its truth as conceived of by realists, their intended evolutionary vindication of our moral beliefs cannot succeed. Evolutionary considerations might show that realism is *compatible* with evolutionary explanations of our moral beliefs, but they do not vindicate it. Hence, evolutionary considerations about the adaptiveness of our substantive moral beliefs do not give us reasons to believe that our moral beliefs are reliable.

The problem persists for evolutionary considerations about the adaptiveness of our metaethical beliefs, which is another way for realists to approach the task of using non-moral beliefs to vindicate their moral beliefs. The thought might be that our beliefs about, say, the nature of moral properties make a behavioural difference only if they are true. The first step in the argument resembles familiar evolutionary explanations of beliefs in absolute objective standards. Such beliefs arguably turned humans into better cooperators (cf. Tomasello 2016). Imagine two people; one believes that 'cheating is wrong because it gets you into trouble if you get caught', while the other believes that 'cheating is morally wrong because one should under no circumstances do it'. Plausibly enough, we can imagine that the beliefs of both agents affect their individual behaviour in different ways. Perhaps the one worried about getting caught will cheat if there is no way of getting caught, while the other, holding a characteristically 'objectivist' conception of morality, might avoid cheating even if there is no chance of getting caught. These

[26] Realists need not accept representationalism, according to which our moral beliefs purport to represent the world, in which case they might have other ways around the reliability challenge (cf. Bogardus 2016). Most proponents of the form of moral realism relevant in this chapter, however, do accept a form of representationalism; see Enoch (2011b), Shafer-Landau (2003), and Wielenberg (2014).

considerations suggest that having objectivist metaethical beliefs might matter in an evolutionary sense. Albert Camus' titular character in *The Stranger* is a case in point: the Stranger thinks value is only ever contingent and, surely enough, partly because of this view he ends up with his head under the guillotine before he had time to procreate. Hence, there seems to be a good case for arguing that different metaethical beliefs might have produced actions of differing adaptiveness and, in particular, that characteristically objectivist metaethical beliefs lead to more adaptive behaviour.

However, apart from the fact that the empirical case to back up this intuition is harder to make (for instance, there is little evidence that beliefs in intrinsic values promote cooperative action; reputational effects are more reliable predictors of proxies for moral behaviour (see Haley and Fessler 2005), so those metaethical beliefs would be adaptive irrespective of their truth. It *may* turn out that moral objectivists act in ways that increase their relative fitness. But, as in the case of substantive moral beliefs, this is compatible with the claim that these beliefs are all false. Indeed, that is the core point of some debunkers (see Ruse 1998): having objectivist moral beliefs would be adaptive even if those beliefs were false (understood on a correspondence model of truth, (see Joyce 2016c: 154ff). Therefore, it seems that using non-moral beliefs to deflect defeaters of our moral beliefs does not provide a way out of the reliability challenge for moral realists either. In that case, the answer to MOON'S QUESTION is that moral realists are not in a situation relevantly similar to the XX-deflector case; hence they cannot take MOON'S WAY OUT.

## 4.6 Concluding Remarks

I carried on with Moon's project to make progress in the debate about the viability of third-factor responses to the metaethical reliability challenge. Moon suggests, through the principle that I called MOON'S WAY OUT, that defeater-deflectors are a valid response to the reliability challenge. He leaves open whether moral realists can make use of third-factor replies so understood.

I argued that we should believe that there is a defeater-deflector in the case discussed by Moon only if externalism is true. Next, I argued that the conditions for a defeater-deflector are not given in the situation of the moral realist, even if

we assume that externalism is true. MOON'S WAY OUT does not work, if it works at all, for moral realists. Hence, moral realists cannot deflect defeat from evolutionary explanations of morality.

For my investigation of Moon's interpretation of the third-factor response, I assumed two points: first, that the defeating power of the reliability challenge is similar to the defeating power of Moon's XX pills. Second, that Moon's arguments against two alternative interpretations of the third-factor account are sound. Realists might challenge both assumptions, either by coming up with alternative takes on the third-factor response or by showing that the reliability challenge does not give us an epistemic problem that is on a par with cognition-destroying pills. So all is not lost for moral realists.

However, if my assumptions hold true, then the considerations offered in this article suggest that the third-factor response in the moral case is unlikely to succeed. What might have seemed like a way out of the reliability challenge turns out to be a bad moon rising for moral realism.

This page intentionally contains only this sentence.

# 5 Defeat by Disagreement

## Reader's Guide

Disagreement is of troubling epistemic potential. In particular, *learning* about a disagreement with people you trust and whose opinion you respect may sometimes give you good epistemic reasons to re-examine your beliefs. Consider the following internal monologue (inspired by Feldman and Warfield (2010):

> I think that human-caused global warming is real. My colleague disagrees. Furthermore, my colleague has examined all the same information as I have, knows as much as I do about the issue, and is as well trained as I am. This worries me. I trust his opinion and would welcome it if he would agree with me on the matter. Is it reasonable for me to retain my belief in light of this disagreement? Or is some adjustment rationally required?

A popular and widely discussed answer to the question raised in the monologue is that revealed peer disagreement (to wit, disagreement between epistemic peers that is acknowledged by at least one peer) rationally requires adjustment about the disputed proposition (Christensen 2007, 2011; Frances 2014; Weatherson 2013). In particular, it is often held that the disputants in the revealed disagreement have epistemic reason to withhold belief about the disputed belief (Elga 2007). So, on this view, some disagreements give rise to undercutting defeat. Might the epistemic significance of disagreement offer a way to raise the evolutionary defeat challenge? I introduced this view as the 'disagreement view' in the main introduction.

In this chapter, I will assess the disagreement view in depth as a substantive proposal for how evolution might undercut all moral beliefs. We have seen in section 1.6.3 of the first chapter that the disagreement view is supposed to be a strong contender for showing how to get from evolutionary explanations of morality to a justificatory loss. If the disagreement view would succeed, it would be a boon for defenders of the evolutionary defeat challenge, especially in light of the

previous discussion. I argued in the previous chapter that if the evolutionary defeat challenge succeeds, the justification of all moral beliefs would be lost and their justification cannot be recovered. But why think that learning about evolutionary explanations of morality raises a defeater in the first place? Thus far, we have seen only problems for this view. We have seen that empirical information alone does not carry the day (in chapter 3) and that existing accounts of undercutting defeat do not imply that moral beliefs are undercut either (in chapter 2). Moreover, when discussing specific proposals about the epistemic principle that gets us from evolutionary explanations of moral judgements to defeat of all moral judgements, such as the claim that our moral beliefs are shown to be coincidentally true and thus defeated, we saw that many prominent accounts did not work (in section 1.6 of chapter 1). It is now time to take up the disagreement view that might yet rescue the 'survival of defeat'.

Several philosophers have suggested that evolutionary explanations of morality are epistemically significant insofar as they raise a problem of counterfactual disagreement. The basic idea behind the disagreement view is that learning about evolutionary explanations of morality shows us that we could have easily endorsed moral judgements that conflict with those that we presently hold. Insofar as this form of disagreement is undercutting, our moral beliefs are undercut. Recall figure 1.1 from chapter 1, which illustrated the two central questions in our quest to determine the survival of defeat. I reproduce the figure here:

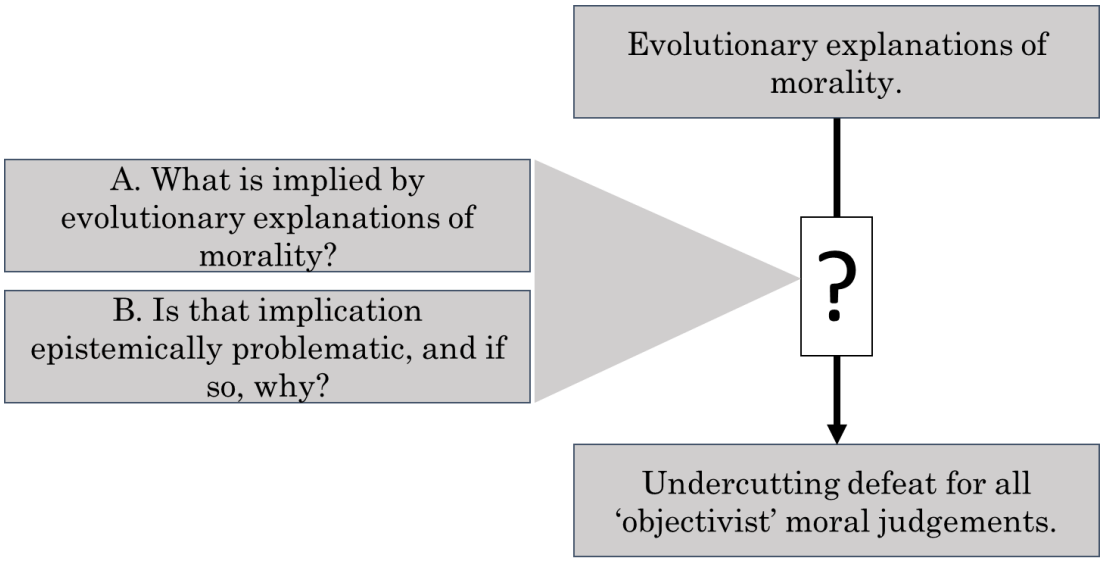


*Figure 5.1 How can evolution defeat objectivist moral judgments? (replicated)*

I will argue that the disagreement view fails: evolution does *not* show that there is undercutting counterfactual disagreement about all our moral beliefs. Hence, I will conclude that the disagreement view does not help us to see how evolutionary defeat could be saved. Before turning to my argument, I provide more details about the epistemic significance of disagreement in general by explaining why I think that the notion of peer-hood is crucial for evaluating the epistemic significance of a given disagreement and how evolution might be relevant in fashioning an argument from disagreement (to wit, an argument which has a premise that refers to the epistemic significance of disagreement).

Many epistemologists distinguish *radical disagreement*, as a form of epistemically significant disagreement, from *peer disagreement* (Wedgwood 2010). A disagreement is radical if it cannot be rationally resolved. Radical disagreement is thought to be a different problem from peer disagreement for two main reasons. Peer disagreement places emphasis on the relation of the disputing parties towards the disputed issue, while a case of radical disagreement may arise independently of the epistemic properties of the disputing parties: it might be that, in virtue of the disputed issue, it could not rationally be settled. Some authors then argue that some instances of peer disagreement are radical disagreement, because peer disagreement does not allow the disputing parties to rationally settle the issue (so both parties have to stick to their guns, or withhold judgement about the disputed issue, rather than settling it).[1] Consequently, insofar as it is rationally permissible to settle or ignore a peer disagreement, it will not count as radical. Most discussions of moral disagreement focus on *radical* disagreement, implicitly restricting their focus to disagreement amongst beings whose moral views matter (Decker and Groll 2013). The implicit assumption seems to be that, in interesting cases of moral disagreement, the disagreeing parties are both equally likely to get things right in moral matters and thus that they are what might be called 'moral peers' and their peerhood is what makes their dispute epistemically significant for them. That such an implicit assumption is present in most discussions of moral disagreement seems evident when you consider that the egoistic tendencies of young children, which are prevalent up to a certain stage of moral development

[1] See Christensen (2007); Goldberg (2013); Kelly (2010).

and which conflict with many considered moral judgements in adults, are not widely thought to constitute an epistemically significant disagreement (Klenk 2017b; Kohlberg and Hersh 1977). A straightforward explanation is that young children are not our 'moral peers' in the relevant, epistemic sense (which is of course independent of whether they are morally relevant more generally). Peerhood in related discussions of non-moral disagreement is also used as a marker of epistemic relevance: it is primarily for that reason that disagreements between evolutionary biologists and creationists, physicists and astrologists, and physicians and charlatans can be regarded as epistemically insignificant (Elga 2010; McGrath 2008, 2011). Peerhood, therefore, seems crucial for evaluating the epistemic significance of disagreement.

The disagreement challenge and the evolutionary defeat challenge that I am concerned with might indeed profit from "considering these challenges side-by-side" (Bergmann and Kain 2014: 1). Let me point out one way in which this might be done. Evolutionary explanations of morality may provide empirical support for an argument from disagreement that has been widely discussed in the literature but often lacked adequate empirical support (Brandt 1944, 1954; Decker and Groll 2013; Doris and Plakias 2008; Gowans 2000; Sinnott-Armstrong 2014; Tersman 2006). The argument from disagreement that I have in mind begins with the apparent widespread disagreement about moral beliefs and claims that the disagreement is rationally irresolvable. It then claims that rationally irresolvable disagreement is incompatible with moral objectivism.

Two problems with raising this challenge are commonly noted: first, distinguishing apparent from real disagreement and, second, establishing that a given disagreement is rationally irresolvable.[2] Consider the problem of distinguishing between apparent and real disagreement. As objectivists point out, although there might be variation in specific moral codes, basic moral codes needn't be disputed too (Mackie 1977: 47). There would be some disagreement, but not about the basic codes. Furthermore, many actual moral disagreements can be explained by the fact that one party to the dispute is irrational, biased, or misinformed about the non-moral facts that underlie the dispute (Tersman 2006).

[2] Which is not to say that this is the only issue regarding raising an argument from disagreement; cf. Tersman (2006).

In this case, the disagreement would be merely apparent and of no epistemic import. One of the problems for an argument from disagreement is thus to establish that there is actual disagreement about a range of moral issues.

Evolutionary explanations of morality might assuage this problem by making it likely that the content of our moral beliefs at least partly reflects human phylogenetic history. If that is correct, then evolutionary explanations of morality provide reasons for thinking that we *might* disagree with a counterfactual human species that took a different evolutionary path and thus developed moral beliefs with different content compared with ours. This might help for the empirical case of the disagreement challenge. Insofar as we are convinced that counterfactual disagreement might be epistemically relevant, evolutionary explanations of morality might thus support a disagreement argument that has played an important role in discussions of moral objectivity (Harman 1977; Loeb 1998; Mackie 1977; Tersman 2014).

The defence of the disagreement view that I consider in this chapter might be called the *direct* route to defending this view. In the next chapter, I will consider another, *indirect* way in which epistemically troubling disagreement might play a role in undercutting all objectivist moral beliefs. Some proponents of the disagreement view defend a stronger thesis by explicitly endorsing the view that evolutionary explanations lead to defeat *if and only if* they reveal an epistemically significant counterfactual disagreement. I call this stronger version of the disagreement view the 'debunking-disagreement thesis'. If that view were right, then of course my rejection of the disagreement view in this chapter would imply that evolutionary defeat fail. As will become clear in chapter 7, however, I think that the debunking-disagreement thesis is ultimately mistaken. The following graph illustrates the structure of my argument in this chapter. Arrows stand for a 'counting-against' relation and can be directed either at theoretical positions or other 'counting-against' relations (indicated by an arrow pointing to another arrow).

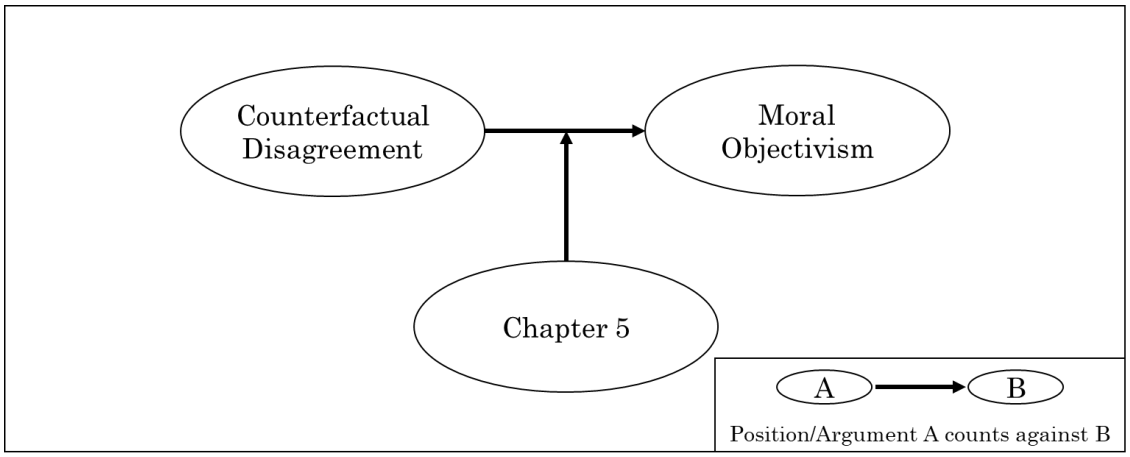


*Figure 5.2 The argument against the disagreement view (direct route)*

Let's turn to the argument against the disagreement view next.

## Abstract[1]

Several philosophers have recently argued that evolutionary considerations undermine the justification of all objectivist moral beliefs by implying a hypothetical disagreement: had our evolutionary history been different, our moral beliefs would conflict with the moral beliefs of our counterfactual selves. This chapter aims at showing that evolutionary considerations do not imply epistemically relevant moral disagreement. In nearby scenarios, evolutionary considerations imply tremendous moral agreement. In remote scenarios, evolutionary considerations do not entail relevant disagreement with our epistemic peers, neither on a narrow nor on a broad conception of peerhood. In conclusion, evolutionary considerations do not reveal epistemically troubling kinds of disagreement. Anti-objectivists need to look elsewhere to fuel their sceptical argument.

## 5.1 Introduction

The burgeoning debate about the metaethical implications of the Darwinist view of morality[2] focuses on which epistemic principle(s) allegedly support debunking arguments against moral objectivism (e.g. Clarke-Doane 2015; Lutz forthcoming; Sinclair forthcoming; Vavova 2015). Moral objectivism is the view that (at least some) moral truths are metaphysically necessary as well as constitutively and causally independent of human attitudes or beliefs (e.g. Enoch 2011b; Shafer-Landau 2003).[3] Though objectivists must, of course, explain how objectivist moral beliefs can be justified in the first place, a central question is whether objectivist moral beliefs can be undermined, assuming that they are at least prima facie justified.

So, what is that 'something' in virtue of which a Darwinist view of morality creates a problem for objectivist moral beliefs? It has been claimed that evolutionary explanations of morality show that moral beliefs are prone to error (Vavova 2015), fail to be modally secure (Clarke-Doane 2015), or that the best

[1] This chapter is based on a paper that is currently under review.

[2] 'The Darwinist view of morality' is shorthand for 'providing an evolutionary explanation of morality'; this will be specified further below.

[3] Objectivism as defined here is not as outlandish a view as it might initially seem. All so-called robust realists explicitly defend it and, arguably, relaxed realists, such as Scanlon (2014), and some naturalists, such as Jackson and Pettit (1995), face a similar challenge, though I cannot argue for this claim here.

explanation of moral beliefs does not entail that they are (mostly) true (Joyce 2016b). None of these theses has found widespread support.

In light of this controversy, a new thesis is quickly gaining currency. A number of philosophers have argued that a Darwinist view of morality is metaethically significant because it shows that moral beliefs are, or could be counterfactually, subject to disagreement (Bogardus 2016; Joyce 2018; Mogensen 2016a, 2017; Sinclair forthcoming; Tersman 2017; White 2010). So, a Darwinist view of morality could yet play a metaethical role if it piggybacks on the epistemic significance of disagreement.[4]

For example, Mogensen writes that any metaethical implications that follow from a Darwinist view of morality "will be due to the epistemic significance of moral disagreement" (Mogensen 2016a: 591). The disagreement in question is *hypothetical* or *counterfactual* disagreement: had our evolutionary history been different, our actual moral beliefs would conflict with the moral beliefs of our counterfactual selves.[5] The consequence of this counterfactual moral disagreement is that the justification of all affected moral beliefs (objectively construed) is undermined, or so these philosophers argue.[6] Let this be the disagreement view:

> **Disagreement View**: Evolutionary explanations of morality imply that there is justification-defeating counterfactual disagreement about all moral beliefs (as conceived of by moral objectivists).[7]

---

[4] Further suggestions about the relevance of disagreement for debunking, though not explicit endorsements, are provided by Street (2016: 314f), Clarke-Doane (2015: 100), Ballantyne (2013: 246–54), and Horn (2017).

[5] According to the ordinary understanding of disagreement, disagreement requires actual disputants and actual disputes. For example, one does not disagree about household chores if one's partner is, perhaps, merely lazy. Thus, on that understanding, whatever is implied by the evolutionary hypothesis seems far removed from disagreement. The relevant idea, however, is that some imaginary disagreements could easily be actual, in which case learning about them seems epistemically significant.

[6] Not all proponents of the disagreement view endorse the view that the metaethical status of moral beliefs is relevant for the challenge. For example, Mogensen (2016a) denies it. My construal of the challenge is more conservative, so I think he could accept my view (without thereby acknowledging that the challenge applies *only* on the assumption that moral objectivism is true).

[7] I understand the relevant type of disagreement as follows: we believe that p, where p is some moral proposition, while our counterfactual selves believe that not-p, where the contents of p are the same for us and our counterfactual selves. I further specify this in section 5.3.

Moreover, some defenders of the disagreement view defend an even stronger version, according to which evolutionary explanations of morality are epistemically significant *if and only* if they reveal such counterfactual disagreement (Bogardus 2016; Mogensen 2014, 2016a):

> **Debunking-Disagreement Thesis**: Evolutionary explanations of morality are epistemically significant if and only if they imply actual or counterfactual justification-defeating disagreement about all objectivist moral beliefs.

The disagreement view rests on two noteworthy claims about the epistemology of disagreement. First, it rests on the claim that counterfactual disagreement about a belief that p might defeat the justification for holding the belief that p. Second, it rests crucially on a concessive view about disagreement (as explained below). The concessive view is controversial (e.g. Enoch 2010). For the purposes of this chapter, however, I will assume that the concessive view is true.

I aim at showing that the disagreement view is false. One route to attacking the view would be to deny that hypothetical disagreement is epistemically significant (Kelly 2005; Tersman 2013). There are good reasons, however, not to place too much weight on the actual/possible distinction in arguments about disagreement.[8] Instead, my argument focuses on the concept of epistemic peer-hood, a rather underexplored issue in recent epistemology and uncharted territory in relation to evolutionary debunking arguments in metaethics.[9] My strategy is to show that evolutionary explanations of morality do not reveal *epistemically significant disagreement* about morality. So, the disagreement view is false. Moreover, if the debunking-disagreement thesis is true, which I doubt but cannot assess in this chapter, then this chapter implies that evolutionary explanations of morality are epistemically insignificant. Independently of that claim, this chapter speaks to what we can and cannot learn about counterfactual moral disagreement from evolutionary considerations. These findings should be of interest to both moral objectivists and their critics. Section 5.2 clarifies the context and the

[8] My main worry is that drawing the actual/possible distinction will depend on counterfactual analyses to explain when possible but absent disagreements are significant, and counterfactual analyses have a bad track record in philosophy.

[9] King (2012) and Gelfert (2011) are notable exceptions.

metaethical significance of the disagreement view. Section 5.3 reconstructs the argument for the disagreement view in greater detail. Sections 5.4 introduces my argument against the disagreement view and sections 5.5 and 5.6 defend the two horns of the dilemma for the disagreement view. I conclude in section 5.7.

## 5.2 Counterfactual Disagreement and Evolutionary Defeat

Evolutionary explanations of morality maintain that the *capacity* for normative guidance or the *content* of at least some of our most fundamental moral beliefs or inclinations is a product of the human evolutionary history (Joyce 2006, 2016b; Street 2006, 2016). For example, bravery appears to be evolutionarily useful, and it is evaluated positively. So, it stands to reason that the positive (moral) evaluation of bravery has an evolutionary origin (Curry 2016). Thus:

> **Evolutionary Hypothesis**: Quite some human moral beliefs are the product of human evolutionary history.

For this chapter, quite a few significant and controversial issues about the evolutionary hypothesis have to be swept under the rug.[10] That is alright, however, because virtually all discussants in the metaethical debate accept two corollaries of the evolutionary hypothesis. First, that the evolutionary determinants of our moral beliefs are *contingent*. Had human evolutionary history been different, human moral beliefs would have been different.[11] Second, that both objectivists and their opponents accept that the objective moral truths were *causally irrelevant* in the evolutionary genesis of our moral beliefs (Enoch 2010; Street 2006). The evolutionary hypothesis, with its two corollaries, provides the basis for so-called evolutionary debunking arguments.

It is imperative to distinguish carefully between different types of evolutionary debunking arguments. First, evolutionary debunking arguments about morality differ in scope. Those of *global* scope target all beliefs of the moral *type*, as opposed to local arguments, which target subtypes (e.g. all moral beliefs moderated by disgust reactions) or tokens (e.g. the belief that torture is wrong).

[10] Cf. chapter 1.

[11] Recall that the claim is an idealisation. As such, it is quite probably false. See Clarke-Doane (2014), who denies the counterfactual. In the present chapter, nothing substantial turns on whether or not the counterfactual is true.

Second, amongst the arguments of global scope, *ontological debunking arguments* proceed by way of Ockham's razor and try to show that sui generis moral facts are explanatorily idle and that they should, therefore, be purged from our ontology (e.g., Ruse and Wilson 1986). This kind of debunking argument is reminiscent of Mackie's argument from queerness, which tries to accomplish as much not via Ockham's razor but via an argument to the effect that, regarding naturalism, objective moral values would be too bizarre to take seriously (Mackie 1977).

In contrast, proponents of the disagreement view are concerned with *epistemological debunking arguments*. Epistemological debunking arguments can be distinguished based on whether their proponents regard the target of the argument as optional or not. Joyce (2016b) argues that morality has some essential features and that the evolutionary hypothesis defeats the justification of all moral beliefs thus understood. Street (2006) argues that the evolutionary hypothesis defeats the justification of all moral beliefs *if* moral objectivism is assumed. But Street is not committed to the truth of objectivism. Despite this difference in regard to the optionality of the target, both variants of the epistemological debunking argument function alike in the following sense: given a certain understanding of morality, the evolutionary hypothesis undermines the justification of all moral beliefs thus understood.

Following the proponents of the disagreement view, the focus of this chapter is on those variants of evolutionary debunking arguments that aim to conclude that all objective moral beliefs are unjustified.

## 5.3 Clarifying the Disagreement View

This section reconstructs the argument for the disagreement view. I clarify each premise as we go along. Mogensen's and Bogardus's defences of some steps differ in the details, but they ultimately reach the same conclusion.

First, they argue that the evolutionary hypothesis implies that our counterfactual selves might have had different moral beliefs from us:

> [H]ad our species evolved elsewhere,—as easily might have happened—and we later formed moral beliefs using the same method we actually used, our beliefs may easily have been incompatible with our actual beliefs, false by our own lights. (Bogardus 2016: 656)

> There is reason to suppose that the moral intuitions of human beings reflect our place on the tree of life: had the conditions for the evolution of moral thought been realized in some distantly related species, their moral outlook would most likely incorporate certain fundamental differences in moral intuition, appropriate to their form of life. (Mogensen 2016a: 607)

Both quotes reflect the idea that the evolutionary hypothesis implies the contingency of at least some of our moral beliefs. To make that idea more precise, let $M_{Actual}$ be the set of moral propositions whose members are our moral beliefs, where 'our' refers to US, the set of all human beings that live or lived in the actual world. Let $M_{Counterfactual}$ be the set of moral propositions believed by THEM, where THEM is the set of all human beings that live or lived in a counterfactual evolutionary scenario.[12] $M_{Counterfactual}$ could, therefore, be different from $M_{Actual}$, or so the evolutionary hypothesis implies.

Second, proponents of the disagreement view claim that the divergence of $M_{Actual}$ and $M_{Counterfactual}$ amounts to hypothetical *disagreement* with our counterfactual selves. Proponents of the direct approach seem inspired by Darwin's famous thought experiment:

> If men were reared under precisely the same conditions as hive-bees, there can hardly be a doubt that our unmarried females would, like the workerbees, think it a sacred duty to kill their brothers, and mothers would strive to kill their fertile daughters, and no one would think of interfering. (Darwin 1871 [2004]: 70)

So, $M_{Counterfactual}$ might radically conflict with the members of $M_{Actual}$ (Mogensen 2016a: 607), and we would find the moral beliefs of THEM "false by our own lights" (Bogardus 2016: 656).

Third, the hypothetical disagreement *with our counterfactual selves* is epistemically significant. Our moral beliefs are based on the same type and quality of evidence, which, Bogardus claims, shows that there is an epistemically significant "evidential *symmetry*" between us and our counterfactual selves.

[12] Counterfactual evolutionary scenarios raise the problem of deciding whether THEY are still human, even on very distant evolutionary paths. It is also not obvious whether THEY are still *our* counterfactual selves, even on very distant evolutionary paths. However, I would defend proponents of the disagreement view here by saying that the real issue is a disagreement in moral belief (see footnote 8 of this chapter), not about whether the disagreement is with *Homo sapiens* or another species, or whether it is really a version of US that we are disagreeing with (thus setting aside issues about identity).

Mogensen, in contrast, takes the evolutionary hypothesis to show that we and our counterfactual selves have *different* evidence, which shows that there is an evidential *asymmetry* between us and our counterfactual selves. Such asymmetry is epistemically significant nonetheless because the moral disagreement implied by the evolutionary hypothesis bottoms out in pure conflicts of intuition (Mogensen 2017: 286ff). So, both Mogensen and Bogardus suggest that there is a systematic conflict between US and our counterfactual selves.

The fourth step in the argument for the disagreement view is to note that *the hypothetical disagreement* with our counterfactual selves is, all else being equal, as epistemically significant as actual disagreement. Bogardus qualifies this by saying that the hypothetical disagreement is "near enough to cause [epistemic] trouble" such that had we run a different evolutionary course, we would have *easily* ended up disagreeing with our counterfactual selves (Bogardus 2016: 657). To clarify the relevance of *hypothetical* disagreement, suppose that you base your belief about the median fertility rate in 2015 in Azerbaijan on the cast of a die. Whatever you end up believing, the basis for your belief might easily have been different, and the mere fact that you *actually* believe that the number is 6 is of no particular epistemic relevance. Relatedly, Mogensen's argues that non-actual disagreement whose absence is caused by an epistemically irrelevant factor is as epistemically significant as actual disagreement (Mogensen 2016a: 598-601).[13] Consider your belief about the fertility rate in Azerbaijan again. Suppose you resolve that the median number is 6. If a mysterious pest had killed off any expert who maintained, correctly, that the answer is 2 there would be no disagreement. Still, learning about the arbitrary absence of the disagreement should give you epistemic pause. Hence, to paraphrase, hypothetical disagreement is relevant if it could be *easily present* (Bogardus) or if it is *arbitrarily absent* (Mogensen).

Fifth, the correct response to disagreement is to withhold judgement about the disputed belief (on the assumption that some counterfactual disagreements are as epistemically significant as actual disagreement, as the previous point suggested). This claim is reminiscent of a *concessive* view about disagreement (Bogardus 2016: 656; Mogensen 2016a: 607). Concessive views imply that

[13] See Kelly (2005: 181) for a similar suggestion.

intractable disagreement amongst interlocutors of comparable epistemic standing undermines the justification of the disputed belief, provided that there is no *independent* evidence in favour of either of the disputed beliefs (Elga 2007; Feldman 2006; Sidgwick 1981 [1874]). The independence principle at the heart of such views implies the following: "[I]n evaluating the epistemic credentials of another's expressed belief about P, in order to determine how or whether to modify my own belief about P, I should do so in a way that doesn't rely on the reasoning behind my initial belief about P" (Christensen 2007: 198, 2011: 1; Elga 2007: 489). To repeat, I will assume for the sake of argument that some such independence principle is valid.

In conclusion, the evolutionary hypothesis implies that there is justification-defeating disagreement about $M_{Actual}$. Given a concessive view about disagreement, and in the absence of independent evidence in favour of $M_{Actual}$, we should give up our belief in $M_{Actual}$.

Of course, objectivism as a metaphysical thesis would still stand. Nonetheless, virtually every objectivist is in fact committed to the possibility of moral knowledge, and so the conclusion of the disagreement view would be a truly devastating result for their view (Enoch 2011b: 166; Shafer-Landau 2003).

## 5.4 The Argument Against the Disagreement View

It is time to introduce the argument against the disagreement view:

**P1** For any domain D, hypothetical disagreement is epistemically significant (to wit, relevant for epistemic justification about D-beliefs) only if it is about a proposition relevant for D and between epistemic peers about D in the narrow or broad sense of peerhood.

**P2** In non-actual nearby scenarios, the evolutionary hypothesis does not imply hypothetical disagreement about objectivist moral beliefs.

**P3** In non-nearby scenarios, the evolutionary hypothesis does not imply hypothetical disagreement with epistemic peers, in the narrow or broad sense, about objectivist moral beliefs.

**P4** So, the evolutionary hypothesis does not imply epistemically significant disagreement in either nearby or remote scenarios.

**C** So, the evolutionary hypothesis does not imply epistemically significant disagreement.

The argument is deductively valid. P1 specifies two necessary conditions for the epistemic significance of disagreement. P2 and P3 state that neither non-actual nearby nor non-actual non-nearby scenarios exhibit epistemically significant disagreement. 'Nearness' is a notoriously vague notion. I do not expect to offer a fully satisfactory account in this chapter. For present purposes, nearby scenarios are those in which our counterfactual selves resemble the members of human societies on the ethnographic record (incidentally, this also implies closeness in time; see Curry 2016). Non-nearby scenarios are those that depart in more or less extreme ways from the known ethnographic record. P4 is a filler premise required for validity and will not be discussed further. The argument's conclusion implies that DD1 is false and that, therefore, the argument for the disagreement view fails.

Before turning to defending the premises, three clarifications about the disagreement view are in order. First, neither Bogardus nor Mogensen specifies the nature of disagreement, so I suggest understanding disagreement about p as follows:[14]

**Disagreement:** There is disagreement about p iff there exists a p such that

(a) S1 believes that p and S2 believes that ~p or S1 believes that p and S2 suspends judgement on whether p.

(b) S1 and S2 have the same understanding of p.

Condition (a) is standard (Kölbel 2004: 54). Condition (b) is sensible to preclude problems with merely apparent disagreement that turns out to be a sort of confusion of tongues that plays a role in the disagreement with our counterfactual selves (Tersman 2006: 22ff).

[14] I leave out complications about differences in credence regarding the disputed proposition between interlocutors. As far as I can see, nothing substantial depends on it in this article. Also, since I assume objectivism I will not discuss possible objections by non-cognitivists, who will reject this account of disagreement because they maintain that S1 and S2 may assign different meanings to p and nonetheless genuinely disagree in their conative attitudes; see Blackburn (1984), Heiphetz and Young (2017), Hare (1963), and Stevenson (1963).

Second, neither proponent of the disagreement view explicitly puts their argument in terms of peer disagreement. Nonetheless, both appeal to cases in which our counterfactual selves appear to be our epistemic peers in the minimal sense that their moral beliefs matter for the evaluation of our own epistemic standing in regard to morality.

Bogardus emphasises, as we have seen above, the "evidential symmetry" between us and our counterfactual selves. This affords the interpretation that he accepts what might be called a narrow conception of epistemic peerhood, which can be understood as follows:[15]

> **Peerhood – Narrow Conception**:[16] S1 and S2 are epistemic peers in regard to p iff S1 and S2 are equals regarding their evidential possession and their evidential processing with respect to p.

Mogensen, in contrast, does not think that we and our counterfactual selves share equal moral evidence. Instead, he thinks that we should treat the moral intuitions of our counterfactual selves as equally likely to be a good guide to the truth (Mogensen 2017: 294ff).[17] This sits very well with what might be called a broad conception of epistemic peerhood, which can be understood as follows:[18]

> **Peerhood – Broad Conception**: S1 and S2 are epistemic peers in regard to p iff S1 and S2 are equally likely to be right about p.

Neither specification of peerhood is fully satisfactory *as a specification of peerhood*. For example, even on a narrow conception, a full specification of peerhood would doubtlessly require further conditions, such as "similar openness to experience" (Frances 2010; Gelfert 2011; King 2012).[19] In the present context,

---

[15] Cf. Cohen (2013: 98); Feldman and Warfield (2010: 2), Kelly (2005: 174–5, 2013: 34), King (2012: 252ff); Matheson (2015); Wedgwood (2010: 226).

[16] We can distinguish between acknowledged peer disagreement and non-acknowledged peer disagreement Kelly (2005: 168); King (2012: 261). In line with an internalist account of defeat (see chapter 1), awareness of the disagreement is required to have an effect on justification. In the definition of peerhood, however, we can leave out this criterion.

[17] Cf. Wedgwood (2010: 241f).

[18] E.g. Vavova Elga (2007); Frances (2014); (2014b: 308).

[19] There are different ways to spell out what it means for someone to be 'equally likely to be right' about some issue. According to Elga (2007), your epistemic peers are those with whom it is rational for you to have the same conditional probability, on the

however, my concern is not so much with a fully accurate specification of the concept of peerhood, but rather with the fixation of our ideas about which interlocutors the proponents of the disagreement view consider to be epistemically relevant. As such, less strict criteria for epistemic peerhood benefit the proponents of the disagreement view, since it would be easier for them to show that there is peer disagreement *on either such conception*.[20]

Third, I will assume, in line with the concessive view, that the burden of proof concerning peerhood is on those who want to *deny* that a particular disagreement is between peers (cf. King 2012: 249). Both Bogardus (2016) and Mogensen (2017) hint at this conception. Thus, perhaps all that is needed for a disagreement between us and our counterfactual selves to be epistemically relevant is for us to be in *doubt* or neutral as to whether our counterfactual selves are our peers, as opposed to having good reasons for thinking that they are. As we shall see in section 5.6.2, even with the benefit of the doubt, there are good reasons not to take all our counterfactual selves as our epistemic peers when it comes to morality.

With these clarifications in place, it is clear that the direct argument for the disagreement view depends on whether the evolutionary hypothesis implies either narrow or broad peer disagreement (or both), as outlined in this section. I will now turn to defending the individual premises of my rebuttal of the direct argument for the disagreement view.

## 5.5 1st Horn of the Dilemma: No Disagreement in Nearby Scenarios

### 5.5.1 Restrictions in Cases of Counterfactual Disagreement

This section aims to make P1 more precise. P1 states two necessary restrictions on the epistemic significance of hypothetical disagreement in general.[21] First, hypothetical disagreement is relevant for domain D only if the disagreement concerns a proposition that is relevant for D. Talk of relevancy and 'domains' is not as precise as one would wish, and there are fuzzy boundaries. But, naturally, if a

supposition that you disagree about some question, for the proposition that they are right as for the proposition that you are right.

[20] See Gelfert (2011) for problems of these accounts.

[21] Naturally, proponents of the disagreement view should consider only disagreements that are plausibly implied by the evolutionary hypothesis.

proposition is irrelevant for D, then disagreement about it should not matter for our D-beliefs. On the other hand, the condition allows that disagreements concerning propositions that are not *within* the same domain are epistemically relevant for beliefs held in both domains. For example, two disputants that disagree about logic might have good reason to reconsider the credentials of their mathematical beliefs too, insofar as logic beliefs are relevant for mathematical beliefs.

Second, the hypothetical disagreement must be between epistemic peers. The narrow sense of peerhood is disjunctively connected with the broad conception, such that two thinkers are peers if they are equals regarding evidential possession or equally likely to get it right (or both). The need to limit the epistemic relevance of hypothetical disagreement in this way is suggested by the potentially devastating effects of combining an uncurbed epistemic relevance of hypothetical disagreement with a concessive view, as suggested in the following example.

Suppose that experts $E_1$ and $E_2$ are, before their encounter, defeasibly justified to believe p and ~p, respectively. According to a simplistic version of the concessive view, $E_1$ and $E_2$ lose their justification for maintaining either belief once they learn of their disagreement. To maintain their belief, they have to appeal to independent evidence for or against p, or find independent evidence that suggests that their interlocutor is not their epistemic peer, to settle the question whether or not p is true. Brushing aside thorny issues about the relevant sense of independence here, suppose that $E_1$ and $E_2$ do find independent evidence, q, about whether or not p. Normally, that would settle the disagreement. But with the suggestion about the relevance of hypothetical disagreement on the table, $E_1$ and $E_2$ cannot yet stop thinking about whether or not p, because it might be that expert $E_3$'s belief that ~q could either be *easily present* or *arbitrarily absent*. In the absence of a reason to think that $E_3$'s disagreement is too modally distant, $E_1$, $E_2$, and $E_3$ would, being diligent adherents of the concessive view, have to consider whether there is independent evidence about whether or not q or about $E_3$'s epistemic status (while $E_1$ and $E_2$ remain agnostic about whether or not p), ad infinitum. So, on the face of it, a concessive view about disagreement paired with a view about the

epistemic significance of hypothetical disagreement leads to a vicious regress that leaves us unjustified in holding any belief at all.[22]

So, lest general scepticism be embraced, the epistemic relevance of hypothetical disagreement must somehow be curbed. Peerhood amongst the interlocutors, in the sense introduced above, is a natural suggestion as a criterion for the epistemic relevance of a given disagreement. More pertinently, in the case of *hypothetical* disagreement, there are countless hypothetical interlocutors, $E_N$, which might be relevant to the dispute existing between any two disputants $E_1$ and $E_2$. Limiting the set of relevant (hypothetical) interlocutors to those who are in equal evidential possession *or* antecedently equally likely to get things right could help to curb the potential regress that is made possible by invoking hypothetical disagreements. Hence, hypothetical disagreement that is epistemically significant must be amongst peers (further defence of this claim follows in section 5.6.2).

The next two sections make the case that there is no scenario in which both criteria, *peer disagreement* about *relevant content*, are jointly fulfilled.

### 5.5.2 Relevant Moral Beliefs

My aim in this section is to narrow down the range of relevant beliefs that objectivists have to defend. According to the proponents of the disagreement view, the evolutionary hypothesis must imply hypothetical disagreement about moral beliefs. However, objectivists need not defend *all* members of $M_{Actual}$ against the evolutionary challenge, and hardly any objectivist aims to do so (Shafer-Landau 2003: 17). This is because $M_{Actual}$ certainly does not contain only true and justified moral beliefs. It contains moral beliefs that reflect biases, conceptual errors, and other infelicities. It also contains highly specific beliefs that refer to idiosyncratic sociocultural or even personal factors, such as the proper conduct on a wedding night. Objectivists do not claim that all of these beliefs are justified.

[22] In fact, considering the possibility of hypothetical disagreement in this sense is rather reminiscent of the Cartesian method of doubt, and I assume both that it would lead to radical scepticism and that proponents of the disagreement view aim for a *targeted* argument against moral objectivism; see Mogensen (2016a).

Rather, defending the justification of some moral beliefs is enough to vindicate objectivism. In particular, objectivists defend the justification (and truth) of the following moral beliefs:[23]

> Survival and reproductive success … is at least somewhat good. (Enoch 2010: 430)

> Pleasure is usually good, and pain is usually bad. (Skarsaune 2011: 232)

> We have rights because we are reflective beings. (Wielenberg 2010: 447)

These platitudes are of a similar structure: some plausibly evolutionarily relevant natural property or event (e.g. being an instance of survival, being painful, being capable of self-reflection, etc.) is related to a moral property such as being good. The normative concept alluded to is always a thin moral concept: GOODNESS, BADNESS, or RIGHT.

Let the set of moral platitudes be $M_{Basic}$. $M_{Basic}$ is a proper subset of $M_{Actual}$. I do not attempt to outline the contents of $M_{Actual}$. It suffices, however, to distinguish $M_{Basic}$ from $M_{Actual}$. The members of $M_{Basic}$ are the moral platitudes that make up a set of moral universals (cf. Curry 2016).[24] Moral platitudes have two components. First, moral platitudes contain *thin* moral concepts. Thin moral concepts are supposed to be purely evaluative moral concepts without descriptive content: RIGHT, GOOD, OUGHT are standard examples. Second, moral platitudes latch onto the non-moral facts that are evolutionarily relevant. For example, Curry (2016) found that every (studied) society has moral rules about problem-centred domains of resource allocation, coordination to mutual advantage, exchange, and conflict resolution. I do not assume whether or not the members of $M_{Basic}$ stand in deductive or inferential relations to each other. The members of $M_{Basic}$ are thus the beliefs that combine thin moral concepts with evolutionarily relevant causal factors, such as pain, procreation, and survival. Judging by the ethnographic record, every society accepts $M_{Basic}$.

This characterisation of the relevant domain in terms of $M_{Basic}$ suggests that in *non-actual nearby scenarios*, we have good reason to suppose that our

[23] Cf. Mackie (1977: 37), who discussed a related idea.
[24] Cf. Setiya (2012: 111), who calls these 'primary ethical facts'.

counterfactual selves will be like individuals in our society or other societies on the ethnographic record. Thus, given the ubiquity of beliefs in the platitudes of $M_{Actual}$, it seems very probable that $M_{Basic}$ is a proper subset of $M_{Counterfactual}$ too. Thus, turning back the wheel of life only a tiny bit will show that there is agreement rather than disagreement about $M_{Basic}$.

This is an important intermediary conclusion: if we confine ourselves to nearby possible scenarios, then the evolutionary hypothesis implies agreement with regard to some moral beliefs that can be explained evolutionarily, rather than disagreement. While the evolutionary hypothesis might suggest disagreement about *some* moral beliefs in nearby possible scenarios, these disagreements are merely disagreements about the application of thick moral concepts, rather than disagreements about the members of $M_{Basic}$ (Barkhausen 2016). As such, these kinds of disagreement need not worry moral objectivists. There is no disagreement about the objectivist moral beliefs, the domain that objectivists care about, implied by the evolutionary hypothesis. Thus, P2 is vindicated.

However, proponents of the direct approach will probably be unimpressed by the lack of disagreement in nearby scenarios. They might argue that considering only nearby scenarios betrays a lack of imagination. Recall Darwin's thought experiment about the bees, which is supposed to illustrate that 'we' could have ended up being very different organisms after all. In that case, it seems that the intersection between $M_{Actual}$ and $M_{Counterfactual}$ will get smaller and smaller as we replay the tape of life until we arrive at a version of our counterfactual selves that does not agree about any member of $M_{Actual}$ and thus, by extension, any member of $M_{Basic}$.

Therefore, P3 might still fail, since having the morality of bees would put $M_{Counterfactual}$ in conflict with $M_{Basic}$. So, the proponents of the direct approach might claim, the evolutionary hypothesis will reveal epistemically significant disagreement in more remote possible scenarios. Let's follow them there.

## 5.6 2nd Horn of the Dilemma: No Disagreement with Moral Peers

### 5.6.1 Disagreement Between Peers on a Narrow Conception

In *non-nearby* scenarios, in which $M_{Counterfactual}$ is not a proper subset of $M_{Basic}$, the evolutionary hypothesis would very be likely to yield some disagreement. However, my aim is to show that any disagreement we may find in non-nearby scenarios is not *peer* disagreement.

Let's consider *narrow* peer disagreement first. Recall that a narrow conception of peerhood says that two persons are peers if and only if they are in equal evidential possession and their processing of the evidence in regard to moral issues is also equal. I begin the assessment by clarifying how to best understand the term 'evidence' in the context of the evolutionary hypothesis.

In order to evaluate the disagreement view, we can take 'evidence' for one's moral beliefs to consist of mental states, such as beliefs, with non-moral content or with moral content, *or both*. That means that moral intuitions may count as evidence for one's moral beliefs.[25] There are two reasons for adopting such an inclusive notion of evidence in this chapter.

First, proponents of the disagreement-debunking thesis have to rely on a strong interpretation of the evolutionary hypothesis, which postulates a tight connection between ancestral environment, evolutionary forces, and the contents of moral intuitions and moral beliefs. Although this is a stark oversimplification of the evolutionary hypothesis, proponents of the disagreement view need it to fend off the objectivist claim that there is a subset of $M_{Basic}$ that is not subject to evolutionary contingency so the evolutionary hypothesis does *not* imply that it is

[25] Speaking of 'evidence' in the context of moral beliefs might raise eyebrows, for example if one a) takes evidence to be closely related to truth and factivity and assumes that there are no moral facts or b) takes evidence to consist of one's sensory input and assumes that moral facts are causally inert; see Quine (1969). However, since we are assuming moral objectivism for the sake of argument, both worries about the use of the term 'evidence' can be assuaged. There are objective moral facts (we assume) and so a factive notion of evidence is not obviously inappropriate when speaking about morality. Moreover, we should not require evidence to be restricted to sensory input – this would beg the question against moral objectivists, who do have a story to tell about what moral evidence is, on their view, e.g. in terms of a proper understanding or moral intuition; see Audi (1997), Huemer (2005), and Roeser (2011).

not also a subset of $M_{Counterfactual..}$.[26] On this view, the evolutionary hypothesis implies that our counterfactual selves' mental states with moral content as well as those with non-moral content will be different on different evolutionary paths. On this view, evolutionary processes influence the raw material based on which we form our moral beliefs, to such a degree that if you change the evolutionary background, you change the raw material and thereby the beliefs that our counterfactual selves are likely to hold (cf. Mogensen 2016a: 593).[27] The implications of the evolutionary hypothesis allow us to adopt an inclusive notion of evidence. To see this, consider that there are at least two factors that influence what moral beliefs it is *rational* for one to hold. The first factor consists of one's non-moral beliefs (and other non-moral mental states). The second factor consists of one's moral intuitions (Wedgwood 2010: 226). Of course, it is a deep and open question whether *both* factors are to be counted as one's evidence for one's moral beliefs (or only one of them). However, the evolutionary hypothesis implies that alternative evolutionary trajectories bring with them *both* different non-moral mental states and differing moral intuitions; *either* factor will be different in counterfactual evolutionary scenarios.[28] That means that irrespective of whether we take evidence for moral beliefs to include mental states with non-moral content (as opposed to only mental states with moral content), we will find different evidence on different evolutionary paths.

The second reason for adopting an 'inclusive' interpretation of evidence for moral beliefs is dialectical. Adopting a restrictive interpretation would be uncharitable for proponents of the disagreement view. On an inclusive interpretation, it is not hard for something to count as evidence. On a restrictive

[26] Denying the stringent relation between sensory input, moral intuitions, and moral beliefs is plausible but not an option for proponents of the disagreement view, as it allows objectivists to counter that if moral beliefs are not determined by moral intuitions which, in turn, are determined by evolutionary forces, then moral beliefs might reliably track moral facts after all (FitzPatrick 2015). Moreover, some have argued that moral beliefs need not be based on moral intuitions as I understand them here, but instead are based on 'direct perception' (Bengson 2015). Adherents of these views would reject my construal of the implications of the evolutionary hypothesis, but they would also not see a problem for moral objectivism in the first place.

[27] Environment is to be widely understood to encompass sociocultural factors.

[28] This is in line with Mogensen (2016a, 2017) and Bogardus (2016).

interpretation, it is hard for something to count as evidence. As an example of a restrictive notion of evidence, consider Williamson's view that your evidence is what you know (Williamson 2000). Williamson's notion of evidence would be restrictive in the present context because then the evolutionary hypothesis could not, per se, imply peer disagreement about objective moral facts, narrowly construed.[29] Either US or THEM would have evidence, but not both, and thus there would not be peer disagreement between THEM and US.[30] Thus, the more inclusive the notion of evidence, the more likely it is that the evolutionary hypothesis implies that there is disagreement with peers that share the same evidence (thus fulfilling the criteria for an epistemically relevant disagreement).[31]

Therefore, on a view of evidence charitable to proponents of the disagreement view, we may take moral intuitions as evidence for moral beliefs and furthermore assume that a) the outputs of our moral faculty, i.e. moral intuitions, are significantly influenced by evolutionary history and b) moral intuitions shape moral beliefs.[32]

Let's consider our counterfactual selves and US and compare the input–output relations of our moral faculty with the input–output relation of our counterfactual selves.[33] The options are exhausted by four cases, where *Input* refers to the forces that shaped moral intuitions in THEM and US ($Input_{US}$ and $Input_{Them,}$ respectively)

[29] Assuming that knowledge requires truth.

[30] Of course, this would be one way to go argue against the disagreement view, though one that I don't pursue here mainly because an adequate discussion of a theory of evidence is beyond the scope of this chapter (and the thesis).

[31] An example of an inclusive notion of evidence is the "dialectical conception of evidence" discussed by King (2012). According to the dialectical conception, "evidence is the sort of thing that is discursive and shareable through articulation" (King 2012: 254; compare van Inwagen 2010). However, problems with *communicating* intuitions make even the dialectic notion of evidence too restrictive for the proponent of the disagreement view. Our counterfactual selves are less likely, the further we go on the tree of life, to possess evidence in a discursive sense.

[32] This understanding of 'evidence' is unorthodox insofar as it does not signify an epistemic support relation: not every determinant of a moral belief is also an epistemically *good* reason for that belief (for some subject); see Huemer (2001: 376).

[33] I do not place too much weight on the term 'information' here. Information theory is a sophisticated and complex field and I wish mainly to exploit the thought of an input–output process whose relation between input and output is systematic, since this is what proponents of the evolutionary hypothesis claim; see Harms (2004) and Dretske (1981) for relevant introductions.

and *Output* refers to the set of moral beliefs (again with the subscript indicating whether they are OUR beliefs or THEIR beliefs):

(1) $Input_{Us} = Input_{Them}$ & $Output_{Us} = Output_{Them}$

(2) $Input_{Us} \neq Input_{Them}$ & $Output_{Us} = Output_{Them}$

(3) $Input_{Us} = Input_{Them}$ & $Output_{Us} \neq Output_{Them}$

(4) $Input_{Us} \neq Input_{Them}$ & $Output_{Us} \neq Output_{Them}$

Cases (1) and (2) signify agreement (since both outputs are identical) and are thus not relevant here. Cases (3) and (4) signify a divergence of $M_{Actual}$ and $M_{Counterfactual}$ and could be relevant for the purposes of assessing the disagreement view. However, case (3) is not implied by the evolutionary hypothesis and case (4) is not relevant disagreement.

Consider case (3) first. Case (3) *is* indeed a relevant peer disagreement. Our counterfactual selves might disagree about some beliefs in $M_{Basic}$, and given that these beliefs are based on the same input, the narrow conception of peerhood tells us that we have a peer disagreement. However, case (3) is *not* implied by the evolutionary hypothesis. Case (3) merely signifies that subjects that base their moral beliefs on the very same input will generate differing beliefs. In other words, the output is not correlated with the input – a sign of a random process. However, the evolutionary hypothesis does not imply that our moral beliefs are the products of a random process. Indeed, a crucial assumption of the evolutionary hypothesis is that moral beliefs are based on moral intuitions to such an extent that changing the moral intuitions would change the organism's moral beliefs.

To illustrate case (3), suppose that our counterfactual selves live in a world exactly like ours in all non-moral aspects. Given that they form their moral beliefs in the same way as we do, by relying on their intuition, there is no indication that their intuitions are any different in a world that is just the same as our world. The point is that disagreement is only a problem insofar as it is not the case that differences in output can be traced to differences in input (Wright 1992: 91ff). Therefore, case (3) does not follow from the evolutionary hypothesis and thus it does not help the proponents of the direct approach. Of course, the assumption that moral intuitions shaped by evolutionary forces determine the content of moral beliefs is a stark idealisation. If the bases of moral beliefs are fully determined by

evolutionary forces, pace the evolutionary hypothesis, then organisms subject to the same evolutionary history might have different moral intuitions and correspondingly different moral beliefs. Note, however, that this line of reasoning is no help for proponents of the disagreement view. Pursuing the same thought about the disconnect between evolutionary influences and moral beliefs, objectivists can argue that truth-conducive methods such as reasoning or understanding can lead to true beliefs based on intuitions that are not influenced by evolutionary forces (Copp 2008; FitzPatrick 2015; Huemer 2016).

Case (4) also shows a disagreement, and the evolutionary hypothesis plausibly implies it. Suppose, for example, that our counterfactual selves live in a world where their overall fitness is increased by sacrificing their children. They might indeed be rather like Darwin's bees. Due to various evolutionary processes, they might have different intuitions about how to treat their children than we do, and consequently they will form moral beliefs that seem to be in conflict with some of the members of $M_{Basic}$. Thus, we certainly have a relevant disagreement in case (4).

However, case (4) does not exhibit peer disagreement, narrowly construed, because evidence is not shared amongst THEM and US. Our counterfactual selves just have different moral intuitions: when they consider whether they should sacrifice their children, they might feel a warm glow of anticipation and a resounding positive attitude towards the thought – quite unlike our moral intuitions about infanticide. But it is not just that our counterfactual selves in case (4) have different intuitions; their intuitions are also formed in response to evolutionary pressures (at least that is the assumption). To explain these differences in intuitions using the evolutionary hypothesis, we have to assume that our counterfactual selves faced different natural and social environments. In other words, they will have based their moral beliefs on different intuitions, although the evidence is still of the same type. Therefore, the counterfactual selves that we disagree with in case (4) are not our peers according to a narrow conception of peerhood (they might be our peers with regard to factual knowledge, but if we understand moral evidence as encompassing moral intuition, and adopt a narrow

conception of peerhood, then THEM being peers about factual knowledge is immaterial to the present issue).[34]

Objection: we should adopt a slightly wider conception of peerhood, according to which there needs to be sameness of the *type* of evidence (i.e. intuition) rather than the same *token* of evidence (i.e. the *same* intuition). In that case, our infanticide-approving counterfactual selves count as our peers, since their beliefs are based on the evidence type 'intuitions', despite having formed different intuitions in response to wildly different environments.

This objection is implausible, however. First, the objection relies on an unorthodox variant of the concessive view about disagreement, which usually takes *tokens* of evidence as relevant and not types of evidence. As such, that need not be a problem yet. It does, however, raise gnarly issues about method individuation familiar from the wider epistemological debate (Conee and Feldman 1998). Second, that variant has abstruse consequences that we should not accept. For example, a hallucinating person would count as our peer, just because the perceptual belief is formed based on the same type of process. So would a deeply depressed economist, who uses his or her cognitive capacities to evaluate the economic data, albeit with an unconscious bias for negative signals in the data. The most natural response is to say that hallucinators or deeply depressed colleagues are not our epistemic peers (relative to the topic in question). This problem might be seen as an artefact of the thin formulation of the narrow view that I proposed earlier. It could be remedied, however, only at the cost of introducing additional criteria for assessing peerhood, which would further complicate the task for the proponent of the disagreement view.

Therefore, on a narrow conception of peerhood that is congenial to the proponents of the disagreement view, the evolutionary hypothesis does not imply relevant peer disagreement in non-nearby scenarios.

Granted, however, the narrow conception of peerhood is not, though congenial to Bogardus's view, the most felicitous conception of peerhood for proponents of the disagreement view. Their argument could yet be saved if there were disagreement

[34] Implications of this view might be taken to be good reasons to reject the narrow conception of peerhood, and I consider the prospects of the broad conception of peerhood next.

on the broad conception of peerhood. In the next section, we stay in non-nearby scenarios but consider whether any of our counterfactual selves are peers on a broad conception of peerhood.

### 5.6.2 Total Disagreement Between Peers on a Broad Conception

Recall that, according to the broad conception of peerhood, our counterfactual selves must be "equally likely to be right" about moral matters (Vavova 2014b: 308). Moreover, giving our counterfactual selves the benefit of the doubt (in regard to whether we bestow default trust on them and so believe they can be our peers in relation to moral matters), the question is whether the evolutionary hypothesis implies that there is disagreement with our counterfactual selves whom we have *no* reason *not* to regard as our peers. Since relevant disagreement is about $M_{Basic,}$ we can imagine two cases: total disagreement and partial disagreement about $M_{Basic}$. Neither case, however, creates a problem for moral objectivism.

Consider total disagreement about $M_{Basic}$ first. The disagreement is total in the sense that our counterfactual selves disagree about *all* beliefs in $M_{Basic}$. Total disagreement about $M_{Basic}$ is a tremendously extreme situation.[35] The extremity of rejecting $M_{Basic}$ altogether implies that familiar cases about 'moral monsters' in the debate about peer disagreement do not straightforwardly apply (Sinnott-Armstrong 2014; Vavova 2014b). In such cases, if they are to be realistic at all, we could at least presuppose agreement about *some* members of $M_{Basic,}$ such that survival is good or that pain is pro tanto bad.

None of this common ground can be found in the case of total disagreement about $M_{Basic}$. Noting the extremity of total disagreement about $M_{Basic}$ is relevant because it suggests that worries about denying peer status all too easily are not warranted in the evolutionary case.[36] For example, Tersman points out that "the

[35] I doubt that we have a clear sense of what an organism would be like that disagrees even about things like 'survival is pro tanto good', without even considering whether such organisms could plausibly evolve.

[36] Compare the discussions of radical moral disagreement with morally deficient individuals by Sinnott-Armstrong (2014: 53) and Ballantyne (2013: 254). Both suggest that moral disagreements with psychopaths would be epistemically significant for the justification of our moral views. The disagreement I am considering is more extreme: psychopaths can recognise, for example, a shared method at arriving at moral beliefs and what constitutes good moral reasoning, they are simply unperturbed by it. Total

mere fact that a person disagrees with us, or is incorrect about the disputed issue, cannot itself count as a shortcoming" (Tersman 2006: 34ff). That is correct, but insofar as our counterfactual selves disagree about *all* our moral beliefs, the charge of undue marginalisation of their moral opinion does not readily apply. My aim in what follows will be to show that total disagreement about $M_{Basic}$ provides good reason not to take our counterfactual selves as our peers.[37]

We can demote our counterfactual selves from their defeasible status as peers in cases of total disagreement about $M_{Basic}$ because morality as a domain of belief is independent of non-moral domains (as I discussed in chapter 4, section 4.5) and because we can assume that we are moral but not that our counterfactual selves are moral (insofar as there is total disagreement). This gives us reason to demote our counterfactual selves from moral peerhood. Proponents of the disagreement view should agree with this view. Let me illustrate why.

Recall that we are assuming, with the proponents of the disagreement view, that the default position is to take others as our peers when it comes to morality (in the broad sense of peerhood) (cf. Mogensen 2017). Prima facie, the fact that our counterfactual selves disagree about the totality of $M_{Basic}$ is not itself a reason to deny their status as our peers in moral matters. But, of course, Mogensen and others grant that default peerhood status can be lost (Mogensen 2017: 294ff).

The specific case that Mogensen discusses to illustrate the loss of default peerhood status in a case of localised disagreement is instructive: suppose that you judge that torture is morally impermissible, but you learn that within a week you will judge that torture is morally permissible. In light of this counterfactual, localised disagreement (with your future self), is your current belief about the impermissibility of torture being defeated (Wedgwood 2010: 241)? No, says Mogensen, because we can justifiably reject the other's peer status based on the following line of reasoning:

---

disagreement with our counterfactual peers cannot even presuppose that much agreement about methods of moral reasoning.

[37] See Goldberg (2013), who calls this "non-localized" disagreement. This kind of disagreement is so radical that defenders of conciliatory views also think that it reduces our claim to peerhood. See Kornblith (2010: 50), who claims that a "homicidal sociopath", does *not* count as a peer in moral matters; see also Elga (2007) for related claims.

> I'm told that my future intuition will favour torture, *and nothing more*. This leaves open a host of questions about what will happen to me and how I will come to hold this view, which I now regard as repugnant. I don't have to rely on the correctness of my present view in order to remain reasonably steadfast in that case. I can rely on the knowledge that, since one of us must be badly mistaken, something must have gone wrong somewhere. I can combine this knowledge with my more detailed knowledge of my present self. This should lead me to assign significant confidence to the hypothesis that I'll be subject to some relevant epistemic fault in [the] future. (Mogensen 2017: 294–5)

This passage implies that the default peerhood status of others (in domains where intuitions count as evidence) can be defeated if considerations *independent* of the disputed proposition let us assign a higher likelihood to us being right about the disputed proposition compared with the likelihood that the interlocutor is right. In other words, Mogensen concedes that if we know *nothing* about an interlocutor who enjoys default peer status, except for the disagreement, the interlocutor loses their status as a peer. It follows that if we know nothing about an interlocutor who enjoys default peer status, except for the fact that there is *total* disagreement, the interlocutor loses their status as a peer.

Of course, knowing *nothing* in the sense of having literally *no* information at all is not really relevant in the case that Mogensen discusses. Rather, it is knowing *nothing of relevance* with regard to the disputed issue that matters. Mogensen agrees that knowing nothing of relevance allows us to demote disputing parties from their status as peers. My point is simply that knowing that they disagree about the totality of $M_{Basic}$ equals knowing *nothing of relevance* when we are looking for reasons to uphold their status as peers.

The main point in support of this claim is that in domains where the only evidence is intuitions (as, in accordance with the assumption, in the objectively construed moral domain), then knowing *more* about the interlocutor who disagrees about the totality of claims in that domain does *not* salvage the interlocutor's status as a peer. The crucial point is that we have to presuppose some standard by which we can ascertain what it means to get moral matters right (Elga 2007: 493ff). If we bracket the contents of $M_{Basic,}$ we have no such standard, and we could not ascertain what it means to get moral matters right. Thus, even granting the claim that we start out with a default recognition of others as peers in domains where

intuitions count as evidence, we can retract that trust upon learning that there is no agreement in the *domain of dispute* at all. Consider three reasons in support of this point.

First, agreement in a domain D where intuitions count as evidence would be a reason for an interlocutor to maintain their status as an epistemic peer about D, all else being equal. In other words, agreement about $M_{Basic}$ constitutes common ground based on which we can assess the likelihood of getting moral matters right. For example, consider whether our counterfactual selves would be our peers if they were like the Neanderthals.[38] Suppose we know nothing about their phylogenetic relatedness, their social habits, or their ventures into early forms of art. Despite ignorance on these matters, a good reason (not necessarily a sufficient reason) to maintain 'default trust' in the moral intuitions of Neanderthals is that they agree about $M_{Basic}$ (Mogensen 2017: 283). If all we know about Neanderthals is that they agreed about $M_{Basic,}$ we would be given no reason to withdraw the default trust we bestowed upon them. Evolutionary considerations also suggest that Neanderthals agreed at least about parts of $M_{Basic.}$ Neanderthals plausibly believed that it is good to take care of one's offspring, so they would have cherished survival and generally avoided pain. Agreement about $M_{Basic}$ thus might give us good reason (in the sense that it does not violate the default trust) to take them as our moral peers, despite the 30,000 years that separate us from them.

Second, consider whether we would *still* have reason to take our counterfactual selves such as Neanderthals as our peers *in the absence of* any agreement about $M_{Basic.}$ To aid imagination, let us conjure up some evolutionary path on which there is *total* disagreement about $M_{Basic}$.[39] Suppose our counterfactual selves do not endorse (or form relevant proto-judgements that lead them to accept) any of the members of $M_{Basic}$ and that they are like the imaginary *Homo sapiens peregrinus*, the strange man. It is difficult to say in what sense your peregrinus-self would still be a counterfactual version of us, given that it, ex hypothesi does not endorse any of our basic moral beliefs. Not only would your

[38] Neanderthals are often considered a subspecies of the genus *Homo*, and they went extinct approximately 30,000 years ago; see Tudge (2006).

[39] Total disagreement is so radical that even defenders of concessive views concede that there are cases in which it loses its epistemic significance; see Kornblith (2010: 50) and Elga (2007).

peregrinus-self be extremely odd from the moral perspective, but it would also be doubtful that your peregrinus-self would be a plausible product of an evolutionary process. But set the biological worry aside – how could we maintain that our peregrinus-self is as likely as us to get things right when it comes to morality *in the absence of* any agreement about $M_{Basic}$? In the absence of agreement about $M_{Basic}$ it is hard to see how we could. Peregrinus had the benefit of the doubt, but lack of even singular agreement about $M_{Basic}$ takes away his prima facie peerhood status.

Third, cases of total disagreement about a domain D where intuitions count as evidence are special because gaining information unrelated to D about the interlocutor does not provide relevant information for evaluating the interlocutor's status as a peer.

To underscore this point, consider *further* criteria, apart from agreement about $M_{Basic}$, which might show that peregrinus-like counterfactual selves are our peers in regard to morality despite their total disagreement about $M_{Basic}$. If any seem plausible, we could dismiss the relevance of even partial agreement about $M_{Basic}$ for evaluations of moral peerhood. I cannot offer an exhaustive case, but three considerations on behalf of the proponent of the disagreement view suggest that the prospects are dim.

To reinstantiate the peerhood of peregrinus-like counterfactual selves, proponents of the direct approach might refer to peregrinus's (1) *cognitive capacities including non-moral beliefs*, (2) *belief-forming method,* and (3) *physical aspects and phyletic relatedness* to support the claim that they are our peers.

- *Cognitive capacities including non-moral beliefs.*[40] Our peers are those who, in general, reason as well as us. They are as good as we are in obtaining factual and scientific knowledge. They compose logical proofs, understand physics, and perform as well on standardised intelligence tests as average humans. The non-moral cognitive capacities of peregrinus, which are similar to ours, make it likely that peregrinus will adopt similar moral beliefs to us.

[40] Capacities are understood here as having the *ability* to function on a certain level.

However, cognitive development might be a necessary condition for counting someone as our peer in moral matters, but it is certainly not a sufficient condition. In other words, it is true that cognitive development of a certain level might function as a kind of 'enabler' for making correct moral judgements (Klenk 2017b). For example, if peregrinus lacked a theory of mind, similar to very young children, he would be prone to making egoistic decisions and would simply lack the ability to recognise that other beings have their own plans and own wishes (Kohlberg and Hersh 1977). It might be thought that cognitive abilities alone do provide a direct reason for expecting peregrinus to be a good *moral* reasoner. But the opposite is true. Just because a specimen of peregrinus can realise that you would be hurt by something he does, this does not imply that he will respect that consideration. Moreover, people can be experts in one area but still be (systematically) wrong in another area, and it is generally the case that assessments of peerhood seem domain specific (Goldberg 2013: 169; Weatherson 2013: 56). Otherwise, it would make good sense to ask expert chess players to sit on ethics committees and top-notch nuclear physicists to weigh-in on Europe's border policies simply *because* their cognitive abilities are taken as evidence of their ethical expertise. These proposals do not look promising, and thus cognitive capacities should not count as a reason in favour of moral peerhood, either. Consider the following alternative instead:

- *Similarity of belief-forming methods.* Our peers are those who use similar methods of belief formation. Peregrinus is as good as we are in obtaining knowledge about non-normative matters. Peregrinus also relies on his intuition in forming moral beliefs, and so do we. So, peregrinus is our peer when it comes to moral beliefs (according to a broad conception of peer-hood).

Notwithstanding the claim that all moral beliefs are based solely on intuitions, focusing on the distinction between the content and the status of evidence does not help defenders of the disagreement view at all. The consequence of taking peregrinus as our peer *based on his belief-forming method* is an inflation of peers. Anything that forms moral judgements based on intuitions follows the same method and thus, irrespective of the contents of resulting beliefs, we should count it as our peer, which seems absurd. Moreover, it would be unclear why we

should restrict peerhood to those things that form *beliefs* based on intuitions. Proto-judgements such as normative inclinations are also based on intuitions, insofar as intuitions are unreflective. Since virtually anything that *lives* makes proto-judgements, all living things (or rather things that lived or can be conceived to live) count as our peers when it comes to moral matters. The mere status of the evidence for our moral beliefs should thus not be the criterion for assessing peerhood. Consider finally the following proposal:

- *Physical aspects and phyletic relatedness.* Our peers are those that have a similar phenotype and those that are close phyletic relatives of us. Bipedal mammals living in small tribes are likely to get things right in morality because they show ample similarity to us in many physical aspects that seem relevant to morality.

Unfortunately for proponents of the disagreement view, not even belonging to the same *species* would be enough to determine epistemic peerhood (Vavova 2014b: 330). If non-moral beliefs are considered to be good determinants of the likelihood that peregrinus will agree on moral matters, then we might reason as follows. Given, for example, the bodily constitution of peregrinus, he will probably want to avoid pain. He will therefore probably form beliefs about the badness of pain. In that case, however, we agree with peregrinus, and there is no disagreement.

Alternatively, we might reason as follows. Peregrinus has a bodily constitution similar to ours, so, in a world like ours, he would want to avoid pain. But in a world different from ours, where it somehow happened that public feats of pain tolerance lead to social appraisal, peregrinus would probably embrace pain and form the judgement that pain is good. In that case, however, peregrinus would have had different information to base his beliefs on, which is why we should not count him as a peer.

Therefore, some of the alternatives to *agreement* about $M_{Basic}$ suggest that there are *no* good reasons to take our counterfactual selves as our peers on a broad conception of peerhood if there is total disagreement about $M_{Basic}$. I have not considered all possible alternatives. But there are good reasons to think that none will be successful. The crucial point is that we have to presuppose some standard

by which we can ascertain what it means to get moral matters right (see Elga 2007: 493ff). The standard by which we can compute the likelihood that others get things right is their agreement about $M_{Basic}$.[41] Recall that, before any disagreement, we take ourselves to have good grounds to think that the members of $M_{Basic}$ are largely true. Thus, we have good grounds to believe that we are getting moral matters right, insofar as we believe in $M_{Basic}$. Our counterfactual selves do not. So, if the evolutionary hypothesis implies total disagreement about $M_{Basic}$, then objectivists need not be concerned.

These considerations suggest that *partial* disagreement about $M_{Basic}$ would not defeat the prima facie status of our counterfactual selves as peers. In the next section, I consider and rebut a final option on behalf of the disagreement view: partial disagreement about $M_{Basic}$.

### 5.6.3 Partial Disagreement Between Peers on a Broad Conception

Still staying in non-nearby scenarios, the much more plausible case is that we rewind the wheel of life, but only to a point where there is still *some* agreement about $M_{Basic.}$ Let partial disagreement be a case in which our counterfactual selves agree about at least one belief that is a member of $M_{Basic}$. We might, therefore, have reason (though perhaps not sufficient reason) to count them as our peers on a broad conception of peerhood. Of course, there are fuzzy boundaries, and I do not suspect that we can say with precision whether agreement about some percentage of the members of $M_{Basic}$ is required for peerhood. But there could be enough agreement to raise the suspicion that "there is no reason to suppose that either party to the dispute is in an evidentially superior position" (Cohen 2013: 98). So, debunking explanations could reveal *local* disagreement with peregrinus. It might concern only some members of $M_{Basic.}$ However, this line of argument does not vindicate the disagreement view for two reasons.

[41] This point resembles a point made by Davidson (1984) about radical interpretation. Davidson argues, roughly, that in cases of radical disagreement about a subject matter, the 'principle of charity' demands that we should regard the other party as talking about a different subject altogether. Since we are concerned with merely conceivable disagreement, I suppose we can conceive that there is *no* talking at cross purposes going on and so we need not be charitable. Still, in agreement with Davidson, I believe that we should not take seriously the disagreement in this case.

First, if we consider just one counterfactual scenario, in which we end up like peregrinus, say, then the most that proponents of the disagreement view could conclude is that the justification of *some* beliefs is challenged. Such a case would not show, however, that all objectivist moral beliefs in $M_{Basic}$ are defeated.

For example, it might be that we cannot determine whether it is morally permissible or impermissible to abort foetuses. But this finding does not imply that all the other moral beliefs in $M_{Basic}$, about which there is agreement, are also unreliable. To reach that conclusion, proponents of the direct approach would have to appeal to a principle of the following sort:

Token-Type: If there is peer disagreement amongst tokens of a type of belief, K, then the type of belief in question is epistemically suspect.

However, the Token-Type principle is certainly false. There may be radical disagreements about matters in physics, but we do not judge all beliefs about physics to be unjustified. Instead, it seems appropriate to judge that the question is beyond our (current) abilities to answer. Objectivists can adopt the same reasoning. There might be peer disagreement about some members of $M_{Basic}$ – and we might want to suspend judgement about those – but that need not compel us to suspend judgement about all beliefs in $M_{Basic}$.[42]

Second, proponents of the disagreement view might argue as follows: if we consider manifold disagreements with manifold counterfactual selves, we could get cumulative total disagreement about $M_{Basic}$. To illustrate, assume that $M_{Basic}$ contains two non-overlapping proper subsets: A and B. We agree with peregrinus about A and disagree about subset B. Now imagine that there is another of our counterfactual species, say *Homo sapiens cerritulus*, the mad man. We agree with cerritulus about B but disagree about A. As a result, there is peer disagreement about all beliefs in $M_{Basic}$, albeit not with the same interlocutor.

However, that response is only initially plausible because it is unlikely that the evolutionary hypothesis implies that such a situation is possible. For one, the contents of the beliefs in $M_{Basic}$ are ecologically related in worlds that are similar to ours. If debunking explanations imply relevant disagreement about $M_{Basic}$ with,

[42] This might imply that there are some moral propositions that are unknowable on an objectivist account of morality (Wright 1992).

say, cerritulus, then cerritulus's world would be very different from ours. Thus, it would be unlikely that cerritulus will agree about the beliefs contained in set B. In other words, disagreement about some beliefs in $M_{Basic}$ raises the probability of disagreement about other beliefs in $M_{Basic}$, such that it is unlikely that there could be a cumulative disagreement about *all* beliefs in $M_{Basic}$. Moreover, given that mere agreement about bits of $M_{Basic}$ can be considered a necessary but not a sufficient condition for peerhood, it is not clear, and certainly not established on the broad conception of peerhood, that imagining many deviant species with whom we have partial agreement establishes that there is relevant peer disagreement.

These considerations suggest that partial disagreement about $M_{Basic}$ is plausible to some extent, but that it does not yield the desired conclusion that all beliefs in $M_{Basic}$ are subject to justification-defeating disagreement. This concludes the case for premise P3. As I said, it is not a conclusive case. In particular, there might be other criteria based on which we could take our counterfactual selves to be our peers despite them disagreeing about all members of $M_{Basic}$. And it might be possible to conjure up scenarios that are biologically possible in which there is a triangulated total disagreement about $M_{Basic.}$ But proponents of the disagreement view have not made that case. As such, the considerations of this section suggest that, no matter how the tape of life is played, we will not find relevant, justification-defeating disagreement about objectivist moral beliefs, so moral objectivism has not been refuted by the disagreement view.

## 5.7 Concluding Remarks

In conclusion, the evolutionary hypothesis does not show that there is relevant hypothetical disagreement that defeats the justification of all our objectivist moral beliefs. As a result, the direct approach to supporting the disagreement view fails. If all alternative interpretations of the epistemic significance of the evolutionary hypothesis fail, as some proponents of the disagreement view claim, then evolutionary debunking arguments fail to have sceptical consequences for moral objectivists.

But even without assuming the radical claim that the evolutionary hypothesis is relevant only insofar as it implies the possibility of counterfactual moral disagreement, this chapter shows that appeals to disagreement do not help

the debunker's case. This takes away one possible route for debunkers to press their sceptical conclusion. For defenders of moral objectivity, this means relief on one front. For their opponents, this means that they need to reinforce efforts to find another epistemic phenomenon to undergird evolutionary debunking arguments against the objectivity of morality. The most promising route, I suggest, lies in finding a way to spell out the epistemic significance of the evolutionary hypothesis in terms of a theory of defeat. Disagreement, in any case, does not help the debunker in the case against moral objectivism.

# 6 Third-Factor Explanations and Disagreement

## Reader's Guide

How can we explain the reliability of our moral beliefs? The most prominent, and most promising, way is to show that a common cause explains both why we hold certain beliefs *and* why those beliefs are true. Providing such an explanation is known as giving a third-factor explanation of the reliability of our moral beliefs. Third-factor explanations, about which I will say more below, are fitting responses to the reliability view, which, you will remember, asks us to explain the reliability of our moral judgments in light of evolutionary explanations of morality. There is an interesting connection between the reliability view and the disagreement view: the epistemic significance of actual disagreement might hamper the prospects of a third-factor account and so disagreement may play an important role in the evolutionary defeat challenge, viewed from the perspective of the reliability view, after all.

In this chapter, I connect the disagreement view that I addressed in the previous chapter with the reliability view. The disagreement view and the reliability view are connected as follows: taking the reliability view, we need to find a way to explain the reliability of our moral judgements (in light of evolutionary explanations of morality). The most promising accounts that provide such an explanation, third-factor accounts, are possibly limited by the degree of actual disagreement. This is how: third-factor accounts depend on a substantive moral claim (which constitutes the common cause, or eponymous third-factor), and if there is actual, irresolvable disagreement relevant to that claim, the best way to resist defeat on the reliability view might ultimately falter because of the epistemic significance of disagreement. Thus, it might turn out that the disagreement view and the reliability view have interesting interconnections, and I will explore whether actual disagreement might hamper the arguably best 'objectivist escape

route' to a defeater on the reliability view. As figure 6.1 illustrates, the objectivist escape route could be blocked by showing that actual disagreement rules out valid third-factor accounts:

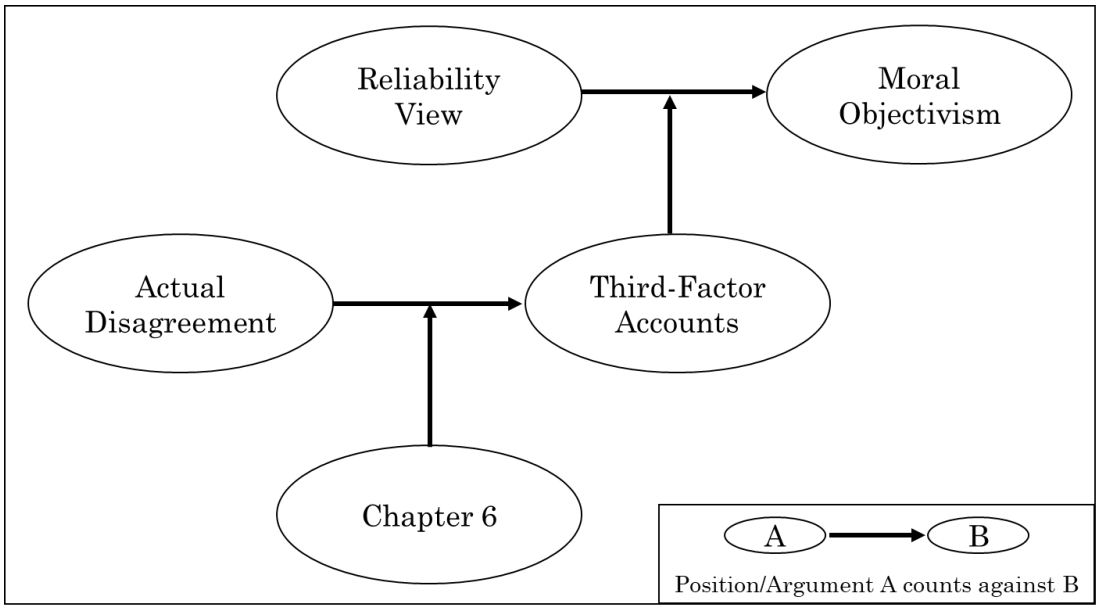


*Figure 6.1 The argument against the disagreement view (indirect route)*

I will argue that actual disagreement does not hamper third-factor accounts. Before turning to my argument, I will use this reader's guide to say more about third-factor accounts. Though I touched on third-factor explanations in chapter 4, I will provide more details about why third-factor explanations are controversial and then preview how they are relevant for the disagreement view and the reliability view.

Third-factor explanations get their name from the third factor that they invoke to explain the correlation between the content of moral beliefs and objectivist moral truths. Such a third factor will be a substantive evaluative claim. That claim, let's call it a bridge law, will attribute a moral property (again, here we have to remember that some objectivists will take this talk of 'property' to be metaphysically non-committal) to a natural property. Then there will be some causal explanation for why humans have beliefs that reflect this claim because the natural property referenced in the bridge law can play a role in an evolutionary explanation. There will also be a philosophical explanation for why beliefs that

reflect the bridge law are true: the bridge law grounds their truth.[1] Third-factor accounts are common-cause explanations, which are familiar forms of scientific explanation (Sober 1984a). An important feature of third-factor explanations is that they might be able to show why objectivist moral beliefs are reliable without invoking the claim that tracking the truth was an adaptation (Vavova 2015: 110). However, it is unclear and a question of considerable discussion whether third-factor accounts are legitimate means to resist defeat in the context of the evolutionary challenge (Behrends 2013; Fraser 2014: 471; Locke 2014; Moon 2017; Street 2016; Vavova 2015: 111).

The controversial feature of third-factor explanations is that they rely on a substantive moral claim, and this has attracted much criticism. I will briefly review the most important objection, and then say how it can be resisted. Many have objected that it is question-begging to assume a substantive moral claim in answering an evolutionary challenge (e.g. Street 2016). I am careful to write 'an' evolutionary challenge because many criticisms of third-factor explanations criticise them in response to *different* kinds of evolutionary challenges. Clearly, in response to some kinds of challenges, assuming the truth of a substantive moral claim might be question-begging, but not in response to other challenges. The crucial questions is whether the evolutionary defeat challenge succeeds so that our moral beliefs are all defeated; in this case, we could not reinstantiate the justification of our moral beliefs, as I argued in chapter 4. Thus, the crucial question is whether evolutionary explanations of morality imply that moral beliefs are *already* defeated or not.

There are thus two possibilities that we have to consider when assessing the legitimacy of third-factor accounts. One possibility is that all moral beliefs are already defeated by evolutionary explanations of morality. As I argued in chapter 4, if that is the case, then no moral belief is left that can be used to fashion a third-factor explanation. Relying on a substantive moral claim, in this case, would be question-begging. However, it would also be unclear *why* one would want to set up a third-factor account in response to a generic evolutionary defeater. Third-factor

[1] Examples of third-factor accounts are Brosnan (2011), Enoch (2010), Locke (2014), Schafer (2010), Locke (2014), Skarsaune (2011), Talbott (2015), and Wielenberg (2010).

explanations are specific responses to the evolutionary defeat challenged viewed through the lens of the reliability view. Third-factor explanations are attempts to explain the reliability of moral beliefs, and as such, they are naturally thought of as relevant only when the reliability of moral beliefs is in question.

The second possibility is that not all moral beliefs are defeated when objectivists try to set up a third-factor account. This seems more plausible to me. On one interpretation of the reliability view, evolution raises a *challenge* only and not a case yet that shows that all moral beliefs *are* defeated. It would make sense to explain the reliability of moral beliefs if evolutionary explanations of morality raise a *worry* about the reliability of moral beliefs, but in that case moral beliefs are not yet defeated. Thus, moral beliefs are assumed to be undefeated when one tries to set up a third-factor account. In that case, it is legitimate to use them in defending a third-factor explanation.

Some have objected to this view by arguing that it is circular or question-begging to rely on substantial beliefs of the type of beliefs whose reliability one wants to explain. Thus, if the challenge is to explain the reliability of moral beliefs, it would be illegitimate to rely on moral beliefs. Since third-factor accounts clearly do this, that might be another reason for rejecting third-factor explanations.

However, to explain the reliability of beliefs of any type, one has to invoke in one's explanation at least some substantive beliefs of that type. Since this is just what third-factor explanations do, they are legitimate, or so the argument goes. This argument in favour of the legitimacy of third-factor accounts relies heavily on the idea that it is impossible to explain the reliability of any faculty in a non-circular way (Berker 2014; Foley 2001; Vavova 2015). As pointed out in the main introduction, it is thus plausible to grant the prima facie plausibility of third-factor accounts as attempts to explain the reliability of moral beliefs.[2]

Therefore, there are good reasons to think that third-factor explanations in response to reliability challenges raised by evolutionary explanations of morality are prima facie legitimate. Assuming that an evolutionary defeater would be due

[2] There might be room for opponents of moral objectivism to argue that explanations of the reliability of perceptual beliefs are less controversial nonetheless. My point here is solely that it is premature to reject third-factor accounts on the ground that they are question-begging.

to a concern with reliability, a third-factor account could yet block evolutionary defeat. It is for this reason that I consider the relevance of disagreement for third-factor explanations in this chapter. If the epistemic significance of disagreement *constrains* or even invalidates third-factor explanations, we would have the following case: there is an evolutionary defeater for all moral beliefs (based on the reliability interpretation) because the only way to resist that defeater is invalidated due to the epistemic significance of disagreement. Disagreement would play a crucial role in evolutionary defeat after all. Let's now turn to Tersman's (2017) recent argument that this is the case.

## Abstract[1]

Several moral objectivists try to explain the reliability of moral beliefs by appealing to a third factor, a substantive moral claim, which explains, first, why we have the moral beliefs that we have and, second, why these beliefs are true. Folke Tersman has recently suggested that moral disagreement constrains the epistemic legitimacy of third-factor explanations in two ways. (1) The moral beliefs that objectivists seek to defend must provide support for the substantive moral claim used in the third-factor explanation. (2) The substantive moral claim must provide a theory of error for the moral beliefs that objectivists seek to defend. This chapter aims at showing that disagreement does not constrain the epistemic legitimacy of third-factor explanations in metaethics. First, Tersman's constraints are impossible to violate. Second, some disagreements are irrelevant, given that they cannot be about beliefs whose reliability the objectivist seeks to defend. Third, actual disagreement about moral beliefs is implausible, given recent ethnographic findings. The chapter thereby weakens the case for the view that epistemic issues that have to do with disagreement are relevant for evolutionary debunking arguments in metaethics.

## 6.1 Introduction

*The Darwinist view of morality* claims that the human propensity to make certain moral evaluations, such as our widespread preference to judge that we have most reason to support our children, can be explained in evolutionary terms. *Evolutionary debunkers of morality* have taken up the Darwinist's claim to draw metaethical conclusions. Most prominently, some debunkers argue that the Darwinist view of morality implies that moral objectivism, the view that there are mind-independent, non-natural moral truths, would commit us to moral scepticism: there would be no reason to think that any moral belief is justified (Joyce 2006, 2013a, 2016c, 2016c; Street 2006, 2016).[2]

A commonly cited construal of the debunker's challenge is what I call the *Reliability View*: the challenge posed by the Darwinist view of morality is to *explain*

[1] The paper on which this chapter is based is currently under review.

[2] The structure and conclusion of both Street's and Joyce's argument is subject to much debate, cf. Berker (2014); Bogardus (2016). For example, Street's (2006) main aim is to show that moral objectivism should be rejected. As such, her argument may be better interpreted as suggesting that we have no reason to think that any objective moral belief is justified, in light of a Darwinist view of morality. This claim, in turn, is plausibly seen as an intermediary step toward the conclusion that all our objectivist moral beliefs are undercut, in light of the Darwinist view about morality.

*the reliability* of objective moral beliefs. Though the assumption is not without problems, it is often assumed that 'explaining the reliability' of some set of target beliefs at least requires one to explain why those beliefs are true more often than chance would predict (cf. Schechter 2018; Street 2016: 305).[3] If an adequate explanation proves to be in principle impossible, the challenge goes, then any prima facie justification that our objective moral beliefs might have is undercut.[4] Several objectivists regard this challenge to be their most arduous test (Enoch 2010; Shafer-Landau 2012; Wielenberg 2014). In effect, the focal point of the current debunking debate, viewed from the perspective of the reliability view, is whether objectivists can adequately explain the reliability of the objectivist moral beliefs (Vavova 2015: 111). It is widely assumed that so-called *third-factor explanations* are the most promising candidate explanations available to moral objectivists (Behrends 2013; Enoch 2010). Third-factor accounts appeal to "bridge principles" that "posit a relation between the facts in virtue of which our moral beliefs are true and the (non-moral) facts to which the evolutionary account attributes them" (Tersman 2017: 765).[5] If third-factor explanations work, then moral objectivists can pass their most arduous test.

This chapter starts by addressing a recent innovative proposal by Folke Tersman about how to determine the epistemic legitimacy of third-factor explanations (Tersman 2017). Tersman argues that third-factor explanations are constrained by *the epistemic significance of disagreement*. More precisely, objectivists rely on a substantive moral claim, in the form of the 'bridge principle', to get the third-factor explanations off the ground, and radical moral disagreement

[3] The correct interpretation of 'explaining the reliability' is an interesting issue of its own that I cannot fully address in this chapter. Common alternative interpretations to the one introduced above invoke modal conditions such as sensitivity or safety, while Tersman (2017) proposes the view that a belief is reliable to the extent that possessing it gives us reason to think its content is true. The correct interpretation of 'explaining the reliability' is an issue that can be set aside in this chapter insofar it may affect whether third-factor accounts offer any help against debunking arguments in the first place, while I focus on whether disagreement constrains third-factor accounts. Since as Tersman regards third-factor accounts to be potentially viable responses to explaining reliability in his sense, too, the minimal characterisation invoked above should suffice for present purposes.

[4] See Pollock and Cruz (1999: 195ff) for a discussion of undercutting defeat.

[5] 'Belief' is ambiguous here: 'true belief' really means that the propositional content of a belief is true, whereas evolutionary explanations show why we have dispositions to hold true certain propositions.

might undermine the objectivist's prima facie justification for maintaining that claim. Tersman's proposal threatens objectivists with a second-order problem: there might be no legitimate moral belief to serve as the bridge principle for a third-factor explanation.

If Tersman's constraints cannot be met, then objectivists lose their most promising answer to the evolutionary debunking challenge. Interestingly, Tersman's view more generally lends support to a novel hypothesis about the general epistemic significance of debunking explanations, as championed by Bogardus (2016), Mogensen (2016a), and White (2010): what's troubling about the causal origins of our beliefs turns out to be a worry about the epistemic significance of disagreement.

The chapter aims at showing that constraints that have to do with disagreement do not pose a problem for objectivists for three reasons: given plausible assumptions about belief formation, one of Tersman's constraints is *impossible* to violate. Another constraint proves *irrelevant* for the objectivist's cause, and even if both constraints were acceptable, there is good empirical reason to conclude that moral disagreement that would violate Tersman's constraints would be *implausible*. The chapter thereby vindicates the legitimacy of third-factor accounts as explanations in metaethics as far as worries about disagreement are concerned.

Section 6.2 introduces third-factor explanations in relevant detail; Section 6.3 reconstructs Tersman's account; Section 6.4 showcases the wider implications for the prospects of the aforementioned disagreement view. Section 6.5 contains my criticism, and in section 6.6, I consider a reply on behalf of Tersman that puts pressure on my objection by appealing to the epistemic significance of higher-order evidence. I show that the rejoinder fails and conclude that worries about disagreement do not constrain third-factor accounts. That's bad news for proponents of the disagreement view but, insofar as third-factor explanations are prima facie legitimate, the conclusion of this chapter should be a boon for moral objectivists.

## 6.2 Third-Factor Explanations

Let M be the moral fact that you have conclusive reason to bring about a state of affairs ϕ. Let N be the non-moral features of ϕ. Suppose that evolutionary considerations explain why actions that brought about states of affairs with property N were adaptive, that is, they increased the actor's relative fitness. The relation of the moral fact, M, and the evolutionary explanation of our tendency to value acts with features N is then explained by a *third-factor* f. f is a bridge law of the form 'N is M' that links the natural state of affairs with a moral value. For example, Enoch's proposed third-factor is that "survival is at least somewhat good" (Enoch 2010).[6] We can assume that organisms that have a tendency to prefer things that aid survival prosper, whereas organisms that don't prefer things that aid survival fail in evolutionary terms. This evolutionary story at least partly explains why humans believe that survival is good. The bridge law, then, explains why such beliefs are also true. Thus, given an evolutionary explanation of why our ancestors favoured N, the third factor thereby also explains why M is instantiated:

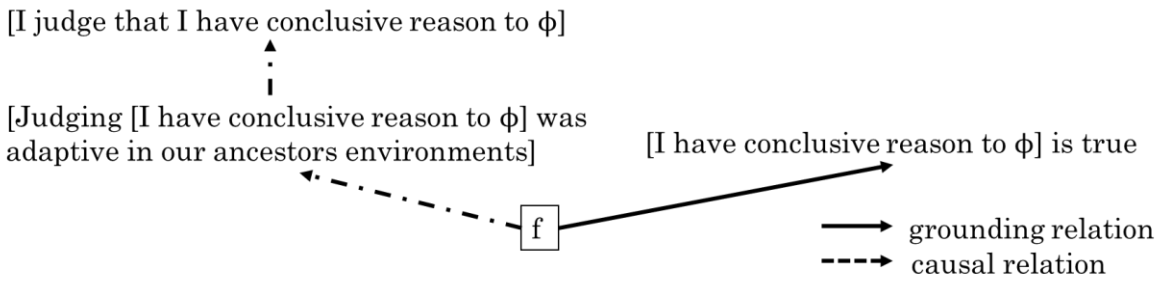


*Figure 6.2 Structure of a third-factor account*[7]

If third-factor accounts work, then there is a "gap in the debunker's evolutionary argument against moral objectivism; a gap through which the non-sceptic might try to sneak out" (Tersman 2017: 767). Crucially, third-factor explanations rely on assumptions about the truth of a substantive moral bridge principle of the form 'N is M'. As we have seen, for example, Enoch assumes that

[6] See Behrends (2013) for a detailed discussion of Enoch's account. For related third-factor accounts see Brosnan (2011), Skarsaune (2011), Talbott (2015), Artiga (2015), Street (2008b), and Copp (2008).

[7] Adapted from Berker (2014: 230).

'survival (N) is at least somewhat good (N)'. Proponents of third-factor accounts do *not* aim to *justify* moral norms, but to *explain their reliability*. This is as it should be: the challenge raised by the evolutionary debunker of morality is that there is *no* explanation of the reliability of moral beliefs; proponents of third-factor accounts try to provide such an explanation.[8]

Not surprisingly, the main discussion point about third-factor explanations themselves is whether assuming the truth of a substantive moral claim, that is, a normative claim rather than a claim about why humans adopt certain moral norms, in response to the debunking challenge is legitimate. Critics claim that it begs the question (Street 2016), but they are met with a *tu quoque* by their opponents: reliability *in any domain of belief formation*, understood as truth-conduciveness, can only be explained by assuming the truth of some beliefs of that domain (Berker 2014; Tersman 2017: 766; Vavova 2014a). For example, consider an explanation of the reliability of our perceptual beliefs. We can explain the reliability of perceptual beliefs by appealing to the theory of evolution, biological and psychological facts about perception, and how perceptual beliefs influence behaviour. It seems very plausible, after all, that our perceptual beliefs must be reliable since creatures that are mostly wrong in such domains "have a pathetic but praiseworthy tendency to die before reproducing their kind" (Quine 1969: 126). But in construing, testing, and defending the theory of evolution, in describing the biological and psychological facts, and in observing how our perceptual beliefs influence behaviour, we rely on those very beliefs. We cannot defend their reliability without relying on those very beliefs in doing so (Tersman 2017: 766).

Hence, it would seem that debunkers rely on the very same assumptions that they deem question-begging if made by objectivists. Tersman, for that matter, accepts the pro tanto legitimacy of third-factor accounts:

[8] Simply claiming that the moral beliefs are justified *because* we hold them would commit a naturalistic fallacy. No proponent of a third-factor account attempts this. Our moral beliefs could be justified by a third-factor account A insofar as it explains why we hold beliefs that are true according to A. Those beliefs would be reliably true and thus justified on a simple externalist account of justification. Again, however, explaining justification is not the issue raised by evolutionary debunkers, because they assume that there is some account that explains how moral judgements are justified and *then* ask what explains the reliability of moral judgements.

> [T]he non-sceptic has a greater room for manoeuvring than one might initially think … [A] non-sceptic can hope to accommodate the evolutionary account by invoking a "bridge principle". (Tersman 2017: 767)

That is, there is a 'bridge' between the evolutionary explanation of why humans endorse certain moral norms and the truth of these norms as conceived by moral objectivists. To discuss Tersman's challenge, I will assume for the sake of argument that the substantive moral assumption at the heart of a third-factor explanation is prima facie legitimate. For reasons of space, we must now sidestep the deep epistemological debate about the legitimacy of using beliefs produced by a faculty in explaining the reliability of that faculty.[9]

That is, there is a 'bridge' between the evolutionary explanation of why humans endorse certain moral norms and the truth of these norms as conceived by moral objectivists. To discuss Tersman's challenge, I will assume for the sake of argument that the substantive moral assumption at the heart of a third-factor explanation is prima facie legitimate. For reasons of space, we must now sidestep the deep epistemological debate about the legitimacy of using beliefs produced by faculty F in explaining the reliability of f.

## 6.3 Constrains for Third-Factor Explanations

Tersman's main contention is that "not just any bridge principle [that] generates the conclusion that the target beliefs are reliable does the trick for the non-sceptic" (Tersman 2017: 767). Hence he proposes three constraints for evaluating the "plausibility" of possible bridge principles on behalf of non-sceptics, of which two constraints are relevant for present purposes.[10]

9 There may be an interesting relation between third-factor explanations and Clarke-Doane's (2016a) response to debunking challenge, which focuses on the modal security of (basic) moral beliefs. The latter may be constrained by the extent of actual moral disagreement, too, as Clarke-Doane himself suggests (2016a: 29). Addressing the relevance of Tersman's constraints for Clarke-Doane's account, however, would require a detailed comparison of third-factor explanations to Clarke-Doane's response that is beyond the scope of this chapter. One apparent disanalogy, however, is that the former aims at showing why a large swath of our current moral beliefs is true, whereas the latter only aims at showing that that our moral beliefs are modally secure assuming that they are true.

10 Tersman's also demands that the objectivist's third-factor account must not be self-defeating. This constraint is entailed by the constraint an acceptable bridge principle must

First, Tersman demands that *an acceptable bridge principle must be supported (i.e. we have reason for accepting it)* by the "beliefs whose reliability is to be established [by the objectivist]" (Tersman 2017: 767). On the one hand, objectivists need some (epistemic) justification to invoke a particular bridge principle, and the required justification must come from the beliefs whose reliability the objectivist wants to defend (Tersman 2017: 768). In other words, there must be a support relation from the contents of the beliefs the objectivist wants to defend to the bridge principle employed in the third-factor account. Moreover, argues Tersman, the content of the beliefs whose reliability is to be established must be "sufficiently varied and rich" and "cannot merely consist of uncontroversial platitudes" because otherwise, the third-factor account would be "underdetermined" (Tersman 2017: 768). Thus, objectivists must feel the pull of two conflicting demands: on the one hand, endorsing a set of moral beliefs that is sufficiently varied and rich raises the probability of radical disagreements about some of the members of that set. Alternatively, limiting the size of the set of moral beliefs whose reliability they want to establish may keep their contents uncontroversial, but then that set might fail to provide the required justificatory support for the bridge principle.[11]

Second, Tersman requires that the bridge principle allow us to provide a theory of error in case there is disagreement between people who endorse moral beliefs that are explained by the third-factor account of choice (the disagreement has to be about *those* beliefs, of course). To illustrate Tersman's second demand, suppose that your belief that 'eating sweets is morally permissible' is influenced by selective pressures towards having a sweet tooth and it is true that 'sweet-tasting food is the best'. If I believe that 'eating sweets is morally impermissible', also because of evolutionary pressures, then we have to be able to explain what went wrong in my belief. If we cannot explain how I came to hold that belief as a cognitive shortcoming, a lack of imagination, or any other relevant failure, then we

be supported by the beliefs whose reliability is to be established insofar as the set of beliefs whose reliability is to be established can only provide sufficient support for the bridge principle unless it is not in tension with the bridge principle.

[11] Note that Tersman does not claim that there *is* radical disagreement about any bridge principles. As such, he poses a challenge to rather than an argument against non-sceptics.

would have to conclude that we are in radical disagreement about whether sweet-tasting food is the best or not. This radical disagreement, Tersman contends, would then cast doubt on the truth of the chosen bridge principle 'sweet-tasting food is the best'. Most importantly, Tersman argues, both constraints combine as follows:[12]

> [A] third-factor account is plausible only if it generates the conclusion that there is a sufficiently varied and rich set of moral claims about which there is no radical disagreement. (Tersman 2017: 769)

Tersman thus concludes that,

> the plausibility of a defence of the reliability of our [objectivist] moral beliefs […] is going to depend on *which types of disagreements actually exist* [and] this is why appeals to disagreement might play a *crucial dialectical role* in the debunkers' strategy (Tersman 2017: 769 emphasis added)

Note that Tersman relies on a *conciliatory* view about disagreement, according to which it is rational to at least reduce confidence in beliefs about which there is radical disagreement (e.g. Elga 2007). That view is controversial, but discussing it would lead us too far afield into the epistemology of disagreement, and thus I will assume it for the sake of argument.

Tersman's challenge is now in full view. Depending on how widespread radical moral disagreement is, the third-factor response to the debunker's challenge might yet turn out to be a dead-end street.

## 6.4 Implications for the Reliability View

Tersman's argument has potential implications beyond the reliability view and the debate about third-factor accounts by providing indirect support for the *Disagreement View*. According to the disagreement view, there are consequences for the epistemic status of our moral beliefs due to the Darwinist view of morality if and only if the Darwinist view of morality implies that there is counterfactual disagreement about human moral beliefs (Bergmann and Kain 2014; Bogardus 2016; Mogensen 2016a; White 2010).

[12] By 'generating the conclusion' Tersman means that there is a bridge principle of the form 'N is M' which is supported by the set of moral propositions that form the content of the beliefs whose reliability the objectivist wants to establish.

If the disagreement view were correct, the most commonly cited explanations of the epistemic significance of the Darwinist view of morality would be mistaken. Contrary to many discussions, the epistemic problems that may arise by uncovering the causal origins of our beliefs would not be due to epistemic insensitivity, a lack of safety, the presence of irrelevant influences, striking contingencies, accidentally true beliefs, or worrying historical variability of moral beliefs.[13] Instead, the problem, if there is one, will turn out to be due to the epistemic significance of counterfactual disagreement, or so the proponents of the debunking-disagreement thesis claim.

Here is how Tersman's challenge may provide indirect support for the disagreement view. Suppose that the reliability view was correct, but third-factor replies work. In that case, undercutting defeat due to evolutionary explanations of morality could be prevented by a third-factor explanation, which could be taken as defeating the evolutionary defeater.[14] However, if Tersman is correct and disagreement constrains or altogether invalidates third-factor explanations, then defeat due to the evolutionary debunking explanation cannot be prevented by third-factor accounts, and our objective moral beliefs are undercut. Thus, disagreement is ultimately crucial for evaluating the epistemic significance of evolutionary debunking explanations, just as proponents of the disagreement view claim. Figure 6.3 depicts the relation of the disagreement view and Tersman's challenge. Arrows in the graph signify an 'attack' or defeat relation, which can be between arguments or theoretical positions (in circles) or between an argument or position and an attack relation.

[13] For reasons of space, I cannot fully introduce these proposals here. For a good overview, see Wielenberg (2016a).

[14] See Pollock and Cruz (1999: 200ff) and chapter 4 for a discussion of defeater-defeaters.

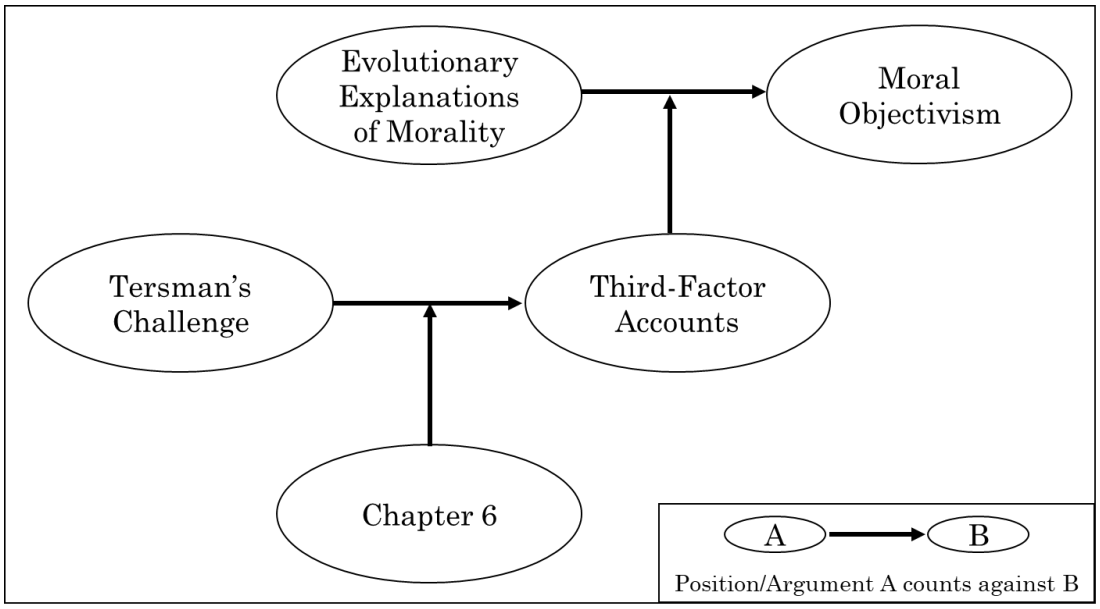


*Figure 6.3 Indirect support for the debunking-disagreement thesis*

To further illustrate the significance of Tersman's challenge, it should be noted that *any* implication that follows from a Darwinist view of morality that a) threatens to undercut all objectivist moral beliefs and b) may itself be defeated by a third-factor explanation could be plucked into the top-left ellipsis in Figure 6.3. Thus, notwithstanding the direct implications of the Darwinist view of morality (of which there might be several valid ones, of course), objectivists could be forced to a sceptical conclusion *only* because of the epistemic significance of disagreement. So, if Tersman's constraints would successfully disallow all third-factor explanations, the disagreement view gains strong indirect support.

However, in the next section, I argue that the constraints proposed by Tersman do not constrain third-factor accounts.

## 6.5 Unconstrained by Moral Disagreement

My criticism has two prongs. First, Tersman's constraints miss the mark against moral objectivists. I show that the first constraint is trivially satisfied on plausible assumptions about the set of beliefs whose reliability objectivists seek to defend and the second constraint is irrelevant to the objectivist's cause. Second, recent ethnographic studies suggest that even if Tersman's criteria were not trivially satisfied, there is a set of moral beliefs that is varied and rich and not subject to any disagreement, thus satisfying Tersman's criteria.

### 6.5.1 Relevant Moral Disagreement is Impossible

Tersman's first constraint is plausible. However, it is unclear at a critical junction. The relevant sense of "support" (Tersman 2017: 768) between the beliefs whose reliability is to be established and the bridge principle needs to be clarified. However, no plausible understanding of support makes it the case that Tersman's constraint could be violated. To support this claim, I will proceed in two steps. First, I consider the set of the contents of the beliefs whose reliability is to be established and how that set gets 'populated'.[15] Next, I consider whether there is a notion of 'support' between the contents of that set and the bridge principle such that there could be radical disagreement between believers who endorse propositions in the set of beliefs whose reliability is to be established (disagreement, that is, about the contents of that set).

The first step is to become clear what the beliefs are whose reliability is to be established. Given that objectivists, by hypothesis, aim to defend the reliability of those beliefs, it stands to reason that they should aim to establish the reliability of only those beliefs that are worth keeping. Any viable epistemology will provide relevant constraints. For example, as an externalist process reliabilist, you would want to keep the beliefs that are formed by a reliable method. As an internalist evidentialist, you would want to keep the beliefs that, very roughly, are sufficiently supported by the evidence available to you or that *seem* to be reliable (to you). Moreover, whichever structure of justification you defend, be it a pyramidal foundationalist structure or a coherentist picture, the justified beliefs that are the beliefs whose reliability is to be established stand in justification-conferring relations to each other. Though it is logically possible that the set of the contents of the beliefs whose reliability is to be established may consist of (sets of) mutually independent propositions, objectivists would plausibly want to defend propositions that are interconnected through some normative theory. Thus, we can assume that the beliefs whose reliability is to be established will be prima facie justified and, given their content, stand in justification-conferring relations to each other. So

[15] I make a distinction between *the beliefs whose reliability is to be established by the objectivist* and *the set of the contents of the beliefs whose reliability is to be established by the objectivist* to make clear that there are justificatory relations between propositions, but disagreement

much for the structural aims of 'populating' the set of beliefs whose reliability is to be established. Before proceeding to the question about support, consider a sceptical worry.

A moral sceptic might ask the following question: how can we select the beliefs whose reliability is to be established so that they stand in justification-conferring relations to each other? However, this is just the problem of determining which moral beliefs we can hold rationally or justifiably. There might be good reason to be sceptical about finding such an account. For example, we might doubt that there is a method through which we can determine what the moral beliefs are that are worth keeping – which ones are worth defending? Should we defend the reliability of the belief that the death penalty is morally impermissible or the reliability of the belief that the death penalty is morally permissible? This is a difficult question, of course, but it is a question that can be bracketed in discussing the debunking challenge, given that the debunking challenge is supposed to be distinct from the general sceptical challenge about having justified moral beliefs *at all*. We should, therefore, assume that once set of beliefs whose reliability is to be established is populated, its members are prima facie justified in mutually reinforcing support relations.

Let us now turn to the second step of my first objection and consider different notions of support. The question is whether the beliefs whose reliability is to be established can *support* the bridge principle (satisfying Tersman's first constraint) and yet it is possible that there be radical disagreement about these beliefs (violating Tersman's second constraint), as is essential to Tersman's argument.

Suppose the relevant support relation between the set of the contents of the beliefs whose reliability is to be established, or the set of target beliefs MT, and the bridge principle f is *deductive* inference, as depicted by the arrow in Figure 6.4:

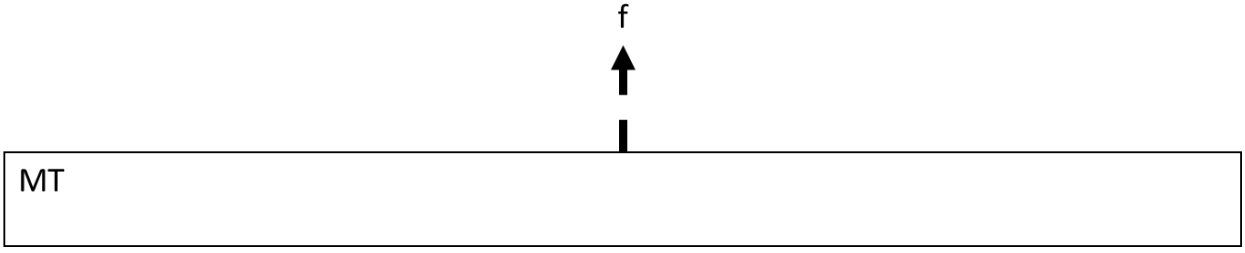


*Figure 6.4 A support relation from moral beliefs to a bridge principle*

In the case depicted in Figure 6.4, the contents of the beliefs whose reliability is to be established provide conclusive reason for the bridge principle. Suppose that all propositions in the set of the contents of the beliefs whose reliability is to be established (MT) are required to deduce the bridge principle f. In that case, there cannot be radical disagreement amongst people who endorse the beliefs whose reliability is to be established. The propositions would have to be inconsistent for radical disagreement to be possible (amongst the believers of those propositions), which is impossible given the assumption that the entire set of contents of the beliefs whose reliability is to be established are part of the bridge principle's premises.

A problem seems to arise if there is another set of propositions that is mutually inconsistent with the set of the contents of the beliefs whose reliability is to be established. In that case, the propositions in both sets taken together would imply anything by material implication, including the bridge principle. Thus, though it is obvious that objectivists should only take consistent propositions to be part of the set of the contents of the beliefs whose reliability they seek to establish, as argued above, it seems problematic that there are many internally consistent but mutually inconsistent sets of (moral) propositions. The union of those sets propositions would imply the bridge principle, but radical disagreement amongst people who endorse the premises of the bridge principle is certainly possible, contrary to what I have argued in the previous paragraph.

The answer to this worry depends on whether the intersection of the deductive closures of the relevant sets is empty or not. If the intersection of the deductive closures of the relevant sets is non-empty, objectivists have an easy way out: they could take beliefs in the contents of the intersection as the beliefs whose reliability they seek to establish. For example, in light of Parfit's (2011b) discussion, we might be 'climbing the same mountain from different sides' in normative ethics, and so it would be plausible that the intersection of the deductive closures of both sets is non-empty. In that case, the argument of the previous paragraph applies, and Tersman's first constraint is trivially satisfied. If the intersection is empty, however, then there are at least two mutually inconsistent sets of moral propositions. To illustrate, suppose that one set contains characteristically deontological claims and the other characteristically utilitarian

claims. If we cannot legitimately choose one set over the other, based on normative theorising, then there will be radical disagreement about moral matters and a conciliatory view about disagreement will imply that no moral belief remains justified. Of course, such fundamental moral scepticism is a serious possibility, but one that can be set aside in the present context. That is because both proponents and opponents of moral objectivism commonly suppose that the evolutionary debunking challenge is an *additional* threat to moral objectivism (cf. Vavova 2015; Wielenberg 2016). If there were fundamental moral disagreement to the extent that we could not rationally chose between different and mutually inconsistent normative theories, then justification of moral beliefs seems forlorn from the start (as long as we assume moral objectivism). Worries about evolutionary debunking arguments would be superfluous (cf. reference omitted B). Thus, it is legitimate to assume that we can legitimately choose one set amongst the set of mutually inconsistent sets of moral propositions that together entail the bridge principle as the set of beliefs whose reliability is to be established. In that case, however, there cannot be radical disagreement about the propositions in that set, as argued above.

Tersman might object as follows: suppose that only *some* propositions in the set of the contents of the beliefs whose reliability is to be established, MT, are required to deduce the bridge principle. Let the propositions required to deduce the bridge principle from MT be the subset $MT_f$. Again, there cannot be radical disagreement about the members of $MT_f$ because of the justificatory structure of MT: the members of MT stand in justificatory relations to each other, and they entail the bridge principle f. Hence, a subset of MT cannot fail to entail f.

It might seem possible that there is radical disagreement between believers who endorse the members of $MT_f$ and those that endorse the members of $MT_f^C$ (the complement of $MT_f$, i.e. the members of MT that are not members of $MT_f$). However, as figure 5.2 illustrates, $MT_f^C$ is a proper subset of MT. If either of the above methods is used to determine the contents of MT, then $MT_f^C$ can only contain members that are consistent with $MT_f$.

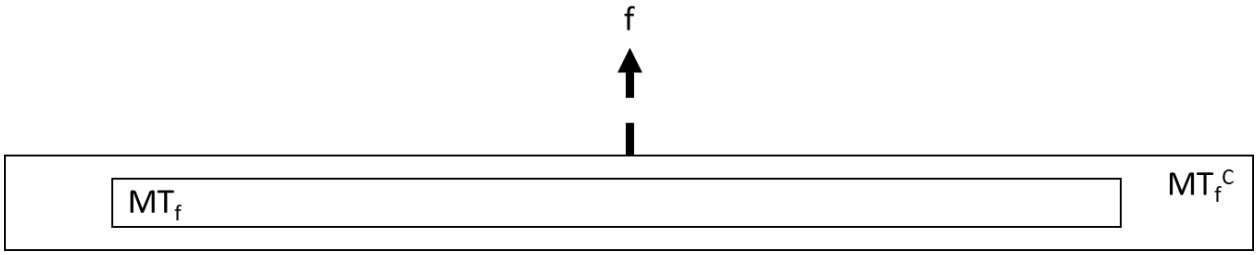


*Figure 6.5 Three ways to support third-factor accounts*

Suppose instead that support is based on an inductive argument that goes from MT to f. The situation would be as in Figure 6.4, but with an inconclusive support relation. If the inductive base is appropriate to confer justification on f, then the inductive base must be consistent. So, as before, there could not be radical disagreement amongst believers who endorse the members of MT. The inductive case can of course be defeated by bringing to light new information that was not but should have been part of the inductive base and that would not serve as a base for the inductive case. However, that would mean that MT is incomplete. But adding a member to MT would follow the same constraints as outlined above, in which case there could not be disagreement either.

Therefore, on any plausible understanding of 'support', if the set of the contents of the beliefs whose reliability is to be established supports a bridge principle, then radical disagreement about the beliefs whose reliability is to be established is impossible. There might be disagreement about beliefs other than those whose reliability is to be established, but it will not be disagreement that could make trouble for third-factor accounts. Let's look at Tersman's second constraint next.

### 6.5.2 Possible Moral Disagreement is Irrelevant

Tersman's second constraint was that there be no radical disagreement about the propositions that form the content of beliefs that are causally explained by the third-factor. However, we will see that beliefs that are causally explained by the third-factor are relevant for anti-sceptics only insofar as they are *also* members of the set of contents of the beliefs whose reliability is to be established.

Tersman's constraint concerns the belief that share a common (ultimate) causal background factor. Let the contents of these beliefs form the set U, as

depicted in Figure 6.6. Tersman demands that there be no radical disagreement about the propositions in U. Clearly, some propositions in U might be inconsistent with the propositions in the set of the contents of the moral beliefs whose reliability is to be established (MT). Thus, U and MT might diverge, as Tersman suggests. However, the intersection of both sets is the only one that should worry objectivists. Given that the intersection will be within the set of the contents of the beliefs whose reliability is to be established there cannot be radical disagreement about the relevant area of the beliefs that are causally explained by the third-factor.

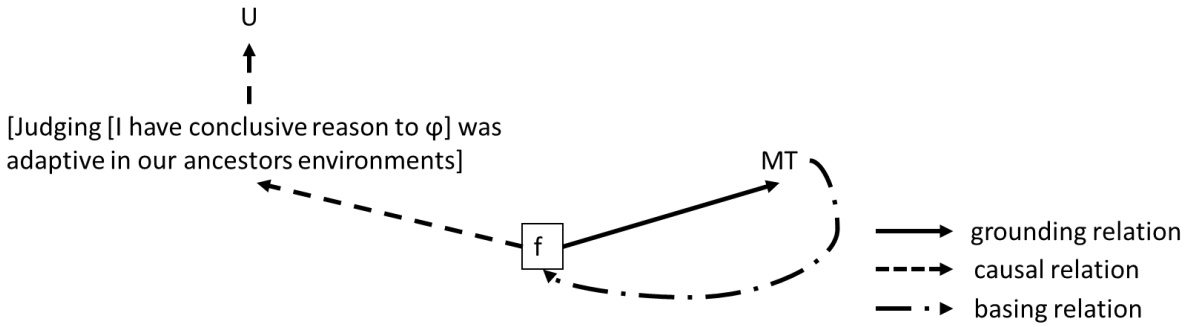


*Figure 6.6 Beliefs that are causally explained by the third-factor*

To support the claim that objectivists need only worry about propositions that are both in U *and* in MT, I show that beliefs caused by the third-factor need not be amongst the beliefs whose reliability the objectivist needs to defend. Thus, disagreement about the beliefs that are causally explained by the third-factor is possible, but that it turns out to be irrelevant for our assessment of third-factor accounts.[16] For example, suppose the causal factor in the third-factor account is 'aids survival.' It might be that some beliefs are ultimately causally explained by actions or events that have this property and nonetheless conflict. A good example is the killing of relatives, which might aid survival when living in polar regions and thus lead to a more favourable opinion of that practice as compared to, say,

[16] There are three further possibilities about the relation between the contents of the beliefs that are causally explained by the third-factor and the set of the contents of the beliefs whose reliability is to be established that I do not discuss in greater detail because they are easily dealt with: a) the former is a proper subset of the latter, b) there are propositions that are in the former but not in the latter but consistent with the latter, c) there are propositions that are in the former but not in the latter and inconsistent with the latter. Given the argument in 6.5.1, there cannot be relevant disagreement in a), objectivists could just adopt the union of both sets in b), and drop the intersection of inconsistencies in c).

the prevailing views in Western countries. The set of contents of the beliefs whose reliability is to be established also has members that are *not* members of the set. For example, take a complex belief like <the right to determine what happens with one's body trumps the concern for the life of a foetus>. The third-factor identified by objectivists may not cause that belief, but it might nonetheless be a member of the beliefs whose reliability is to be established by objectivists. At the same time, the set of contents of the beliefs that are causally explained by the third-factor and the set of contents of the beliefs whose reliability is to be established cannot be *disjunct* because that would violate the constraint that the bridge principle must be supported by the beliefs whose reliability is to be established.

It is possible that the contents of the beliefs that are causally explained by the third-factor form a proper subset of the set of contents of the beliefs whose reliability is to be established (U⊆MT). Thus, the relevant beliefs, about which there must not be radical disagreement, are the beliefs that are causally explained by the third-factor and amongst the beliefs whose reliability is to be established. The considerations above have shown, however, that radical disagreement about those beliefs is impossible given plausible assumptions about the population of the set of beliefs whose reliability is to be established.

It might be objected that the set of beliefs whose reliability is to be established might be a proper subset of the set of beliefs that are (ultimately) causally explained by the third-factor (U⊇MT). The beliefs about the contents of U might get seem to 'ascribed reliability' by the bridge-principle insofar as objectivists have identified the relevant causal factor as conducive to reliability. In that case, some beliefs that are causally explained by the third-factor might conflict with the beliefs whose reliability is to be established (and yet all seem supported by the bridge principle). That seems to create the predicament of a radical disagreement that cannot be easily resolved. For example, if both your belief that 'killing your elderly relatives is permissible' as well as my belief that 'it is not the case that killing your elderly relatives is permissible' could be explained in reference to the fitness-enhancing effects of environmentally-sensitive family management. Here the *causal* explanation is in line with the third-factor account, but the resulting beliefs are not because we have two inconsistent sets of beliefs. That situation is possible if 'ascribing reliability' – or the support relation from the bridge principle

to the beliefs whose reliability is to be established and the beliefs that are ultimately causally explained by the third-factor – is understood as an explanation that shows that the evolutionary process that influenced our moral beliefs creates beliefs that are mostly, but not always, true. Hence, the process that leads to the beliefs whose reliability is to be established can be reliable even if it gives rise to some false beliefs.

However, even if the contents of the beliefs whose reliability is to be established form a proper subset of the contents of the beliefs that are ultimately causally explained by the third-factor, there is no special problem in sorting out the possible disagreement. The answer is clear once we recognise that explaining the reliability of our beliefs by means of an evolutionary process is different from explaining the support that beliefs that are ultimately causally explained by the third-factor gain from the bridge principle. Evolutionary considerations might explain why we tend to form the beliefs that are explained by the third-factor. However, even if the contents of the evolutionarily-explained beliefs conflict with the contents of beliefs whose reliability is to be established, the bridge principle will either imply that the evolutionarily-explained-beliefs are unsupported, or reason to include the evolutionarily-explained beliefs it supports amongst the beliefs whose reliability is to be established.

Therefore, there are no reasons to think that there could be radical disagreement about the beliefs that underwrite the objectivist's third-factor account. Thus, disagreement between thinkers who endorse beliefs that are causally explained by the third-factor account is possible, but only insofar as those beliefs are not amongst the beliefs whose reliability is to be established. In that case, however, the disagreement is irrelevant for objectivists.

### 6.5.3 Actual Moral Disagreement is Implausible

Even if Tersman's criteria were not trivially satisfied, recent ethnographic research provides strong support for the view that there is a rich and varied set of foundational moral beliefs that is not subject to radical disagreement. In other words, even if Tersman's conditions could work, it is *implausible* that there would be actual radical disagreement amongst the members of MT. So, objectivists would pass Tersman's test.

I should make clear that this section addresses the question whether actual moral disagreement helps the debunker's case, assuming that Tersman's main point (that disagreement has an important role to play in debunking arguments) is correct. My arguments in section 6.5.1 and 6.5.2 might fail, and disagreement could be relevant in the way Tersman describes and potentially negative for moral objectivists, but here I argue that disagreement does not have negative effects for moral objectivists. Of course, since I am making a claim about the actual extend of moral disagreement, this section is to some extend speculative: efforts are under way, but the empirical investigations required to establish the degree of actual moral disagreement are still in their infancy.

A widely endorsed hypothesis about the evolution of morality is that moral beliefs serve a cooperative function. The view is endorsed by authors across the metaethical spectrum (Gibbard 2003; Joyce 2006; Kitcher 2011; Street 2006; Wielenberg 2014). A recent defence of the hypothesis comes from Curry (2016). Using a game-theoretic approach, Curry predicts that moral beliefs will be concerned with four problem-centred domains: (1) the allocation of resources to kin; (2) coordination to mutual advantage; (3) exchange; and (4) conflict resolution (Curry 2016: 30). For reasons of space, the interesting details of these considerations cannot be recounted here. However, in a recent study, Curry, Mullins, and Whitehouse show that their predictions are born out in the ethnographic record of 60 societies (Curry et al. 2017). Behaviour that fell into the four categories specified by Curry (2016) was regarded as morally good by all studied societies. The finding implies that there is tremendous agreement about several moral domains.

Of course, there might be cases of intractable, radical disagreement about moral matters too. For example, Doris and Plakias (2008) suggest, based on the work of Nisbett and Cohen (1996), that "North/South differences in attitudes toward violence and honour might well persist in ideal discursive conditions" (Doris and Plakias 2008: 331). However, as they themselves acknowledge, perhaps a tad too modestly, "one case is not an induction base" (ibid.). More importantly in the present context, the cases of radical disagreement discussed in the literature (the locus classicus being Brandt's (1954) discussion of Hopi ethics) do not concern the domains of shared moral beliefs identified by Curry et al. and it is an open

possibility that the apparent moral disagreements turn out to depend on non-moral disagreements.

Therefore, leaning on the hypothesis of morality as stimulating cooperation, objectivists can argue that widely shared positive evaluations of kinship relations, cooperation, exchange, and conflict resolution provide an adequate basis for coming up with a third-factor account that meets Tersman's criteria.

## 6.6 Rejoinder: Higher-Order Evidence

Disagreement does not seem to impugn the epistemic legitimacy of third-factor responses. However, Tersman's proposal seems suggestive of a wider problem that has to do with higher-order evidence. The thought goes as follows. The fact that there is actual disagreement about even deeply held moral beliefs, often even with those whom we consider to be our peers, provides evidence that our moral beliefs are not reliable. Debunking challenges provide us with higher-order evidence, not about particular moral beliefs, but rather about *our general ability to form reliable moral beliefs*, at least concerning the topics about which we disagree.

Perhaps interpreting the problem of disagreement as a higher-order evidence problem helps Tersman's account. According to Christensen (2010), a distinctive feature of sceptical higher-order evidence about p is that it requires bracketing of one's evidence for p in assessing whether one should or should not endorse p. If that is correct, and if we can interpret the debunker's challenge as a higher-order evidence challenge, then the support from MT for f must be bracketed in evaluating disagreement about f. The question becomes how likely we would be to be correct in picking a bridge law f, once we bracket the available evidence for f (such as our moral intuitions, several hundred years of philosophising about f, and so on). This response would avoid the problem with the critical notion of support that is at the heart of Tersman's original argument.

However, the appeal to higher-order evidence confuses two challenges, and so the appeal to higher-order evidence does not help Tersman's account. The debunker challenges moral objectivists to vindicate a third-factor account given the assumption that at least some moral beliefs have positive epistemic credentials. The higher-order evidence challenge to moral objectivism, in contrast, casts doubt precisely on the validity of that assumption.

Moreover, higher-order evidence challenges seem to break down in the limiting case where the reliability of beliefs about an entire domain is called into question. If higher-order evidence is applied to all beliefs of a type and there is no possibility of using other beliefs to vindicate the justification of that type, then higher-order evidence reduces to a general sceptical challenge. Tersman himself acknowledges that beliefs of the moral type are properly 'isolated'. Thus, given that the debunking challenge is supposed to be distinct from the general sceptical challenge, the higher-order evidence interpretation cannot save Tersman's disagreement-based constraints on third-factor explanations.

## 6.7 Concluding Remarks

Once the prima facie legitimacy of third-factor accounts is granted, as many scholars do, disagreement is not a problem for third-factor explanations, contrary to Tersman's suggestion.

In particular, plausible assumptions about the structure of justification and the nature of the debunking challenge show that there cannot be radical disagreement amongst the beliefs supported by the third-factor account. And even if it were possible, there is good empirical support for thinking that there is no actual radical moral disagreement.

In conclusion, worries about disagreement do not rationally constrain third-factor explanations in metaethics. Plausibly, this finding generalises to other domains, such as mathematics. More generally, Tersman's proposed constraints on third-factor explanations do not serve to support the debunking-disagreement thesis. Insofar as third-factor accounts are prima facie legitimate, this is good news for moral objectivists.

# 7 Defeat, Reliability, and the Etiquette Conception of Defeat

## Reader's Guide

Will the reliability view save the day for the survival of (evolutionary) defeat? According to the reliability view, you will recall, evolutionary explanations of morality show us that we can explain quite well the content of many moral beliefs without assuming their truth, and this compels us to explain the reliability of our moral beliefs (in the sense of why they are true significantly more often than chance would predict). In conjunction with the epistemic principle that says that we ought to give up any belief whose reliability is in principle impossible to explain, the reliability view might explain how evolution undercuts all objectivist moral beliefs (Benacerraf 1973; Field 1989, 2001).

The reliability view is a very prominent account of evolutionary defeat (e.g. Schechter 2010; Tersman 2016; Wielenberg 2016a) and, in light of the previous discussion, it seems to the last hope for the survival of defeat. In the previous two chapters, I argued that the disagreement view of the evolutionary defeat challenge falters, and news about the survival of (evolutionary) defeat has not been gloomy in the initial two chapters, either: we saw that an objectivist account of defeat is required but unavailable to explain undercutting defeat of non-empirically justified beliefs (in chapter 2) and that a posteriori information shovelled up by evolutionary explanations of morality cannot by itself undercut our moral beliefs (in chapter 3).

However, we have also seen that there are problems for the reliability view in section 1.6.3 of chapter 1. Though it might seem plausible that we have to be able to explain the reliability of any of our beliefs (it must at least in principle be possible), this leaves open, first, whether our moral beliefs are shown to be unreliable by evolutionary explanations of morality and, more importantly, whether we ought to give up beliefs after noting a problem with their reliability

(Clarke-Doane 2015, 2017c).[1] There is more bad news to come for the survival of defeat.

In this chapter, I will show that the problem for the reliability view is a serious one indeed and that the reliability view therefore fails. It is false that evolution shows that all our moral beliefs are unreliable and it is false that evolution could undercut all our moral beliefs if it could show that their reliability is lacking. However, I also show that there is hope for evolutionary defeat. My main project in this chapter will be to assess in detail an assumption about the nature of undercutting defeat that grounds the challenge to the reliability view and to argue that the challenge fails.

In addressing the challenge to the reliability view, I will show that we need a more encompassing conception of undercutting defeat if we are to keep open the survival of defeat. I will argue that we have to accept the etiquette conception of undercutting defeat that I already mentioned in the main introduction. Recall that the etiquette conception says that a belief can be undercut by information that implies that the belief does not qualify as knowledge without giving us reason to doubt that the belief is true as well as sensitive and safe.

In this reader's guide, I will briefly touch on the core obstacle to the reliability view and two crucial assumptions that solidify the obstacle. I will then turn to my assessment of the challenge and how to overcome it in section 7.1.

The obstacle for the reliability view can be reconstructed as follows. The reliability view relies on the view that the in-principle impossibility of explaining a type of belief gives one reason to withhold belief in that type of belief. However, why would that the case? Justin Clarke-Doane (2017c) argues that for one to give up one's belief B in light of new information E, E has to reveal there to be a problem with the epistemic sensitivity or the epistemic safety of B. Epistemic safety and epistemic sensitivity are counterfactual conditionals. Recall that, according to sensitivity, S's true belief that p is sensitive if and only if the following

[1] It is worth repeating that proponents of *conditional* debunking arguments, like Street (2006), might claim that *no* moral beliefs ought to be given up, only metaethical beliefs. As pointed out in the section on Method and Presuppositions in the main introduction, however, I assume that moral objectivism is true and that if moral objectivists are correct about the contents of moral beliefs, then we'd have to give up our moral beliefs if the evolutionary defeat challenge succeeds.

counterfactual is true: <had p been false, S would not have believed that p>. According to safety, S's true belief that p is safe if and only if the following counterfactual is true: <if S believes that p, p is true>. Intuitively, Clarke-Doane's demand can be understood as a demand for a kind of epistemic accuracy[2] in the sense that if our beliefs are sensitive and safe, we know that they are true and could not have easily been wrong. This seems to satisfy what we would want, epistemically, from a belief. If new information does not raise doubt about this being the case, Clarke-Doane attests that there is little reason for giving up the belief. That seems plausible, certainly on the orthodox view of defeat that we already encountered in chapter 2. As I will show in greater detail later, Clarke-Doane argues, furthermore, that evolutionary explanations of morality do not threaten the epistemic sensitivity or safety of moral beliefs (if anything, they strengthen the case for the safety and sensitivity of our fundamental moral beliefs). Even if that were not the case, a concern with the reliability of our moral beliefs that does not show how it impugns their epistemic safety or sensitivity does not undercut them, or so the challenge goes.

I will show why the reliability view succumbs to Clarke-Doane's challenge. This will make good on a promise made in the main introduction. It will make clear in greater detail why the reliability view fails. In short, given the assumptions commonly taken in the field, at least some moral beliefs are true in most nearby possible worlds and on most accounts of reliability such beliefs should count as reliable. If reliability is understood in this way, then evolutionary explanations of morality do not raise a problem for the reliability of our moral beliefs.

However, I will also show where Clarke-Doane's attack on the reliability view goes wrong, and though my rebuttal does not reinstate the reliability view, it opens the door for a novel type of undercutting defeat based on the etiquette conception of defeat.

In the remainder of this reader's guide, I want to discuss a couple of assumptions that I presuppose in the discussion to come and which are commonly shared by discussants in the debate. The most relevant assumptions are that at

[2] I do not mean epistemic accuracy in the technical sense as described by Pettigrew (2016).

least some moral propositions are metaphysically necessary and that the standard view on the semantics of counterfactual conditionals is correct. I will follow conventional practice and grant both assumptions for the sake of argument, but briefly clarify them now.

Let's begin with the metaphysical necessity of at least some 'basic' or 'fundamental' moral propositions. After pointing out how metaphysical necessity is commonly understood, I will discuss in somewhat greater detail why moral propositions could be regarded as metaphysically necessary.

Metaphysical necessity is a modality typically considered to be stronger than physical necessity but weaker than logical or conceptual necessity. Traditional examples of metaphysical necessity involve theoretical identity statements such as 'Water is $H_2O$' and 'Gold is the element with the atomic number 79'; both are physically and metaphysically necessary, but not logically or conceptually necessary. The view is most closely associated with Kripke (1984), though there is an abundance of recent discussion (see Cameron 2010).

Many believe that at least some fundamental moral truths are metaphysically necessary (e.g. Enoch 2011b: 172). The metaphysical necessity of at least some moral propositions is grounded in the supervenience of supervenient properties on their supervenience bases . A strong supervenience principle says that if two possible entities are alike in every non-moral respect, they are alike in every moral respect. Strong supervenience seems plausible to many (cf. Rosen 2018). If supervenience is true, then ordinary moral propositions seem metaphysically necessary. Suppose we provide a complete description of the non-normative aspects of some act A, D(A). Suppose also that, given D(A), A seems obviously morally wrong and we (me and you, suppose) therefore judge that it is impermissible to do A. We can express this as ~Per(A). It seems natural to put the case as follows: given that A is so-and-so, D(A), it is impermissible to do A, ~Per(A). Our conditional would have the form

1) $D(A) \rightarrow \sim Per(A)$

Clearly, (1) is a moral proposition: given that A is so-and-so, it is impermissible to do A! It follows from strong supervenience that (1) must be a metaphysically necessary truth: if A is exactly as D(A) says, then it must be the

case that ~Per(A), even though this is not analytic. Thus, though (2) is imaginable, where it is *permissible* to do A, it cannot possibly be true:

2) D(A) → Per(A)

Both worlds (1) and (2) are logically and analytically possible. Are both worlds *metaphysically* possible? Not if one believes that things have essences (cf. Fine 1994). In the same vein, we could have a world where water is made of lithium and not hydrogen. Such a world would also be logically and analytically possible but metaphysically impossible: it is the nature of lithium to have three atoms, or so the argument goes. Of course, these brief remarks do not constitute a defence of the metaphysical necessity of the truth of (at least some) moral propositions. But they should make plausible the view that accepting strong supervenience gives one reason to accept the metaphysical necessity of the truth of (at least some) moral propositions. In any case, this is what most proponents of moral objectivism accept too. Let me now turn to the truth values of counterfactual conditionals, which play a significant role in the rejection of the reliability view.

The standard way to interpret conditionals such as epistemic sensitivity and epistemic safety is as counterfactual conditionals (cf. Lewis 1973). For example, consider the counterfactual <had Oswald not shot Kennedy, someone else would have>. In ordinary language, this claim signals, dubiously, that Kennedy's assassination was inevitable. It is clear, though, that the conditional <if Oswald did not kill Kennedy, someone else did> is different in meaning from the counterfactual considered earlier. The latter conditional is surely true, while the counterfactual depends on whether we believe that there is, say, a conspiracy of assassins that aimed to kill Kennedy. The truth values of counterfactuals are commonly evaluated according to Lewis's possible world semantics (Lewis 1973). On this view, the standard semantic for counterfactual conditionals, only *metaphysically possible* worlds are considered when evaluating the truth value of counterfactuals. Moreover, truth values are assigned, similarly to the material conditional familiar from standard logic, by making the assignment of truth values to the antecedent and the consequent independently of any causal relationship between the two. In ordinary language, both indicative conditionals such as <if I am hungry, I eat> and counterfactual conditionals such as <had I been hungry, I

would have eaten> signal a causal relation between antecedent and consequent. It makes sense to eschew the ordinary way of assigning truth values to the safety and sensitivity counterfactuals, because demanding a causal relation between fact and belief as a requirement for epistemic justification or knowledge is eschewed by virtually all contemporary epistemologists (Ichikawa and Steup 2017). Nonetheless, these critical considerations should be kept in mind when assessing the argument in what follows. Note also that I discuss Clarke-Doane's challenge in respect to the reliability challenge specifically, and not in regard to the evolutionary defeat challenge. In light of my discussion in chapter 2, the relevance for the evolutionary defeat challenge should be clear. Finally, it should also be noted that the chapter implies that the findings of my thesis extend beyond the moral domain, as we will see in the argument against Clarke-Doane's challenge. Let's have a look.

## Abstract[1]

When does new information undercut a belief? According to the 'modal security' principle, a belief can only be undercut by information that raises doubt that the belief is both epistemically sensitive and safe. If true, modal security would immunise domains where beliefs are taken to be true in virtue of mind-independent facts, such as moral objectivism in metaethics or Platonism in mathematics, from undercutting defeat. Extant criticisms of modal security do not show where it goes wrong. This chapter aims at disproving modal security. I defend two novel claims. First, it can be shown that learning that a type of belief fails to qualify as knowledge can undercut those beliefs without raising doubt about their sensitivity and safety. This finding contradicts modal security. Second, rejecting modal security commits us to what I have previously called an etiquette conception of undercutting defeat, according to which even invariably true beliefs ought sometimes to be given up. This finding explains why modal security is false. The etiquette conception is a radical departure from current conceptions of undercutting defeat and, as the chapter shows, the reliability challenge depends on it.

## 7.1 Introduction

Since Gettier's paper of (1963), virtually all epistemologists seem to agree that accidentally true beliefs do not amount to knowledge. For example, Riggs writes that "the immunity-from-luck requirement [for knowledge] is virtually the only thing in the theory of knowledge about which we can claim consensus" (Riggs 2007: 330). Partly as a way to cost the absence-of-luck requirement, many philosophers have proposed a sensitivity requirement for knowledge. According to a simple version of sensitivity, subject S knows that p only if, were p false, S would not believe that p (Becker 2012; Nozick 1981: ch. 3). Sensitivity-based analyses of accidentality have trouble with necessary truths because every belief in a necessary truth automatically comes out as sensitive and thus non-accidental, at least if we apply the standard material interpretation of conditionals. But clearly, believing the truth might be accidental even though the propositional contents of the belief are necessary truths, such as when you just guess the solution to a mathematical question. A similar, though less widely recognised, problem with analysing knowledge of necessary truths holds in the case of safety. According to a simple version of safety, S knows that p only if S could not easily have falsely

[1] The paper on which this chapter is based is under review at the time of writing.

believed that p (Pritchard 2005; Sosa 1999; Williamson 2000). However, a number of cases indicate that S could be manipulated into believing p safely but, intuitively, fail to have knowledge nonetheless (Greco 1999). As Roland and Cogburn put it, "on safety-based accounts of knowledge, knowledge of necessary truths is as cheap, easy, and universal as it is on sensitivity-based accounts" (Roland and Cogburn 2011: 554). These authors argue that satisfying either safety or sensitivity does not seem to be enough for a belief to qualify as knowledge. I suggest in this chapter that the conjunction of safety and sensitivity is also insufficient for a true belief to qualify as knowledge, in particular when we consider propositions that are necessarily true.

My main point in this chapter will be that this rather underappreciated problem of epistemology turns out to be of crucial importance for the widely held metaethical view of moral objectivism.[2] Moral objectivism is the view that (at least some) explanatorily basic moral truths are constitutively and causally independent of attitudes or beliefs. It is often held, as a corrolary of the view, that moral truths are metaphysically necessary. Moral objectivists face the 'reliability challenge': what explains that moral beliefs are reliable, given that the relevant states of affairs that account for the moral truths are causally inert and constitutively independent of humans? Learning that the reliability of objectivist moral beliefs is in principle impossible to explain, the challenge goes, undercuts the justification of those beliefs. Although formulated in paradigmatically externalist parlance, the reliability challenge arises for epistemic internalists too, insofar as it is problematic for a believer to *learn* that he or she cannot explain how his beliefs are reliably connected to the truth (cf. Enoch 2010). Many moral objectivists regard the reliability challenge as their most arduous test (Enoch 2010, 2011b; Parfit 2011a; Scanlon 2014; Wielenberg 2014). Some have argued that the reliability challenge is the real problem behind so-called evolutionary debunking arguments in metaethics (Enoch 2010; Klenk 2017c).

In a series of influential papers, Justin Clarke-Doane has argued forcefully that the reliability challenge can be met (Clarke-Doane 2015, 2016a, 2016b, 2017b, 2017c). In particular, he argues that 'explaining the reliability' of moral beliefs

[2] In fact, it is of crucial importance for objectivist accounts of mathematics, logic, and modality too. For simplicity, I will restrict my discussion to moral objectivism.

means showing that moral beliefs are epistemically safe and sensitive. Roughly, a belief is safe, in Clarke-Doane's sense, if the belief's content does not vary in nearby possible worlds and sensitive if it would not be held in worlds where the belief would be false. According to his 'modal security' principle, beliefs can only be undercut by showing that they fail to be both safe and sensitive. The reliability of moral beliefs is 'explained' in the relevant sense because the relevant set of moral beliefs turns out to be modally secure: they are both safe and sensitive (Clarke-Doane 2016a). If Clarke-Doane is right, then moral objectivists escape the most serious challenge to their view. The alleged rebuttal of the reliability challenge has received considerable attention in the recent literature. An increasing number of philosophers accept that successfully defending a principle such as MODAL SECURITY would be enough to answer the reliability challenge (Baras 2017; Barkhausen 2016; Handfield 2016; Hill 2016; Warren 2017). Others have criticised the principle, but have not shown where it goes wrong (Dogramaci 2017; Faraci 2016; Jonas 2017; Lutz forthcoming; Schechter 2018; Tersman 2016, 2017: 757; Woods 2016).[3]

This chapter aims at disproving MODAL SECURITY by showing that a belief can be undercut despite being modally secure.[4] The strategy is to identify a lack of warrant[5] that is undercutting but not costed in modal terms. I defend two sets of novel claims. First, MODAL SECURITY is false because (1) having a modally secure belief is not sufficient for knowledge and (2) it can be shown that learning that a belief does not count as knowledge undercuts that belief *without giving one a reason to doubt that the belief is both safe and sensitive.* The falsity of MODAL SECURITY, however, does not reinstate the reliability challenge for moral objectivism because the truth of some moral beliefs can be shown to be stable in nearby possible worlds and thus they are reliable in that sense. Second, rejecting MODAL SECURITY based on my argument raises a different epistemological challenge for moral objectivism, which commits us to a novel conception of

[3] Linnebo (2006) has independently addressed the issue, but his solution can be shown to fail; see Berry (2017).

[4] A belief is modally secure iff it is both epistemically sensitive and epistemically safe.

[5] I use 'warrant' in the technical sense to mean *whatever must be added to a true belief to qualify as knowledge.*

undercutting defeat. According to that view, some undercutting defeaters compel us, epistemically, to withhold belief when we learn that the belief fails to qualify as knowledge even though we might have no reason to doubt that the belief is true. I call this novel conception of defeat the *etiquette conception* of defeat because it implies that the aim of belief is *knowledge* rather than mere (non-accidental) truth.[6]

The main upshot of this chapter is that rejecting modal security can be rejected by committing us to the etiquette conception of defeat. The etiquette conception is a radical departure from the orthodox conception of defeat, which suggests that all undercutting defeaters must cast doubt on the truth of the target belief (cf. Pollock 1995). Nonetheless, the etiquette conception explain what's wrong, epistemically, with views such as moral objectivism. The problem is one of epistemic ability, and if that is correct, then focusing on concerns about the truth of moral beliefs is tilting at windmills: the problem with moral objectivism is not epistemic risk, but a problem with believing for the right reason. This shows that there are unobvious, and as of yet unexplored, relations between the burgeoning literature on the reliability challenge, the nature of defeat, and discussions about the norms of belief.

Section 7.2 introduces the reliability challenge and Clarke-Doane's defence of MODAL SECURITY in greater detail. Sections 7.3 introduces my anti-modal security argument and I defend it in sections 7.4 to 7.6. Section 7.7 presents the etiquette conception of defeat and discusses independent reasons for accepting it. I conclude in section 7.8.

## 7.2 The Reliability View and Modal Security

Hartry Field wonders how it could be that, almost always, when mathematicians believe that p, then p is true, and when they believe that p is false, then it is indeed false that p (Field 1989: 25–30). Similarly, David Enoch wonders how, very often, when we accept a normative judgement j, then j is also considered to be true, and very often when we reject j, it is indeed false that j (Enoch 2010: 421). Thus, there is a need to explain the correlation between truths and beliefs.

[6] I say more about this choice of terminology in section 7.7.

According to Clarke-Doane, the reliability challenge is the challenge to provide the required explanation of how we are reliable regarding morals (Clarke-Doane 2015: 87; Schechter 2018: 453ff). Objectivists seem to agree (Enoch 2010; Shafer-Landau 2012; Wielenberg 2014). Clarke-Doane makes three relevant qualifications. First, the challenge is non-sceptical because it is assumed for the sake of argument that at least some moral beliefs are true and that they are defeasibly justified (Clarke-Doane 2017d: 17). Second, he takes the challenge, and his defence, to apply only to the *explanatorily basic moral beliefs*, such as <pain is pro tanto bad>, which form the basis of thicker, more context-dependent moral beliefs (Clarke-Doane 2015: 93). Finally, in line with common practices in the literature, he assumes that at least some of the explanatorily basic moral truths are metaphysically necessary (Clarke-Doane 2012: 320, 2016a: 26, 2016b: 31, 2017c: 28).[7]

The question, then, is what it takes to 'explain the correlation' in the relevant sense. Clarke-Doane argues that the only relevant sense of explaining the reliability must have to do with the epistemic sensitivity and safety (of the explanatorily basic moral beliefs). He understands epistemic sensitivity and safety as follows:[8]

> **Sensitivity$_{CD}$**. Had the contents of our D-beliefs [formed via method M] been false, we would not have believed them [by using M]. (Clarke-Doane 2016a: 26–8)

> **Safety$_{CD}$**. It is false that we might easily have had false D-beliefs [formed via method M].(ibid.)

[7] It should be noted that the assumption that some moral truths are metaphysically necessary is often taken to be crucial in supporting the view that if 'explaining adequately the reliability of moral beliefs' is correctly interpreted as 'showing that moral beliefs are modally secure' then the reliability of moral beliefs can be explained (e.g. Jonas 2017). For reasons outlined in the reader's guide of this chapter, however, I will not question this assumption. Note also that my point against modal security does not rely on the idiosyncrasies of necessary truth, but on general considerations about knowledge. So, there only remains the question whether moral beliefs might *also* fail because of general reliability concerns if we reject the necessity assumption.

[8] Sensitivity and safety are not equivalent contrapositives since they are counterfactual conditionals rather than material conditionals. Note that the problem is thus that there are no worlds in which the fundamental moral belief that p is false (on the assumption that p is true, as in the context of the reliability challenge).

Method M is a placeholder for whatever belief-forming method accounts for moral knowledge. As stated before, it is a crucial assumption of the reliability challenge that some such method is capable of giving us moral knowledge. Both counterfactual conditions differ from familiar safety and sensitivity conditions on knowledge insofar as Clarke-Doane's formulations concern *kinds* of beliefs rather than particular tokens of beliefs. Based on these conditions, Clarke-Doane proposes the following necessary condition on undercutting defeaters:

> **Modal Security**: If information, E, undercuts all of our beliefs of a kind, D, then it does so by giving us reason to doubt that our D-beliefs are both sensitive and safe. (Clarke-Doane 2016a: 31)

Applied to moral beliefs, modal security implies that, for example, an evolutionary explanation of our moral beliefs can only undercut our moral beliefs (objectively construed) if it gives us some reason to doubt that moral beliefs are both safe and sensitive.

Moreover, Clarke-Doane argues that modal security thwarts the reliability challenge even if we accept its ramifications for justification.[9] The explanatorily basic moral beliefs are metaphysically necessary, and so they are sensitive on a standard semantics (Clarke-Doane 2017c: 35; Lewis 1973).[10] We are also led to adopt the explanatorily basic moral beliefs in all nearby possible worlds, or so evolutionary explanations of morality suggest, and so moral beliefs are safe (Clarke-Doane 2017c: 35).[11] Hence, our explanatorily basic moral beliefs are

[9] Essentially the same reasoning applies to the question whether failing the reliability challenge gives us reason to doubt the probability of our moral beliefs, so that's another closed door for defenders of the reliability challenge; see Baras (2017). An adequate interpretation of 'explaining the reliability' cannot demand a causal interpretation, either, not only because that would beg the question against realists but because it relies on an implausible causal theory of knowledge.

[10] One might consider counter-possible worlds to avoid this issue (e.g. Collin 2017). As Clarke-Doane points out, however, such a view would imply scepticism even regarding dry-goods judgements; see Clarke-Doane (2014). Also, it would seem as if metaphysical impossible worlds are very remote from the actual world. It's not clear why explaining the reliability of moral beliefs would require showing that they are true in such remote worlds.

[11] 'Nearby' is vague in this context; see Baumann (2008). One might try to exploit this and argue that fundamental moral beliefs are *not* safe; see Handfield (2016). Partly because the question is vague, I doubt that much progress can be made by pursuing this avenue; see Joyce (2016d).

modally secure and thus failing to solve the reliability challenge cannot undercut them.

## 7.3 The Anti-Modal Security Argument

Given the assumption that the truth of our fundamental moral beliefs is metaphysically necessary, a standard semantic, and the available evidence from evolutionary explanations of morality, it is true that our moral beliefs are reliable in the sense of being modally secure. It is false, however, that the fundamental moral beliefs are off the hook when it comes to undercutting defeat, as I aim to show in the following sections.

Clarke-Doane's defence of modal security relies on the assumption that sensitivity and safety are sufficient for non-underminability. So,modal security can be disproven by showing that we might learn some information about moral beliefs that gives us no reason to doubt their modal security but undercuts them nonetheless. Clarke-Doane asks:

> Why would anyone believe Modal Security? Because it is hard to see why we should give up beliefs in light of information that neither tells 'directly' against their contents, nor against the 'security' of their truth. (Clarke-Doane 2016a: 31)

I will show that we sometimes have reason to give up beliefs in light of information that neither tells against their contents nor the 'security' of their truth, because that information can imply that such beliefs fail to qualify as knowledge.[12] To anticipate my argumentation, my strategy is to exploit a problem with theories of knowledge that rely solely on modal criteria, which are incapable of giving a complete account of knowledge. Leaning on the view that there is a necessary *credit* or *ability* requirement for knowledge,[13] I will show that there are general conditions under which it is guaranteed that a belief is modally secure but not knowledge and that we can learn information that implies as much. Specifically, I defend the following ANTI-MODAL SECURITY argument:

[12] Cf. Williamson (2000); Hirvelä (2017).

[13] Cf. Pritchard (2010, 2012), Pritchard et al. (2012), Sosa (2005, 2007), Zagzebski (2002).

**P1:** It is possible that S gains new information, E, that shows that no D-belief qualifies as knowledge without giving S a reason to doubt that D-beliefs are modally secure.

**P2:** If S gains new information, E, which shows that no D-belief qualifies as knowledge and some D-beliefs ought to be held only if they qualify as knowledge, then all D-beliefs that ought to be held only if they qualify as knowledge are undercut even though they might be modally secure.

**C:** So, it is possible that S gains new information that undercut's S's D-beliefs without giving S a reason to doubt that their D-beliefs are modally secure.

P1 is supported by cases that show that a belief can be modally secure but still epistemically lacking such that the belief is not knowledge. Moreover, I show that these cases can be generalised using two criteria that I will explain below. P2 is based on the assumption that the norm of belief is knowledge, that is, that one should believe that p only if one knows that p and the fact that objectivists are committed to the possibility of moral knowledge. The conclusion contradicts MODAL SECURITY, which said that it is *not* possible to undercut a belief without showing that it is not both safe and sensitive.

## 7.4 Reliability Without Knowledge

It should not be surprising that some modally secure beliefs do not qualify as knowledge, because problems with the sufficiency for knowledge of sensitivity or safety are much discussed in the epistemological literature. However, MODAL SECURITY concerns justification and so it is one thing to note that a thinker may have a modally secure belief that fails to be knowledge and quite another to show that the thinker can learn about his or her lack of warrant such that his belief is undercut. Showing that this is possible is my main contribution towards defending P1.

My defence of P1 proceeds in three steps. First, I discuss two cases to illustrate that sensitivity, or safety, or the conjunction of safety and sensitivity is not sufficient for knowledge. I then show that the problem generalises: there are

cases in which modal analyses are bound to mistakenly imply that the believer has knowledge. Third, I consider and reject recent attempts at salvaging safety-based accounts of knowledge. This will show that believers can learn that their belief does not qualify as knowledge without thereby putting into doubt the modal security of their belief and thus vindicate P1. [14]

### 7.4.1 Sensitivity

Problems with having sensitivity as a necessary condition for knowledge in general have been highlighted by many writers, and I shall not repeat these complaints here.[15] The sensitivity condition is vacuously satisfied in the case of necessary truths (since it is based on a counterfactual conditional) if only possible worlds are considered and thus sensitivity cannot be a sufficient condition for knowledge in the case of necessary truth. The converse of sensitivity, called 'adherence', might be considered for analysing necessary truths. According to adherence, when S knows that p, then in most nearby possible worlds where p is true, S believes p (Nozick 1981: 186–7). However, adherence does not solve the problem. Apart from it being ad hoc to demand a special criterion for knowledge in the case of necessary truths, adherence is no requirement for knowledge in general because we can know something even when we use a method that makes us stay agnostic most of the time. For example, you may choose a very demanding method of belief formation, such as writing a philosophy article, and so come to believe only a fraction of the truths that you could endorse on less thorough reflection, and yet those that you come to believe should count as knowledge.[16] Hence, neither sensitivity nor its adherence variant suffice for a belief to count as knowledge. Let's turn to safety next.

[14] For externalists, defeat depends on gaining information about the reliability of one's belief-forming process. Hence, I ask what the agent *learns* about his or her beliefs. For internalists, the justificatory status of beliefs depends on the mental states of thinkers, so it is obvious that it matters what thinkers learn.

[15] Cf. Becker (2012); Bogardus (2014); Levy (2014: 26ff); Schafer (2014: 3892ff).

[16] See Setiya (2012).

### 7.4.2 Safety and Virtue Requirements on Knowledge

Safety is not sufficient for knowledge, either. Consider the following case by Schafer (2014: 384):[17]

> The Little Prince: The crown prince, Etienne—purely out of a deep sense of arrogance—believes that he is the strongest boy of his age in Paris. As a matter of fact, his belief is correct, but solely because his father has decreed that no stronger boy should be allowed to live in the city—a decree that the king's secret police are extremely efficient at carrying out.

In most nearby possible worlds in which the little prince believes that he is the strongest boy in Paris, his belief is true because the king's secret police is extremely efficient at making this the case. Moreover, his father is strongly disposed to issue the decree, so we may assume that he may not easily have failed to give the command. Hence, the little prince's belief is safe.

Nonetheless, argues Schafer, the little prince's belief is not knowledge because "while his belief is true (and safe), its truth (and safety) cannot be attributed to him in the sense that knowledge seems to require" (Schafer 2014: 385). This shows, argues Schafer, that "someone can have a safe belief about some matter, even though this belief is based on an epistemically horrible method, when the negative effects of this method are canceled out by good environmental epistemic luck" (Schafer 2014: 385). Schafer thus analyses the little prince case as involving a kind of knowledge-excluding *luck* or *accidentality*, and he relies on the wide consensus amongst epistemologists about the incompatibility of luck (or accidentality) and knowledge that I mentioned in the introduction.[18] Like early discussions of the Gettier problem in epistemology, which tried to identify warrant (in the technical sense as *that which must be added to true belief in order to secure knowledge)* by filling in the missing variable in 'JTB (justified true belief) + X' conditions of knowledge, the point of Schafer's case is that warrant will require more than JTB and safety.

[17] See Lackey (2008) for a related argument.
[18] For present purposes it is unproblematic to treat these terms synonymously.

The lack of warrant in the prince's case seems best explained by the lack of *credit* that the prince earned for his cognitive success.[19] Proponents of a credit requirement of knowledge suggest that in crediting an agent with knowledge, we are crediting his or her with having a relevant cognitive ability which played some key part in the production of the target true belief (Pritchard 2010: 135; Pritchard et al. 2012: 20).[20] According to this view, to say that someone knows is to say that believing the truth can be credited to her. It is to say that the person got things right due to her own abilities, efforts, and actions, rather than due to dumb luck, or blind chance, or something else. When one knows, then one's cognitive success should be creditable to one's cognitive ability (Greco 2010: 111; Pritchard 2012: 247–8). Of course, to assess adequately whether the prince lacks knowledge on a virtue-theoretic account, a more thorough assessment of 'cognitive abilities' and the possession conditions for such abilities would be required.[21] But to make the case for P1, it is enough to rest with the intuitive sense in which the prince's true belief cannot be credited to his ability.

Thus, what I want Schafer's case to illustrate is that there is sometimes a sense of accidentality (due to a lack of cognitive control exerted by the believer) in the truth of one's beliefs that is not exhausted by the simple safety principle that we have discussed thus far and therefore that safety is not sufficient for knowledge. This is important, because it makes the criticism of MODAL SECURITY apply wider than just to a criticism of modal analyses of knowledge applied to necessary truths.

---

[19] Though Schafer discusses his crown prince case as a case of knowledge preventing *accidentality*, I do not think this analysis works. First, there are reasons to doubt that we have a good account of luck or accidentality (Hales 2016; e.g. Lackey 2008; Morillo 1984). Second, many accounts of luck or accidentality are modal and they can be shown to collapse into safety and sensitivity, which makes them uninteresting for present purposes (Levy 2014; Pritchard 2005, 2014; Pritchard and Whittington 2015; Unger 1968; Yamada 2011). Third, there are cases that suggest that the best currently available accounts of luck do not exhaust the conditions for knowledge, so we would have to look for another factor to reach warrant anyway (Pritchard 2012). Finally, beliefs that are modally secure are, in one sense of the term, surely not accidental.

[20] I wish I could say more about the problem of characterising the kind of epistemic virtue (which is to be regarded as different than 'having a justified belief', 'having a reliable belief', etc.), which is in itself an interesting problem. Relevant to the thesis defended here is that if any conception of epistemic virtue is such that a lack of virtue entails a lack of safety, then the argument using cases that I have defended thus far does not work. I am not aware of such discussions, however (there are some who argue that virtue entails safety, which is different, of course).

[21] See Palermos (2011).

Four points about Schafer's case demand clarification. First, the case does not show why safety *and* sensitivity are not sufficient for knowledge (it might be that the conjunction is sufficient, even though the conjuncts are neither necessary nor sufficient). Second, why should we take the prince's beliefs to be justified in the first place? Third, well-known concerns about individuating methods make assessing the case difficult. Finally, it is unclear whether a better version of safety would not solve Schafer's problem with accidentality. In the next section, I dispel these worries.

### 7.4.3 Sensitivity and Safety

Let's begin by showing that the conjunction of sensitivity and safety is insufficient for knowledge too. Consider the following case:

> The Geeky Prince: Etienne's brother, Estephan the geeky prince, believes that all propositions expressed by the sentences in his green book are true – purely out of love for the colour green. As a matter of fact, all these beliefs are correct, but solely because his father has decreed that only true mathematical statements should be written in the green book – a decree that the king's court mathematicians are extremely efficient at carrying out.

Since the geeky prince's beliefs are about necessary truths, rather than about contingent truths, as in Schafer's case, the geeky prince's beliefs are clearly sensitive, at least on a standard semantic of the sensitivity conditional (and I mentioned above why adopting a standard semantic is prudent). The same holds for safety. We can assume that the king's court mathematicians are as efficient and vigilant as the king's secret police and so they will never make a mistake in filling the green book with true sentences. Moreover, not only is the king disposed to cater to the interests of his geeky son, and so will always decree that there be only true propositions in the book that is to his son's liking, he has also devised a sure-fire way to predict which book his son will like. If his son goes on to believe only the propositions in the red book instead, the king will adjust his decree accordingly. The geeky prince's method is therefore extremely reliable and there are no relevantly nearby possible worlds where he forms false mathematical beliefs. Hence, the geeky prince's beliefs are safe.

Nonetheless, the geeky prince's cognitive success is as accidental as that of the crown prince and his epistemically horrible method is cancelled out by environmental luck. Therefore the geeky prince does not have knowledge of any of the propositions written in his green book, although his beliefs are both sensitive and safe. This shows that the conjunction of sensitivity and safety is insufficient for knowledge.[22] This view can corroborated by showing that the geeky prince's beliefs are indeed safe and that the intuition that there is something amiss (epistemically) in the prince's case has to do with the insufficiency of safety for knowledge. To corroborate my claim, it is handy to return to the three problems I mentioned in the previous section: first, whether the prince's beliefs are justified to begin with; second, whether there are problems with individuating methods; and third, whether a better version of safety could indicate that the prince's beliefs are *not* safe.

First, one might worry that the geeky prince's beliefs are not justified because relying on one's colour preferences does not seem like a method that leads to justified beliefs about mathematics. Hence, any 'discomfort' that arises about the epistemic properties of this case might be due to a lack of justification, rather than due to, as Schafer and I argue, the accidental correctness of the little prince's belief.

The concern about justification can be dispelled on both externalist and internalist accounts of epistemic justification. An externalist notion will, at least on the interpretation favoured by Clarke-Doane, depend on a modal criterion such as safety or sensitivity.[23] In both cases, the little prince's belief is indeed justified,

[22] My proposed interpretation of the two prince cases depends on how one should understand the closeness of possible worlds. One might argue that both princes have false beliefs in nearby possible worlds, depending on one's criterion of closeness. However, using Lewis's (1979) widely accepted guidelines for assessing the closeness of possible worlds suggests that the nearest possible worlds are those in which the physical laws are held fixed and in which there is a near-maximal match of the spatio-temporal region. In the cases of the princes, keeping fixed the dispositions of the little prince, the geeky prince, and their respective fathers will count as a match in spatio-temporal region. As to the recipe for creating relevant counterexamples, it can easily cook up counterexamples in nearby possible worlds because there are beliefs that we take to be necessarily true and that are also deeply engrained in our psychology. Naturally, there is some vagueness here, but I take it that this is a general problem with nearness relations, not with the cases discussed here.

[23] As opposed to a non-modal, probabilistic notion of reliability, which gives faulty results in even simple cases; see Grefte (2017).

and the worry disappears. On an internalist notion, all depends on whether we take the mental life of a geeky prince, son of an all-powerful king, to provide good reasons, from his point of view, to believe that taking a fancy to a certain colour is a good indicator of mathematical truth. Assume that the prince's method has never failed him: whenever he compared what he believed about mathematics due to his 'colour-matching' method to the reputable court mathematicians, he was right. Never did he believe that 'p must be true because it was written in the green book' and find someone disagree with p. On an evidentialist account of justification, for example, it seems plausible that the geeky prince is justified in believing the truth of the mathematical sentences. All available information points towards this being the case, and for evidentialists it does not matter whether that seeming is also a correct representation of reality (Feldman and Conee 1985: 15). Therefore, the geeky prince has a justified belief, but he does not know, at least if Schafer is right about knowledge.

Next, one might object that the prince's beliefs turn out to be unsafe (as desired) depending on how one individuates methods. For example, one could describe differently the method employed by the prince (e.g. 'relying on whatever one fancies') to arrive at the verdict that the prince's belief is not safe, which would mean that these cases cannot work as counterexamples to the claim that safety and sensitivity are sufficient for non-accidentality. That is just the objection, however, that my argument depends on the solution to the generality problem for reliabilism (Conee and Feldman 1998). All I require is that i*f* there is some case in which the method turns out to generate safe but accidental beliefs, then this is sufficient for supporting the case for P1. No matter whether we individuate methods externally (to wit, what the agent believes is the way she formed her belief is not necessarily the way she formed it) or internally (the denial of external individuation), there are at least some cases in which a belief is modally secure and yet the truth of the belief is accidental.[24]

Finally, one might object that a better account of safety gets the right result in the geeky prince's case. There are systematic problems with safety analyses of

[24] See Pritchard (2005: 152) for a case of external method individuation; Baker-Hytch and Benton (2015: 48ff) and Setiya (2012: 93ff) defend internal method individuation, and both accounts are compatible with my argument.

knowledge, however, that should make us sceptical about whether safety will give the right result in cases like that of the geeky prince. Consider the distinction between *safety of beliefs* and *safety of methods*, which is key and yet often overlooked in the metaethical literature. To elaborate, consider how a defender of what might be called the METHODS APPROACH to safety might respond to the geeky prince case:

> Let p be a mathematical truth in the geeky prince's green book, and therefore a necessary truth. Of course, there is no possible case in which the prince falsely believes p. Yet, one can still doubt p by doubting the reliability of the prince's *method* that led him to believe p (rather than the reliability of his *belief*). For example, perhaps the prince could easily have chosen another method and decided to believe all propositions in the red book, full of falsities? If the prince believes p because it is written in his green book, he does not know p, although he could not have believed p falsely. His belief fails to count as knowledge because the method by which he reached it *could just as easily have led to a false belief in a different proposition*. (based on Williamson 2000: 181–2)

The METHODS APPROACH focuses on the safety of the relevant belief-forming method, M, that produced the belief whose safety is in question in the actual world, and requires that in nearly all possible worlds, beliefs produced by the method M are true (e.g. Pritchard 2005: 146, 2009; Weatherson 2004).[25] Although there are differences between individual accounts, the shared basic idea of the METHODS APPROACH is to check whether the belief about the "target proposition" continues to be true in nearby possible worlds (Pritchard 2005: 239). Thus, according to the METHODS APPROACH, the relative frequency of nearby worlds in which falsehoods are believed is the correct approach to showing why some beliefs about necessary truths are accidental and thus fail to qualify as knowledge.

However, the METHODS APPROACH does not rescue a purely modal approach to knowledge. Some methods produce true beliefs in counterfactually robust ways and yet the truth of the beliefs they produce seems uncomfortably lucky, as

[25] Clarke-Doane explicitly associates his view with that of Pritchard, whom we can count as a defender of the METHODS APPROACH (cf. Clarke-Doane 2016: 28).

demonstrated by the geeky prince, whose method is bound to lead to true beliefs given his father's strong disposition to make it the case and his fantastical ability to do so. Thus, proponents of safety cannot correctly analyse the geeky prince's case by emphasising methods rather than beliefs.

Even if doubts remain about the safety of the geeky prince's beliefs, more general considerations indicate that a purely modal safety condition for knowledge does not seem to capture all there is to know. To illustrate, consider a case in which the METHODS APPROACH seems to get it right:

> 6-sided die: Alf wonders whether p, where p is a proposition whose truth value is necessary. To decide, Alf casts a six-sided die, using the method <believe p when the dice lands 1>.

The METHODS APPROACH correctly implies that Alf does not know and the explanation is that there are five nearby possible worlds in which Alf uses that method and forms a false belief (because the die lands on any other number than 1, in which case he does not believe that p is the case). However, increasing the number of nearby worlds in which the belief is false does not seem to make up for *all* the relevant epistemic difference:

> 10-sided die: Bert wonders whether p, where p is a proposition whose truth value is necessary. To decide, Bert casts a ten-sided die, a pentagonal trapezohedron, using the method <believe p when the die lands 1>.

Again, the METHODS APPROACH correctly implies that Bert does not know whether p and the explanation is that there are even *more* nearby possible worlds in which Bert ends up with a false belief compared to Alf. However, there is something amiss, epistemically, even apart from the modal nearness of Bert's failure. There is an epistemic badness to Bert's method that is not reducible to the number of nearby worlds in which the method leads to false beliefs. To see this, consider the following:

> Loaded 6-sided die: Carla wonders whether p, where p *is necessarily true*. To decide, Carla casts a loaded die that has 1 on each side using the method <believe p when the die lands 1>.

Carla cannot form a false belief using her method. The METHODS approach implies that she knows. But that is mistaken. Carla's method is as bad as the method used by Alf and Bert. The purely modal METHODS APPROACH will explain the badness of Carla's method in terms of nearness of error, focusing on whether there are nearby worlds where Carla considers a falsehood using her method. But tweaking the case so that Carla automatically and highly reliably, though without being aware of it, uses her method and that loaded die *only* if she considers a necessary truth makes her belief seem as epistemically bad, but the METHODS APPROACH cannot tell us why.

The die cases are meant to illustrate that there is a remnant of epistemic concern even in cases where there is no 'modal elbow room', that is, no nearby possible world in the method used leads to a false belief. Whenever a method leads to true beliefs that are stable – in the sense that the belief's content could not have easily been different – Williamson, Pritchard, Clarke-Doane and other proponents of pure modal conditions for knowledge have to conclude that the belief is in fine epistemic standing as far as safety is concerned.[26] What this goes to show is that epistemic accidents that prevent a belief from qualifying as knowledge can happen even if there is no 'modal elbow room'. These cases are not meant to contain a counterexample to MODAL SECURITY. The point was to illustrate *how* modally secure beliefs may fail to qualify as knowledge. And it should warm us up to the idea that there is more to knowledge than 'JBT + sensitivity and safety'. If that is so, then we should expect that *learning* about the lack of warrant for one's beliefs can show that we don't have knowledge but modally secure beliefs nonetheless, which I defend in the next section.

## 7.5 Lack of Knowledge Undercuts

In this section, I argue that thinkers can *learn* that their belief fails to qualify as knowledge without giving them reason to doubt their belief's safety and sensitivity. This will vindicate premise P1, which was as follows:

[26] Compare variants of Greco's (1999) demon case, where the truth-conduciveness of the belief-forming method is upheld by a demon, which are counterexamples to the METHODS APPROACH analysis that Clarke-Doane adopts.

**P1**: It is possible that S gains new information, E, that shows that no D-belief qualifies as knowledge without giving S a reason to doubt that D-beliefs are modally secure.

Consider the geeky prince's cases again. The king's role in the thought experiment is to ensure that the beliefs of his son are 'bound to be true' and by that I simply mean that there is no nearby scenario in which the geeky prince forms a false belief (about the sentences in his maths book). Clearly, the explanation of the reliability of the prince's beliefs is that the king ensures that the prince's beliefs are reliable (in the sense of being modally secure). Now, the fact that the king ensured that his son is guaranteed to have true, reliable beliefs without the prince's own doing is what explains why the prince does not have knowledge. The truth of his beliefs has nothing to do with his abilities and everything with a fortuitous circumstance of his princely life. This is what I identified as a lack of cognitive control above.

Here's the rub: the very conditions that make the prince's beliefs reliable (the king's interventions) are responsible for the intuition that the prince's beliefs do not qualify as knowledge *because* they make it the case that the prince is not creditable for his cognitive success:

- On the one hand, the prince's beliefs are sensitive because they are necessarily true and safe because the king makes them so.
- On the other hand, the prince's beliefs do not count as knowledge because the king only, rather than the geeky prince, is creditable for the prince's cognitive success.

Thus, the king makes it the case both that the prince's beliefs are modally secure *and* that they fail to be knowledge. When the geeky prince *learns* about his fortuitous cognitive circumstances he will learn that his beliefs about the sentences written in the green book are all but guaranteed to be true and he would have no reason to doubt that his beliefs are modally secure. At the same time, he would learn that his beliefs about the sentences written in his green book fail to qualify

as knowledge.[27] Now, the geeky prince's case is surely fantastical, but a roughly analogous point holds in the case of our moral beliefs, objectively construed.

In the moral case, two conditions are 'king': they make our beliefs reliable and they make it the case that they do not, all else being equal, qualify as knowledge:

FIXED TRUTH VALUE: True D-beliefs are true in all possible worlds.

FIXED CONTENT: S holds true D-beliefs in all nearby possible worlds because of factor F, where F is not creditable to S's cognitive agency.

Beliefs in domains that satisfy FIXED TRUTH VALUE and FIXED CONTENT do not qualify, all else being equal, as knowledge because there is no information that those beliefs are virtuously formed. Both factors make for a world in which whatever the beliefs the thinker forms, they are bound to be true. To see this point, we have to imagine a thinker who knows nothing about his or her moral beliefs but their contents – he or she is oblivious to the way they were formed, unaware of the logical relations amongst the contents of his beliefs, and of whether other people hold similar beliefs and so on.

Of course, our beliefs *might* qualify as knowledge even though they satisfy FIXED TRUTH VALUE and FIXED CONTENT because neither condition rules out that our beliefs are formed virtuously. Surely most people also know more about their moral beliefs, but these complications can be reintroduced when we want to know whether information can undercut a belief without giving reason to doubt its modal security.

So, suppose that you consider the fundamental moral beliefs and you learn that they satisfy FIXED TRUTH VALUE and FIXED CONTENT by metaethical reflection. If *all you know* about the fundamental moral beliefs is that they satisfy FIXED TRUTH VALUE and FIXED CONTENT because of some factor F then you know that they will be modally secure (because both factors jointly entail sensitivity and safety). Indeed, satisfaction of both conditions is just what Clarke-Doane claims about the fundamental moral beliefs (Clarke-Doane 2015: 95): their truth values are fixed and their content is stable across the relevant possible worlds (given evolutionary explanations of morality). More generally, all beliefs whose content is

[27] Should the prince give up his beliefs? I address this in the next section.

necessarily true fit FIXED TRUTH VALUE. All beliefs that are sufficiently basic such that we can expect organisms in relevantly similar possible worlds to hold them satisfy FIXED CONTENT. If you learn that D-beliefs satisfy FIXED TRUTH VALUE and FIXED CONTENT, you will know that D-beliefs will be modally secure. If you are granted the assumption that some of your D-beliefs are true, you will know that some of your D-beliefs are modally secure. You will have no reason to doubt that your D-beliefs are modally secure.[28] At the same time, if that is all that you know about your D-beliefs, then learning that the D-beliefs satisfy FIXED TRUTH VALUE and FIXED CONTENT because of F implies that your D-beliefs are not knowledge.

Therefore, information might imply that a type of belief fails to qualify as knowledge *precisely* in cases where there are no reasons to suspect that the belief in question is not both safe and sensitive, and this type of lack of warrant can hold for all members of a type of belief. This is enough to vindicate premise P1.

## 7.6 Believe That p Only if You Know That p

Thus far, I have shown that it is possible to learn that some types of belief fail to be knowledge without having reason to doubt that beliefs of that type are modally secure. The next step is to show that, sometimes, we ought to give up our beliefs when we learn this. This will vindicate P2, which said the following:

> **P2**: If S gains new information, E, which shows that no D-belief qualifies as knowledge and some D-beliefs ought to be held only if they qualify as knowledge, then all D-beliefs that ought to be held only if they qualify as knowledge are undercut even though they might be modally secure.

According to P2, all tokens of a type of belief are undercut if we learn that beliefs of that type do not qualify as knowledge while also being committed to the

[28] We should not understand 'giving reasons' in the sense in which I can give you reasons for thinking that it is raining presently by pointing out that the street is wet. We might both see that it is not presently raining, and still, the street's being wet is some reason to think that it is raining. The counterfactual sense of giving reasons is not applicable in the case of modal security, where one needs to be convinced that at least one of the constituents of the modal security of one's beliefs is in jeopardy. Once we have established safety and sensitivity, then giving reasons to the contrary, as in 'suggesting that it is otherwise', is not relevant any more.

view that some members of that type of belief 'aim at knowledge' in the sense that one ought to believe that p only if one knows that p. There are two ways to interpret the claim that some D-beliefs ought to be held only if they qualify as knowledge. Narrowly, by appealing to moral objectivism's commitment to the actuality of moral knowledge. Or more broadly, by appealing to the claim that the norm of belief is knowledge (to wit, you ought to believe that p only if you know that p).

Consider the narrow interpretation first. Suppose we learn that all moral beliefs, objectively construed, satisfy FIXED TRUTH VALUE and FIXED CONTENT. If that is all we know about objectivist moral beliefs, then, all else being equal, those moral beliefs lack warrant and therefore do not qualify as knowledge. But moral objectivists are committed, not logically but as a matter of fact, to the view that we can in principle attain moral knowledge (e.g. Enoch 2011b: ch. 3; Parfit 2011a; Scanlon 2014: ch. 4; Shafer-Landau 2003: ch. 12). This commitment underlies their claims that most disagreements in ethics can be resolved, which requires an imbalance in knowledge between both disputants and not just an imbalance of justification (Enoch 2011b: ch. 8). Moreover, moral objectivists themselves argue that defending moral objectivism without the possibility of moral knowledge "has just about zero appeal" (Shafer-Landau 2012: 1) and should be avoided if at all possible (Enoch 2011b: 189). Importantly, the *possibility* of knowledge is not at issue, but not its *actuality*. For example, Sherlock Holmes might have good evidence, and thus justifiably believe, that Moriarty did it without knowing that Moriarty did it, and pointing out that Holmes does not know should not undercut his justification. Though Holmes does not have knowledge, it is possible that he could attain it. In the moral case, we might learn that we cannot in principle attain moral knowledge and then objectivists should accept that learning this compels us to withhold moral judgement. What the above has shown is that the reason why our moral beliefs may in principle fail to qualify as knowledge might precisely be because of those conditions that ensure that they are modally secure. Of course, we might learn more about our fundamental moral beliefs other than that they satisfy FIXED TRUTH VALUE and FIXED CONTENT, but unless we do, objectivists are committed to the view that we should withhold moral belief.

Clarke-Doane foresees such an objection and suggests that objectivists better avoid commitment to the possibility of knowledge than to hold on to the thesis that

the possibility of knowledge is a central motivation of the objectivist view (Clarke-Doane 2017c). That is, we might learn that our beliefs do not qualify as knowledge but the only consequence to draw is that we ought to give up the belief "my belief that p is knowledge" but *not* the belief that p (Clarke-Doane 2017c: 36). Others have suggested, similarly, that we can live with the concession that we don't know some of the things we believe to be true provided that we can still maintain that we are justified in believing these things (Wright 1991: 88). Thus, could objectivists escape the objection by weakening their commitment to the possibility of moral knowledge?

However, they cannot escape this objection because P2 gains support from more general epistemic principles too.[29] The most relevant principle is most closely associated with Williamson, who thinks that it is plausible that knowledge is the norm of belief (Williamson 2000: 249ff). On such a view, if we learn that our belief that p does not qualify as knowledge, this gives us reason to no longer believe that p, even though we are *not* given reason to think that p is false. Whether knowledge is the norm of belief is far from uncontroversial, but many epistemologists are inclined to accept the view.[30] If the norm of belief is knowledge, then learning that some type of belief fails to qualify as knowledge gives one reason to abandon those beliefs, precisely because one should only hold on to those beliefs insofar as they qualify as knowledge. This goes to show that P2 receives support from both the actual commitments of moral objectivists and, if these commitments are dropped, the general epistemological considerations about the norms of belief. Whether the norm of belief is knowledge, rather than truth, is of course an open question. But given its widespread support, it can be marshalled to support the ANTI-MODAL SECURITY argument.

Therefore, the idea that MODAL SECURITY is a general constraint on undercutters is false. Put in terms of undercutting information, the argument's conclusion can be restated as follows: you can learn that your belief B fails to qualify as knowledge *precisely because* B is modally secure and that undercuts B.

---

[29] One recent example is Gardiner (2017), who argues, congenially to my position, that a pure modal account of knowledge does not adequately explain the value of knowledge.

[30] See Baker-Hytch and Benton (2015); Hirvelä (2017); Huemer (2007); Jackson (2012); Littlejohn (2013); Sosa (2011).

Insofar as the norm of belief is knowledge you ought to give up your belief. There is thus a possible undercutter that does not conform to the constraints of modal security and so MODAL SECURITY as a general constraint on undercutters fails. Moral objectivism, therefore, is not off the hook when it comes to challenges that might show how all moral beliefs are undercut.

## 7.7 The Etiquette Conception of Undercutting Defeat

Moral objectivism still faces the possibility of undercutting challenges. To keep this possibility alive, we had to reject MODAL SECURITY. Rejecting MODAL SECURITY commits us to the etiquette conception of undercutting defeat. As pointed out in the main introduction, the most significant feature of the etiquette conception of undercutting defeat is that it goes beyond mere concern with forming failsafe beliefs and says that the permissibility of holding a belief can be affected by information that does not put into doubt the truth of the belief in question. Though I cannot provide a detailed defence of the etiquette conception in this chapter, I will briefly state its most fundamental assumption, defend my choice of terminology, and discuss some reasons for accepting the etiquette conception by showing how it deals with paradigmatic cases of undercutting defeat and by contrasting it with the orthodox conception of undercutting defeat.

The most fundamental assumption required to make the etiquette conception work is that the norm of belief is knowledge and not mere (non-accidental) truth. This would explain why we have epistemic reason to give up beliefs whose truth is 'safe'. The discussion in sections 7.4 and 7.5, however, lends some support to the view that the norm of belief is knowledge and not mere non-accidental truth because mere modal security of beliefs can be shown to leave open epistemic woes.

Calling the view an 'etiquette' conception is meant to bring out that the view prescribes a behavioural norm that is respected in some but not all circles and is rejected in other circles. Those who regard (non-accidental) truth to be the paramount norm of belief in epistemology will view the prescription to give up one's belief B upon learning that B does not qualify as knowledge as excessive in a similar way to how someone might consider certain table manners as excessive. This can be illustrated by two ways of looking at the norms of belief that are roughly analogous to two ways of looking at etiquette.

The first view is dismissive of etiquette and sees it as an inessential distraction from more important concerns. For example, you might consider table manners to be a mere distraction if your main concern is to satisfy your hunger. On this view, etiquette is an *add-on* to more important concerns. In a rough analogy, many philosophers conceive of truth as the fundamental norm of belief (that is, you ought to believe p only if p is true) and a concern with truth plays a pivotal role in epistemology, philosophy in general, and science (cf. Chan 2014; Duhem et al. 1906 [1991]; Lynch 2009). Such a concern with truth also explains why modal criteria such as sensitivity and safety are often considered of epistemic importance: satisfying those criteria minimises error. If (non-accidental) truth is the main goal for forming beliefs, we should expect conditions such as safety and sensitivity to be important. Everything else, arriving at the truth rationally, with justification and so on, is a nice addition, but not essential when we are in pursuit of our epistemic goals.

On an alternative view on etiquette, etiquette is of central importance rather than a mere distraction. Going back to the example about table manners, it may not be your goal merely to still your hunger but to do so graciously, in proper fashion. Similarly, you may believe that how you talk, dress, or travel into town may be futile if not done properly.[31] On such a view, 'etiquette' is essential to doing things and not a mere distraction. Again in a rough analogy, merely believing the truth in a way that is error prone is not enough for one's belief to be in good epistemic standing. More than modally secure belief is required to be permitted to hold a belief. We have to accept this view if we want to reject MODAL SECURITY. Though it might seem odd to give up one's beliefs because they do not qualify as knowledge, even though one has no reason to doubt their truth, the plausibility of the view just depends on the norms that govern belief.

Let's consider the details of the etiquette conception of undercutting defeat. The etiquette conception of undercutting defeat contrasts with the orthodox conception because it postulates a relation between knowledge and justification that runs counter to the usual conception of the relation between knowledge and justification. According to the etiquette conception, information relevant for

[31] For a view of how a form of etiquette can be of central importance, see Appiah (2011).

whether or not your beliefs qualify as knowledge affects whether or not your beliefs are justified, rather than the other way around:

> **The Etiquette Conception of Defeat**: Information, E, undercuts all our beliefs of a kind D based on reasons R, *if*: E implies that basing D-beliefs on R does not qualify D-beliefs as knowledge.

Thus, the etiquette conception specifies a sufficient condition for undercutting that applies to the support that a (type of) belief gains from its base. As such, the etiquette conception species defeat as it applies to doxastic justification, whereas propositional justification is not per se affected by the view.[32] This is as it should be: the orthodox view of undercutting defeat is properly concerned with doxastic justification, too, as we have seen in chapter 2. The 'basing' relation denotes a relation that obtains between a creature's belief, on the one hand, and the reasons for which she holds the belief (Neta 2011: 110). So, if R are the reasons *for which* we maintain D-beliefs, and D-beliefs are justified, R give us positive justification for maintaining D-beliefs. Information E, however, might show us that basing D-beliefs on R does not imply that our D-beliefs qualify as knowledge. This does not imply that D-beliefs are false, and so an etiquette defeater does not count as a rebutter (Pollock 1995: 85). Instead, the etiquette defeater is an undercutter insofar as it removes the support for maintaining D-beliefs without showing that D-beliefs are false (ibid.).

We can get an etiquette defeater in a number of ways. New information E may simply be a statement to that effect (e.g. 'you do not know whether p') or E might be information that implies it (e.g. 'your belief that p satisfies FIXED TRUTH VALUE and FIXED CONTENT' or, more colloquially, 'no matter what you do, you would believe that p truly'). As in the orthodox conception of defeat, *whether* E defeats our D-beliefs depends on the credence that we ought to bestow on D-beliefs and E, respectively, and whether there are defeaters for E.

The etiquette conception is more encompassing than the orthodox conception of defeat. The orthodox conception maintains that new information undercuts the support conferred on a belief by its base by implying that *the content of the belief*

[32] Cf. Turri (2010) and Melis (2017) for a discussion of the relation between doxastic and propositional justification.

*might misrepresent the fact*s. The etiquette conception, in contrast, maintains that new information undercuts the support conferred on a belief by its base by implying that *the way the thinker formed her belief might not qualify the belief as knowledge.*

The etiquette conception identifies paradigm cases of undercutting defeat quite well. Recall our earlier example from the main introduction. Suppose you are on a factory visit looking at what appear to be red wedges on a conveyor belt. The wedges *seeming* red to you, and your background assumption that perceptual conditions are normal, justify your belief that the wedges *are* red. When the foreman tells you that the wedges are illuminated by a red light for technical reasons, you learn that basing you beliefs about the wedge's colour on your *seemings* does not imply that you *know* that the wedges are red. Of course, this case can be explained by pointing out that your seeming is no reliable guide to reality in this case and thus the content of your belief might misrepresent the facts, and therefore the orthodox conception also does well in paradigm cases of defeat.

The etiquette conception is, moreover, capable of dealing with undercutting defeat in cases where the orthodox conception fails. If D-beliefs satisfy FIXED TRUTH VALUE and FIXED CONTENT, there is no way in which new information could call into question the justification of all D-beliefs without showing that they are false. Adherents of the orthodox conception would therefore have to say that there is no undercutting defeat in domains such as morality (objectively construed). It is possible, however, that beliefs that satisfy FIXED TRUTH VALUE and FIXED CONTENT do not qualify as knowledge and the etiquette conception explains why even such 'failsafe' beliefs ought to be given up.

The radical contrast to the orthodox conception of defeat becomes clear by seeing that the orthodox conception of undercutting defeat conceptualises undercutting defeat solely in terms of the relation of undercutting information and the truth of the target belief (Pollock 1970, 1995; Pollock and Gillies 2000). The orthodox conception is dominant in the current discussion of the reliability challenge and related genealogical debunking arguments. Virtually all discussions of the 'unsettling feeling' that sometimes arises when we discover the mechanisms that produced our beliefs diagnose that feeling as a case of what might be called *alethic anxiety*, a worry about the truth of one's beliefs. For example, Silva (2016)

argues that historical variability is evidence that the factors that influenced one's belief that p are "disconnected" from the truth about whether or not p (Silva 2016: 3). Similarly, Ballantyne (2013) argues that discovering the historical variability of your beliefs makes it the case that "you lack reason to think that … you now believe truly and not falsely whether p" (Ballantyne 2013: 252; DiPaolo and Simpson 2016). Braddock (2016) interprets causal debunking arguments of morality as indicating that it is false that moral judgements are "likely to be true" (Braddock 2016: 845), as do Leben (2013) and Lutz (forthcoming).

We should now be in a position to see that the focus on the orthodox conception of defeat is misguided in discussions of the reliability challenge and related debunking arguments. *Alethic anxiety* is unwarranted when we consider the fundamental moral beliefs: they are bound to be true. However, *alethic anxiety does not* exhaust the epistemic phenomena that we can be anxious about, as shown by the discussion above. Moreover, alethic anxiety *should* not exhaust the epistemic phenomena that we can be anxious about. If alethic anxiety is the sole epistemic malaise that we ought to have, many genealogical debunking arguments aimed at beliefs that satisfy FIXED CONTENT and FIXED TRUTH VALUE would be inefficacious.

Hence, there are some reasons for accepting the etiquette conception of undercutting defeat. I have already mentioned one; if we reject it, there could be no undercutting defeat in domains that satisfy FIXED TRUTH VALUE and FIXED CONTENT. It is implausible that proponents of objectivism in those domains have an easy time answering the reliability challenge. This should not be taken as a very strong reason, however, because it is motivated mainly by the presupposition that there must be something wrong with moral objectivism and related views. Independent motivation for the etiquette conception is a bit harder to find. One avenue to pursue to find independent motivation for the etiquette conception is that there seem to be *contexts* in which knowledge is the norm of belief and where learning that one's belief is not knowledge epistemically compels one to give up one's belief. One such case might be the philosophical seminar room. Suppose someone is defending the sound argument of an authoritative philosopher that he or she has learned by heart, without fully grasping it. He or she might insist on the truth of the argument's conclusion, but he or she should, epistemically, give it

up.[33] In a similar way, when we are challenged to defend the reliability of the fundamental moral beliefs and we learn that the conditions in which we formed them guarantees them to be true but without qualifying them to be knowledge, we might have a good epistemic reason to give them up.

Moreover, it is doubtful that moral objectivism is per se excluded from the threat of undercutting defeat, and the etiquette conception offers an explanation of why this is so. Since the phenomenon of undercutting defeat is a stalwart of fallibilism, and fallibilism is the epistemology of choice for virtually all epistemologists today, it would be a cost to moral objectivism if it implied that fundamental moral beliefs cannot be undercut. So, moral objectivists have good reason to accept the etiquette conception of defeat too.

Of course, these considerations do not constitute a full defence of the etiquette conception. For one, it appears odd to give up one's beliefs because they do not qualify as *knowledge* even if one has no reason to doubt that they are bound to be true. In response, I can only repeat a suggestion made above: this is just like a question of etiquette insofar as it is an open question what the norm of belief turns out to be (just as it is an open question what the norm of eating turns out to be: satisfying one's hunger or doing so 'in style'?).

To convince adherents of the orthodox conception of defeat, a defence of the etiquette conception of defeat must take into account more points than I can consider here. What is clear, however, is that those who want to raise an undercutting challenge to moral objectivism have to get busy in providing said defence.

## 7.8 Concluding Remarks

This chapter showed, first, that MODAL SECURITY as a general constraint on undercutters is false: learning that a belief is invariably true because of its content and the way the environment aligns with it may undercut that belief despite the fact that the belief is modally secure. The second finding was that rejecting MODAL SECURITY based on this argument commits us to the etiquette conception of undercutting defeat. Though this does not rescue the reliability challenge to moral

[33] Focusing on context might also help to avoid the swamping problem that is relevant in this context (cf. Kvanvig 1998; Pritchard et al. 2012; Zagzebski 2002).

objectivism (insofar as reliability is understood as 'modally secure'), it captures what I take to be the spirit of the challenge: there is something wrong, epistemically, with moral objectivism (and related views about logic, mathematics, or modality) and this can be brought out by raising a defeat challenge for the view. Since rejecting MODAL SECURITY is required to reinstate the possibility of such a challenge, it makes good sense to accept the etiquette conception.

For the debate about undercutting defeat challenges in metaethics, this means that moral objectivists cannot avert those challenges simply by pointing to the modal security of our moral beliefs. However, to put pressure on objectivists, proponents of such challenges must now show what it is that reveals that our moral beliefs violate a virtue-epistemological constraint on knowledge. Showing this, and defending the etiquette conception of defeat, must be left to another project.

This page intentionally contains only this sentence.

# Conclusion

We have seen that we can't escape our evolutionary past and its influence on our moral cognition. Whether we, therefore, have good epistemic reason to give up our moral judgements depends on a deeper epistemological question about the nature of undercutting defeat. Ought we (epistemically) to give up beliefs that are justified and true when we learn that they do not qualify as knowledge? Answering 'yes' commits us to what I called an etiquette conception of undercutting defeat. According to the etiquette conception, learning that a belief fails to qualify as knowledge for a reason beyond your control undercuts your doxastic justification for holding the belief, even though you might have no reason to doubt that the belief is justified and true. Thus, my answer to the main question of my thesis is that evolutionary explanations of morality can undercut our moral beliefs only if two conditions are met.

The first condition is that the etiquette conception of defeat must be true. I supported this point in two ways. On the one hand, I argued that there are good reasons for rejecting the disagreement view (according to which evolutionary explanations of morality are undercutting in virtue of the epistemic significance of disagreement), which seemed to be the most promising interpretation of the evolutionary defeat challenge. On the other hand, I argued that to make the evolutionary defeat challenge stick, we cannot appeal to explanatory, modal, probabilistic, or causal conceptions of defeat since they either beg the question against moral objectivism or do not result in a defeater for moral beliefs. I argued that the disagreement view should be rejected and that orthodox conceptions of defeat were unavailable, either because they cannot deal with undercutting defeat of necessary truth or because they would imply that objectivist moral beliefs are not justified to begin with. Therefore, I was led to espouse the etiquette conception of defeat as a necessary condition for the success of the evolutionary defeat challenge. I am not sure that I have reviewed every possible account and so there might be others that could get the evolutionary defeat challenge off the ground. I

am sure, however, that I have considered the best accounts currently offered in the metaethical literature.

The second requirement for the evolutionary defeat challenge to succeed is that evolutionary explanations of morality must show us that our beliefs are true for reasons beyond our own cognitive ability. I supported this view by arguing that objectivist moral beliefs are not epistemically lacking in modal terms, nor in explanatory terms, but that the best explanation for why it still seems suspicious to rely on them would be that they fail an ability or control condition on knowledge. This condition must be shown to be violated by evolutionary explanations of morality for the evolutionary defeat challenge to succeed. Whether the best available evolutionary explanations of morality do so is an open question, and I have already suggested, briefly, that the answer depends on how to best understand what our cognitive abilities are in an epistemically relevant sense. Put simply, the central finding of my thesis is that *evolutionary explanations of morality undercut all moral beliefs only if the etiquette conception of belief is true.*

The etiquette conception of defeat will be hard to accept for those who believe that the norm of belief is non-accidental truth (to wit, only believe that p if p is true), not knowledge (to wit, only believe that p if you know that p). They will resist the view that a belief that is justified, and all but guaranteed to be true, ought to be given up. Whether their resistance is rational will depend on wider epistemological considerations for and against knowledge as the norm of belief. What I have done in this thesis is to make the case for taking seriously the etiquette conception of undercutting defeat as the turning point of the success of the evolutionary defeat challenge. This was mainly done by considering the two strongest interpretations of the epistemic significance of evolutionary explanations of morality, the reliability view and the disagreement view. In addition, I clarified the nature of the challenge by assessing its reliance on a priori and a posteriori considerations (arguing that it depends on the former, not the latter) and defended the view that *if* moral beliefs are undercut, their justification cannot be reinstated, and by showing that the orthodox view of defeat cannot account for evolutionary defeat. Therefore, this is how to undercut all defeasibly and non-empirically justified moral beliefs:

(a) Vindicate the etiquette conception of defeat.

(b) Vindicate an ability or control requirement for knowledge.

(c) Demonstrate that the best evolutionary explanation of our moral beliefs implies that they are true for reasons that are insufficiently connected to our cognitive abilities or control (and that our moral beliefs thus fail to satisfy a necessary ability requirement for knowledge).

In the remainder of this conclusion, I provide a synthesis of the research results and describe the contribution of the thesis towards answering the central research questions. I will briefly recount the individual steps of my argument before discussing some open questions and avenues for future research.

## Review of the Thesis

**Chapter 1** prepared the ground for my thesis by elaborating on three themes: first, that there are solid empirical accounts of the evolution of morality; second, that metaethical evolutionary ethics must be distinguished from the notorious (and mistaken) prescriptive evolutionary ethics of the 20th century; and third, that existing discussions of the metaethical implications of evolutionary explanations of morality suggest that the survival of evolutionary defeat is in jeopardy. The problem is, I argued, that we need a valid epistemic principle and an indication based on evolutionary considerations that moral beliefs violate this principle to draw legitimate metaethical conclusions. A brief literature review then showed that the survival of defeat hinges on the disagreement view (which said that evolution defeats moral beliefs in virtue of the epistemic significance of disagreement) and the reliability view (which said that evolution defeats moral beliefs in virtue of the epistemic significance of explaining our moral reliability). The upshot of chapter 1 was that we need an answer to my main research question to evaluate the strength of the two prominent evolutionary debunking arguments of Sharon Street and Richard Joyce. In that way, chapter 1 provided motivation for the search for an answer to how the evolutionary defeat challenge can succeed.

I started out my search in **chapter 2** by considering a general question about epistemic defeat: what are the conditions for undercutting defeat? I argued that

there must be objective, perspective-independent criteria for establishing whether new information undercuts a belief. In support of this claim, I argued that it is untenable to maintain a pure subjectivist view of undercutting defeat, according to which S's belief that p is defeated if S takes her belief that p to be defeated. Such a view would make justification come too cheap, because one could *take* new evidence to be undercut and thus shield one's beliefs from defeat. Philosophers such as Bergmann (2006) and Plantinga (2002) try to avoid this conclusion by supplementing their subjectivist accounts of defeat with an idealised element that allows them to say that a thinker *should* take new information to be undercutting, even if the thinker does not actually do so. I argued that the subjectivist and idealist element in these accounts are motivated by incompatible explanations and that they yield conflicting verdicts about whether a belief is undercut or not. Thus, subjectivist idealist views ought to be rejected.

The findings of chapter 2 were important for my main research question for two main reasons. First, they supported the view that looking for an account of undercutting defeat is required to assess properly the evolutionary defeat challenge. Moreover, by showing that existing accounts of undercutting defeat do not yield the conclusion that defeasibly and non-experientially justified moral beliefs are undercut, chapter 2 lends plausibility to the view that the success of the evolutionary challenge will depend on the correct account of undercutting defeat.

In **chapter 3**, I turned from the conditions of defeat to a question about the kind of information that would be defeating in the case of the evolutionary defeat challenge. I started out by noting that the evolutionary defeat challenge is often considered as relying on information about the *causal* origins of our moral beliefs. Thus, the challenge seems to be based on empirical, a posteriori considerations about how we arrived at our moral beliefs. I resisted this common conception and argued that if evolutionary explanations of morality defeat moral beliefs, they do so on the basis of a priori considerations about, for example, the in-principle impossibility of explaining the reliability of moral beliefs. A posteriori information about the causes of our beliefs is insufficient to undercut them because we are considering non-empirically justified beliefs. Thus, I concluded that, strictly speaking, a posteriori considerations about the causal origins of our moral beliefs are inessential to defeating defeasibly and non-empirically justified moral beliefs.

Moreover, the chapter demonstrated why it is not possible to show that *all* moral beliefs are false by appealing to their causal background.

Thus, chapter 3 has shown that the success of the evolutionary defeat challenge, which deals in a posteriori information about the evolutionary origins of our moral beliefs, depends on whether there are good a priori grounds to give up non-empirically justified beliefs. Of course, this is not to say that evolutionary explanations of moral beliefs are irrelevant in an effort to undercut all moral judgements. Their relevance is restricted, however, to pointing towards an a priori problem with all moral beliefs, and the problem itself will have to be defended on a priori grounds.

Chapters 2 and 3 put the evolutionary defeat challenge into sharper relief: the grounds for giving up our moral beliefs will be a priori, and we must find an adequate, objective account of undercutting defeat. Chapter 4 then answered a question about the objectivist's options for responding to the challenge. Contrary to the position thus far discussed in my thesis, some philosophers have thought that moral judgements are defeated by the evolutionary challenge.

I started **chapter 4** by considering how defeat could be resisted and then picked up on a recent suggestion by Andrew Moon (2017) about how moral objectivists could respond to the threat of the defeat of all moral beliefs. Although defeaters can usually be defeated themselves, I pointed out that defeat of *all* moral beliefs would disable objectivists from reinstating the justification of moral beliefs. Moon suggests an alternative way to resist defeat: it might be possible that a defeater D for the support that evidence E provides for the belief B can be *deflected* by a defeater-deflector, DD. Hence, had S not known DD, D would have defeated the support that E provides for B. Moon suggests that the evolutionary defeat challenge might *raise* a defeater, but that the defeater could be *deflected*. Moon also suggests a number of conditions for when a subject S has a defeater-deflector, and I argued that none of them are satisfied in the case of moral beliefs. Thus, I concluded that *if* the evolutionary challenge raises a defeater, it could neither be defeated nor deflected by moral objectivists.

Chapter 4 thus made a jump in the dialectic by *assuming* that all moral beliefs could be undercut by evolutionary explanations of morality, whereas I had not yet established whether or how this could be the case. In doing so, chapter 4

contributed to a better understanding of the evolutionary defeat challenge by considering how devastating its effects would be, should it succeed. This was particularly important because I conceive of the challenge as a challenge of defeat. Insofar as defeaters can easily be defeated themselves, it would be premature to assume that showing that justification for objectivist moral beliefs is *lost* would mean that it cannot be regained. As I showed in chapter 4, however, defeat of all moral beliefs seems definite: justification cannot be regained. Thus, up until chapter 5, I left open the main question that I raised in chapter 2: *why* are evolutionary explanations of moral beliefs defeating?

Beginning in **chapter 5**, I set out to answer this question. I first discussed what seemed to be the most promising candidate for raising an undercutting defeater: the epistemic significance of disagreement. I say 'most promising' for two reasons. First, chapter 1 has shown that all other interpretations of the case for evolutionary defeat face serious difficulties, with only the disagreement view being free from counterarguments and objections. Second, several philosophers have already pointed out the possible relevance of considerations regarding disagreement for the debate about evolutionary defeat.

In chapter 5, after describing *how* disagreement is generally thought to be epistemically significant, I introduced the view that evolutionary explanations of morality might *show* that there is such disagreement about all moral beliefs. If that were correct, evolutionary explanations of morality would undercut all moral beliefs in virtue of the epistemic significance of disagreement. My strategy in chapter 5 was to argue that only disagreement between epistemic peers (in regard to the dispute at hand) is epistemically relevant and then to see whether evolutionary explanations of morality imply that there is such disagreement. I considered both a broad and a narrow view of peerhood and argued that the evolutionary hypothesis fails to imply relevant moral disagreement on either conception of peerhood. Peers on a narrow conception must have the same evidence and be equally good at processing the evidence. On evolutionary paths where we would have different moral beliefs, we would not have the same evidence insofar as we count moral intuitions as evidence. By portraying moral intuitions as a function of an organism's evolutionary path, proponents of the disagreement view are committed to the view that the evolutionary hypothesis, at the very least,

cannot imply narrow peer disagreement. Peers on a broad conception must be equally likely to have true beliefs (in their domain of peerhood). I granted that other organisms are considered peers until we have a reason *not* to consider them to be peers. Even so, I argued, if we learn that someone *completely* disagrees about all relevant moral beliefs, this gives us reason to consider him or her *less* likely than us to get moral matters right. Finally, if there is only some disagreement about moral beliefs, we might thus have reason to withhold judgement about *some* moral beliefs. The evolutionary hypothesis, however, does not give us reason to give up all moral beliefs. Put simply, any disagreement implied by evolutionary explanations of morality will either not encompass *all* moral beliefs or fail to be amongst peers.

With respect to my main question, chapter 5 implied that evolutionary explanations do not undercut our moral beliefs in virtue of the epistemic significance of disagreement. If the disagreement view were indeed the only hope of undercutting all objectivist moral beliefs, that project would fail. The thesis defended in chapter 5 might have implications beyond the evolutionary debunking debate. A core idea is that in domains where intuitions count as evidence, anyone with different intuitions will fail to be a peer (on a narrow conception of peerhood). Moreover, I hope to have shed more light on the extent to which evolutionary considerations imply something about hypothetical moral disagreement. Since moral judgements fulfil a function, they will be clustered, and it will be unlikely that any conceivable organism will seriously dispute all moral beliefs. Of course, the latter point is a substantial empirical question, and it would demand an empirically supported answer to fully settle this point.

In **chapter 6**, I considered another way in which disagreement may play a role in answering my main research question: the evolutionary hypothesis might raise a worry about the reliability of moral beliefs and moral objectivists might cope with this challenge by providing a third-factor explanation. I assumed in this chapter that a concern with the reliability of moral beliefs might be the problem with the evolutionary hypothesis. That is, any defeater generated by the evolutionary hypothesis will be due to the *inability* to explain the reliability of moral beliefs. On this assumption, third-factor explanations are crucial for the success of the evolutionary defeat challenge. If third-factor explanations are

legitimate and successful, the reliability of moral beliefs could be explained, and moral beliefs would not be undercut. I began the chapter by pointing out some controversies about third-factor explanations and then considered a recent proposal by Tersman (2017), who argued that third-factor explanations are constrained by the actual amount of radical moral disagreement. I argued that the constraints proposed by Tersman are satisfied given the assumption granted in the context of the reliability challenge. If we are allowed to assume the truth of some moral beliefs (to support a third-factor account), then there cannot be rational disagreement about these beliefs and disagreement about beliefs other than those is irrelevant for an assessment of third-factor accounts. Hence, third-factor accounts should be immune from defeating disagreement.

The upshot of chapter 6 for my main question was that it substantiated the view that disagreement does not play a role in showing how the evolutionary defeat challenge works. Moreover, in chapter 6 I engaged with the debate about the legitimacy of third-factor explanations and argued that moral disagreement does not curb third-factor accounts.

As illustrated in the literature review discussed in the introduction, the starting point of the thesis was that the evolutionary challenge succeeds only in virtue of its connection to the problem about disagreement. If I am right that evolutionary explanations of morality fail to raise a problem about disagreement, the prospects of the evolutionary defeat challenge depend on finding another objective criterion for undercutting defeat. I discussed such a criterion in the final chapter, chapter 7.

In **chapter 7**, I took on a challenge for all who maintain that the evolutionary hypothesis raises a problem for moral beliefs by either undercutting it directly or via the inability to explain their reliability. I showed that a number of philosophers, like Justin Clarke-Doane (2017a), have argued that there is no plausible interpretation of the demand to 'explain the reliability' of moral beliefs such that it seems in principle impossible to do and yet also undercutting. If that view were correct, then the evolutionary defeat challenge would falter. I argued that this view is false for reasons that have to with the proper analysis of knowledge. By considering a number of cases, I constructed the view that pure modal analyses of knowledge will fall short of giving us a full account of knowledge.

It followed from these cases that moral beliefs can be epistemically sensitive and safe but nonetheless fail to be knowledge. The second step in my argument involved showing that learning that a belief fails to be knowledge may provide us with reason to give up that belief, even though we are given no reason to doubt that the belief is justified and true. Finally, I argued that this is so based on the assumption that the norm of belief is knowledge, rather than (mere) truth.

I showed that rejecting the orthodox conception of defeat commits us to the etiquette conception of undercutting defeat. This chapter provided the answer to my main question by delineating a way in which the evolutionary challenge might work. To get the evolutionary defeat challenge off the ground, we need to adopt the etiquette conception of defeat.

Though I did not conclude *that* moral beliefs are undercut by learning about evolutionary explanations of morality, I argued how it might be possible. In pointing out the relevance of the etiquette conception of defeat, I have also established a result that should be of interest beyond the metaethical debate. We seem to form non-empirically justified beliefs about necessary truths whose content is not analytic in domains such as mathematics, logic, or philosophy and it is important to say how these beliefs could be undercut. The etiquette conception of defeat can help us understand *how* beliefs in these domains might be undercut, which is a desideratum of any fallibilist epistemology and a crucial precondition of making good sense of the metaethical implications of the many recent findings about the origins of our moral beliefs.

## Outlook

Future work should be directed along two main pathways: first, defending the etiquette conception of defeat. Second, inquiring whether evolutionary explanations of morality imply that moral beliefs are undercut according to the etiquette conception of defeat.

The most controversial assumption regarding the etiquette conception of defeat is that the norm of belief is knowledge rather than mere truth. Given the support for this view in current epistemology, I believe that the prospects are good for vindicating the etiquette conception of defeat. Moreover, the etiquette conception might itself provide support for the view that the norm of belief is

knowledge. This is because it is prima facie plausible that even non-empirically justified beliefs can be subject to undercutting defeat. On the assumption that non-empirically justified belief is possible in the first place, there should be some way in which such beliefs can be undercut and the etiquette conception explains how that would be the case. Reasoning along this line could provide independent support for the view that the norm of belief is knowledge. It seems fruitful to extend this line of research because many philosophers are concerned with elucidating the (predominantly sceptical) epistemic consequences of revealing the causal influences of our beliefs (e.g. Avnur and Scott-Kakures 2015; Hricko and Leben 2017; Sher 2001; Silins 2014), arguments that appear to share a "family resemblance" (Machuca 2018), and at the same time, the very nature of defeat, including its existence as an actual phenomenon, is questioned by others (e.g. Casullo 2008; Janvid 2008; Lasonen-Aarnio 2013, 2014; Neta 2009). There are very few connections between both debates and I believe that my thesis provides a useful starting point that shows how these debates could be fruitfully connected. It will also be interesting to establish links to projects that try to model moral reasoning in machines. John Pollock's OSCAR project (1995, 2001), for example, is such an attempt and it relies on the orthodox view of defeat. Insofar as non-empirically justified beliefs are an integral part of our moral reasoning, it would seem fruitful to inquire how an etiquette conception of defeat could help to model moral reasoning more appropriately.

I have not addressed the question *whether* our moral beliefs are undercut and I want to end on briefly suggesting a way forward in regards to this question: are our moral beliefs undercut by evolutionary explanations of morality if the etiquette conception of defeat is true? As pointed out above, we have to understand in greater detail what a control or ability requirement on knowledge entails and whether our moral beliefs might violate this requirement.

The control or ability requirement on knowledge can be understood as follows. Virtue epistemologists tend to think of the abilities that account for knowledge-acquisition as belief forming abilities and to think of successful belief as true belief irrespectively of how the belief is formed. We thus can have success in believing, true belief, without exercising the right kind of ability, that is the kind that accounts for knowledge when suitable conditions are satisfied. There can be a

manifestation of the right kind of ability in the formation of false belief and there can be a manifestation of the wrong kind of ability in the formation of true belief, as in the prince cases that I discussed in chapter 7. Hence, there can be manifestation of ability and success but not knowledge (cf. Millar 2009: 227ff). Virtue epistemologists insist, therefore, that for knowledge the true belief must be sufficiently due to the subject's manifesting the right kind of ability. John Greco (2009) sums up the view by saying that knowledge is achievement: success due to ability. Sosa (2007) says that knowledge is true belief due to competence.

Apart from the obvious fact that accounts of epistemic ability or control come with problems of their own, it will be crucial to assess whether all virtue theoretic accounts are created equal in regards to their ability to support an evolutionary defeat challenge. For example, Sosa (2007: 27) writes that a cognitive competence or ability is "a disposition resident in the agent, one that would in appropriately normal conditions ensure (or make highly likely) the success of any relevant performance issued by it". Sosa's account puts epistemic success *brought about by an agent's disposition* into focus. The question in vindicating an evolutionary undercutter based will therefore be whether the (ultimate) influence of evolutionary forces can be understood in such a way that our disposition to form moral judgments can be seen as harming our cognitive control or not. If it does, then evolution might undercut all our moral beliefs. Pursuing this question in appropriate detail has to be left to another project. The way ahead, however, is clear.

This thesis has shown, above all, that evolutionary defeat is a possibility by inquiring deeper into the nature of defeat and the prospects of information about our evolutionary origins to instantiate defeat. Evolutionary defeat has thus been taken off the list of endangered phenomena. If we want to find it in the wild, we now know where to look.

This page intentionally contains only this sentence.

## References


Abarbanell, L., and Hauser, M. D. (2010), 'Mayan morality: an exploration of permissible harms', *Cognition*, 115/2: 207–224.

Alexander, R. D. (1987), *The Biology of Moral Systems* (London: Routledge).

Allhoff, F. (2003), 'Evolutionary Ethics from Darwin to Moore', *History and philosophy of the life sciences*, 25/1: 51–79.

Alston, W. P. (1989), *Epistemic Justification*: *Essays in the Theory of Knowledge* (Ithaca, NY: Cornell University Press).

—— (2002), 'Plantinga, Naturalism, and Defeat', in J. K. Beilby (ed.), *Naturalism defeated? Essays on Plantinga's evolutionary argument against naturalism* (Ithaca, NY: Cornell University Press), 176–203.

Anonymous (1871), 'Darwin on the Descent of Man', *The Edinburgh Review*, 1871: 195–196.

Appiah, K. A. (2011), *The Honor Code*: *How Moral Revolutions Happen* (New York, NY: Norton & Company).

Artiga, M. (2015), 'Rescuing Tracking Theories of Morality', *Philosophical Studies*, 172/12: 3357–3374.

Audi, R. (1997), *Moral Knowledge and Ethical Character* (Oxford: Oxford University Press).

—— (2004), *The Good in the Right*: *A Theory of Intuition and Intrinsic Value* (Princeton, NJ: Princeton University Press).

—— (2013), *Moral Perception* (Princeton, NJ: Princeton University Press).

Avnur, Y., and Scott-Kakures, D. (2015), 'How Irrelevant Influences Bias Beliefs', *Philosophical Perspectives*, 29/1: 7–39.

Axelrod, R. (2006), *The Evolution of Cooperation* (New York, NY: Basic Books).

Axelrod, R., and Hamilton, W. D. (1981), 'The Evolution of Cooperation', *Science*, 211/4489: 1390–1396.

Ayer, A. J. (1971 [1936]), *Language, Truth, and Logic* (Harmondsworth: Penguin Books).

Baker-Hytch, M., and Benton, M. A. (2015), 'Defeatism Defeated', *Philosophical Perspectives*, 29/1: 40–66.

Ballantyne, N. (2013), 'The Problem of Historical Variability', in D. E. Machuca (ed.), *Disagreement and Skepticism* (New York, NY: Routledge), 239–58.

Ballantyne, N., and Thurow, J. C. (2013), 'Moral Intuitionism Defeated?', *American Philosophical Quarterly*, 50/4: 411–421.

Baras, D. (2017), 'Our Reliability is in Principle Explainable', *Episteme*, 14/2: 197–211.

Barkhausen, M. (2016), 'Reductionist Moral Realism and the Contingency of Moral Evolution', *Ethics*, 126/3: 662–689.

Baumann, P. (2008), 'Is Knowledge Safe?', *American Philosophical Quarterly*, 45/1: 19–30.

Baumard, N. (2016), *The Origins of Fairness*: *How Evolution Explains our Moral Nature* (Oxford: Oxford University Press).

Becker, K. (2012), *Epistemology Modalized* (New York, NY: Routledge).

Bedke, M. (2009), 'Intuitive Non-Naturalism Meets Cosmic Coincidence', *Pacific Philosophical Quarterly*, 90/2: 188–209.

—— (2014), 'No Coincidence?*', in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics*. *Volume 9* (Oxford: Oxford University Press), 102–25.

Beebe, J. R., and Sackris, D. (2016), 'Moral Objectivism Across the Lifespan', *Philosophical Psychology*, 29/6: 912–929.

Behrends, J. (2013), 'Meta-normative Realism, Evolution, and Our Reasons to Survive', *Pacific Philosophical Quarterly*, 94/4: 486–502.

Benacerraf, P. (1973), 'Mathematical Truth', *The Journal of Philosophy*, 70/19: 661–679.

Bengson, J. (2015), 'Grasping the Third Realm', in T. S. Gendler and J. P. Hawthorne (eds.), *Oxford Studies in Epistemology*, Volume 5 (Oxford: Oxford University Press), 1–38.

Bergmann, M. (2005), 'Defeaters and Higher Level Requirements', *The Philosophical Quarterly*, 55/220: 419–436.

—— (2006), *Justification without Awareness*: *A Defense of Epistemic Externalism* (Oxford: Oxford University Press).

—— (2009), 'Rational Disagreement after Full Disclosure', *Episteme*, 6/03: 336–353.

Bergmann, M., and Kain, P. (2014) (eds.), *Challenges to Moral and Religious Belief*: *Disagreement and Evolution* (Oxford: Oxford University Press).

Berker, S. (2014), 'Does Evolutionary Psychology Show That Normativity Is Mind-Dependent?', in J. D'Arms and D. Jacobson (eds.), *Moral Psychology and Human Agency*. *Philosophical Essays on the Science of Ethics* (Oxford: Oxford University Press), 215–52.

Berry, S. E. (2017), *Coincidence Avoidance and Formulating The Access Problem*, Manuscript.

Bicchieri, C. (1990), 'Norms of Cooperation', *Ethics*, 100/4: 838–861.

—— (1993), *Rationality and Coordination* (Cambridge: Cambridge University Press).

—— (2006), *The Grammar of Society* (Cambridge: Cambridge University Press).

—— (2017), *Norms in the Wild*: *How to Diagnose, Measure, and Change Social Norms* (Oxford: Oxford University Press).

Blackburn, S. (1984), *Spreading the Word*: *Groundings in the Philosophy of Language* (Oxford: Clarendon Press).

Boehm, C. (2001), *Hierarchy in the Forest*: *The Evolution of Egalitarian Behavior* (Cambridge, MA: Harvard University Press).

—— (2012), *Moral Origins*: *The Evolution of Virtue, Altruism, and Shame* (New York, NY: Basic Books).

Bogardus, T. (2013), 'The Problem of Contingency for Religious Belief', *Faith and Philosophy*, 30/4: 371–392.

—— (2014), 'Knowledge Under Threat', *Philosophy and Phenomenological Research*, 88/2: 289–313.

—— (2016), 'Only All Naturalists Should Worry About Only One Evolutionary Debunking Argument', *Ethics*, 126/3: 636–661.

BonJour, L. (1998), *In Defense of Pure Reason*: *A Rationalist Account of A Priori Justification* (Cambridge: Cambridge University Press).

BonJour, L., and Sosa, E. (2003) (eds.), *Epistemic Justification*: *Internalism vs. Externalism, Foundations vs. Virtues* (Malden, MA: Wiley-Blackwell).

Boudry, M., and Vlerick, M. (2014), 'Natural Selection Does Care about Truth', *International Studies in the Philosophy of Science*, 28/1: 65–77.

Bowles, S. (2016), *The Moral Economy*: *Why Good Incentives are No Substitute for Good Citizens* (New Haven, CT: Yale University Press).

Bowles, S., and Gintis, H. (2011), *A Cooperative Species*: *Human Reciprocity and its Evolution* (Princeton, NJ: Princeton University Press).

Boyd, R. (1988 [1995]), 'How to be a Moral Realist', in G. Sayre-McCord (ed.), *Essays on Moral Realism* (Ithaca, NY: Cornell University Press), 181–228.

Braddock, M. (2016), 'Evolutionary Debunking. Can Moral Realists Explain the Reliability of Our Moral Judgments?', *Philosophical Psychology*, 29/6: 844–857.

—— (2017), 'Debunking Arguments from Insensitivity', *International Journal for the Study of Skepticism*, 2017: 91–113.

Brandon, R. N. (1990), *Adaptation and Environment* (Princeton, NJ: Princeton University Press).

Brandt, R. B. (1944), 'The Significance of Differences of Ethical Opinion For Ethical Rationalism', *Philosophy and Phenomenological Research*, 4/4: 469–495.

—— (1954), *Hopi Ethics: A Theoretical Analysis* (Chicago, IL: University of Chicago Press).

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

Brosnan, K. (2011), 'Do the Evolutionary Origins of Our Moral Beliefs Undermine Moral Knowledge?', *Biology and Philosophy*, 26/1: 51–64.

Buchanan, A., and Powell, R. (2015), 'The Limits of Evolutionary Explanations of Morality and Their Implications for Moral Progress', *Ethics*, 126/1: 37–67.

—— (2016), 'Toward a Naturalistic Theory of Moral Progress', *Ethics*, 126/4: 983–1014.

Cameron, R. P. (2010), 'The Grounds of Necessity', *Philosophy Compass*, 5/4: 348–358.

Casullo, A. (1988), 'Revisability, Reliabilism, and A Priori Knowledge', *Philosophy and Phenomenological Research*, 49/2: 187.

—— (2003), *A Priori Justification* (Oxford: Oxford University Press).

—— (2008), 'Defeasible A Priori Justification. A Reply to Thurow', *The Philosophical Quarterly*, 58/231: 336–343.

—— (2016), 'Pollock and Sturgeon on Defeaters', *Synthese*, 2016.

Chan, T. H. W. (2014), *The Aim of Belief* (Oxford: Oxford University Press).

Chandler, J. (2013), 'Defeat Reconsidered', *Analysis*, 73/1: 49–51.

Chisholm, R. M. (1964), 'The Ethics of Requirement', *American Philosophical Quarterly*, 1/2: 147–153.

—— (1989), *Theory of knowledge* (London: Prentice Hall).

Christensen, D. (2007), 'Epistemology of Disagreement. The Good News', *Philosophical Review*, 116/2: 187–217.

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

—— (2011), 'Disagreement, Question-Begging and Epistemic Self-Cristicism', *Philosopher's Imprint*, 6/11: 1–22.

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

—— (2014), 'Moral Epistemology. The Mathematics Analogy', *Noûs*, 48/2: 238–255.

—— (2015), 'Justification and Explanation in Mathematics and Morality', in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics*. *Volume 10* (Oxford: Oxford University Press), 80–103.

—— (2016a), 'Debunking and Dispensability', in U. D. Leibowitz and N. Sinclair (eds.), *Explanation in Ethics and Mathematics* (Oxford: Oxford University Press), 23–36.

—— (2016b), 'Objectivity in Ethics and Mathematics', in B. Colburn (ed.), *Proceedings of the Aristotelian Society*. *The Virtual Issue Vol 3*, 98–109.

—— (2017a), 'Debunking Arguments. Mathematics, Logic, and Modal Security', in M. Ruse and R. J. Richards (eds.), *Cambridge Handbook to Evolutionary Ethics* (Cambridge: Cambridge University Press), 202–9.

—— (2017b), 'Objectivity and Reliability', *Canadian Journal of Philosophy*, 47/6: 841–855.

—— (2017c), 'What is the Benacerraf Problem?', in F. Pataut (ed.), *Truth, Objects, Infinity*. *New Perspectives on the Philosophy of Paul Benacerraf* (Dordrecht: Springer), 17–44.

—— (2017d), *Set-Theoretic Pluralism and the Benacerraf Problem*, Manuscript.

Clifford, W. K. (1879), *Lectures and Essays* (London: Macmillan).

Climenhaga, N. (2017), 'Knowledge and Certainty', PhD Thesis (Notre Dame, IA, University of Notre Dame).

Cohen, S. (2013), 'A Defense of the (Almost) Equal Weight View', in D. Christensen and J. Lackey (eds.), *The Epistemology of Disagreement. New Essays* (Oxford: Oxford University Press), 98–117.

Collin, J. H. (2017), *Epistemic Luck for Necessary Truth*, Manuscript (Academia.edu).

Conee, E., and Feldman, R. (1998), 'The Generality Problem for Reliabilism', *Philosophical Studies*, 89/1: 1–29.

—— (2001), 'Internalism Defended', *American Philosophical Quarterly*, 38/1: 1–18.

Copp, D. (2008), 'Darwinian Skepticism about Moral Realism', *Philosophical Issues*, 18: 186–206.

Crow, D. (2016), 'Causal Impotence and Evolutionary Influence. Epistemological Challenges for Non-Naturalism', *Ethical Theory and Moral Practice*, 19/2: 379–395.

Cruz, H. d., Boudry, M., Smedt, J. d. et al. (2011), 'Evolutionary Approaches to Epistemic Justification', *Dialectica*, 65/4: 517–535.

Cummins, D. (1996), 'Evidence for the Innateness of Deontic Reasoning', *Mind & Language*, 11/2: 160–190.

Cuneo, T. (2007), *The Normative Web*: *An Argument for Moral Realism* (Oxford: Oxford University Press).

—— (2014), *Speech and Morality*: *On the Metaethical Implications of Speaking* (Oxford: Oxford University Press).

Cuneo, T., and Shafer-Landau, R. (2014), 'The Moral Fixed Points. New Directions For Moral Nonnaturalism', *Philosophical Studies*, 171/3: 399–443.

Curry, O. S. (2016), 'Morality as Cooperation. A Problem-Centred Approach', in T. K. Shackelford and R. D. Hansen (eds.), *The Evolution of Morality* (Dordrecht: Springer), 27–51.

Curry, O. S., Mullins, D. A., and Whitehouse, H. (2017), *Is it Good to Cooperate?*: *Testing the Theory of Morality-as-Cooperation in 60 Cocieties*, Manuscript / Revise & Resubmit.

Darwall, S. L. (1995), *The British Moralists and the Internal 'Ought'*: *1640 - 1740* (Cambridge: Cambridge University Press).

Darwall, S. L., Gibbard, A., and Railton, P. (1992), 'Toward Fin de siecle Ethics. Some Trends', *The Philosophical Review*, 101/1: 115–189.

Darwin, C. (1838), 'M Notebook. Metaphysics on Morals and Speculations on Expression' <http://darwin-online.org.uk/content/frameset?itemID=CUL-DAR125.-&viewtype=text&pageseq=1>, accessed 28 Jan 2018.

—— (1871 [2004]), *The Descent of Man, and Selection in Relation to Sex*, A. Desmond and J. R. Moore (London: Penguin Books).

Davidson, D. (1984), *Inquiries into Truth and Interpretation* (Oxford: Oxford University Press).

Decker, J., and Groll, D. (2013), 'On the (In)Significance of Moral Disagreement for Moral Knoweldge', in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics. Volume 8* (Oxford: Oxford University Press), 140–67.

Deem, M. J. (2016), 'Dehorning the Darwinian Dilemma for Normative Realism', *Biology and Philosophy*, 31/5: 727–746.

Dennett, D. C. (1995), *Darwin's Dangerous Idea*: *Evolution and the Meanings of Life* (New York, NY: Simon & Schuster).

Descartes, R. (1998 [1637]), *Discourse on Method and Meditations on First Philosophy*, D. A. Cress (Indianapolis, IN: Hackett Publishing).

DeScioli, P., and Kurzban, R. (2013), 'A Solution to the Mysteries of Morality', *Psychological bulletin*, 139/2: 477–496.

DiPaolo, J., and Simpson, R. M. (2016), 'Indoctrination Anxiety and the Etiology of Belief', *Synthese*, 193/10: 3079–3098.

Dogramaci, S. (2017), 'Explaining our Moral Reliability', *Pacific Philosophical Quarterly*, 98/S1: 71–86.

Doris, J. M. (2009), 'Genealogy and Evidence: Prinz on the History of Morals', *Analysis*, 69/4: 704–713.

Doris, J. M., and Plakias, A. (2008), 'How to Argue about Disagreement: Evaluative Diversity and Moral Realism', in W. Sinnott-Armstrong (ed.), *Moral psychology. The Cognitive Science of Morality. Intuition and Diversity* (Cambridge, MA: MIT Press), 303–31.

Doris, J. M., and Stich, S. P. (2012), 'As a Matter of Fact. Empirical Perspectives on Ethics', in S. P. Stich (ed.), *Knowledge, Rationality, and Morality* (Oxford: Oxford University Press), 247–84.

Doughterty, T., and Rysiew, P. (2009), 'Fallibilism, Epistemic Possibility, and Concessive Knowledge Attributions', *Philosophy and Phenomenological Research*, 78/1: 123–132.

Dretske, F. I. (1981), *Knowledge and the Flow of Information* (Cambridge, MA: MIT Press).

Dugatkin, L. A. (2000), *Cheating Monkeys and Citizen Bees*: *The Nature of Cooperation in Animals and Humans* (Cambridge, MA: Harvard University Press).

Duhem, P., Wiener, P. P., and Vuillemin, J. (1906 [1991]), *The Aim and Structure of Physical Theory* (Princeton, NJ: Princeton University Press).

Dworkin, R. (1996), 'Objectivity and Truth. You'd Better Believe It', *Philosophy & Public Affairs*, 25/2: 87–139.

Dwyer, S. (2006), 'How good is the linguistic analogy?', in P. Carruthers, S. Laurence, and S. P. Stich (eds.), *The Innate mind. Culture and Cognition*, Volume 2 (Oxford: Oxford University Press), 237–56.

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

—— (2010), 'How to Disagree about how to Disagree', in R. Feldman and T. A. Warfield (eds.), *Disagreement* (Oxford: Oxford University Press).

Endler, J. A. (1986), *Natural selection in the Wild* (Princeton, NJ: Princeton University Press).

Enoch, D. (2009), 'How is Moral Disagreement a Problem for Realism?', *The Journal of Ethics*, 13/1: 15–50.

—— (2010), 'The Epistemological Challenge to Metanormative Realism. How Best to Understand It, and How to Cope With It', *Philosophical Studies*, 148/3: 413–438.

—— (2011a), 'Not Just a Truthometer. Taking Oneself Seriously (but not Too Seriously) in Cases of Peer Disagreement', *Mind*, 119/476: 953–997.

—— (2011b), *Taking Morality Seriously*: *A Defense of Robust Realism* (Oxford: Oxford University Press).

Faraci, D. (2016), *Knowledge, Necessity and Defeat*, Manuscript.

Farber, P. L. (1994), *The Temptations of Evolutionary Ethics* (Berkeley, CA: University of California Press).

Fehr, E., and Fischbacher, U. (2003), 'The Nature of Human Altruism', *Nature*, 425/6960: 785–791.

Feldman, R. (2006), 'Epistemological Puzzles about Disagreement', in S. C. Hetherington (ed.), *Epistemology Futures* (Oxford: Clarendon Press).

Feldman, R., and Conee, E. (1985), 'Evidentialism', *Philosophical Studies*, 48/1: 15–34.

Feldman, R., and Warfield, T. A. (2010) (eds.), *Disagreement* (Oxford: Oxford University Press).

Field, H. (1989), *Realism, Mathematics and Modality* (Oxford: Wiley-Blackwell).

—— (2001), *Truth and the Absence of Fact* (Oxford: Oxford University Press).

Fine, K. (1994), 'Essence and Modality. The Second Philosophical Perspectives Lecture', *Philosophical Perspectives*, 8: 1–16.

Fisher, M., Knobe, J., Strickland, B. et al. (2017), 'The Influence of Social Interaction on Intuitions of Objectivity and Subjectivity', *Cognitive Science*, 41/4: 1119–1134.

FitzPatrick, W. J. (2008), 'Robust Ethical Realism, Non-Naturalism, and Normativity', in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics. Volume 3* (Oxford: Oxford University Press), 159–205.

—— (2014a), 'Skepticism about Naturalizing Normativity. In Defense of Ethical Nonnaturalism', *Res Philosophica*, 91/4: 559–588.

—— (2014b), 'Why There is No Darwinian Dilemma for Ethical Realism', in M. Bergmann and P. Kain (eds.), *Challenges to Moral and Religious Belief. Disagreement and Evolution* (Oxford: Oxford University Press), 237–55.

—— (2015), 'Debunking Evolutionary Debunking of Ethical Realism', *Philosophical Studies*, 172/4: 883–904.

Flew, A. G. N. (1970), *Evolutionary Ethics* (London: Macmillan).

Foley, R. (2001), *Intellectual Trust in Oneself and Others* (Cambridge: Cambridge University Press).

Forgas, J. P., Jussim, L. J., and van Lange, P. A. M. (2016) (eds.), *The Social Psychology of Morality* (New York, NY: Psychology Press).

Frances, B. (2010), 'The Reflective Epistemic Renegade', *Philosophy and Phenomenological Research*, 81/2: 419–463.

—— (2014), *Disagreement* (Malden, MA: Polity Press).

Frank, R. H. (1988), *Passions within Reason*: *The Strategic Role of the Emotions* (New York, NY: Norton & Company).

Frankena, W. K. (1939), 'The Naturalistic Fallacy', *Mind*, 48/192: 464–477.

Fraser, B. (2014), 'Evolutionary Debunking Arguments and the Reliability of Moral Cognition', *Philosophical Studies*, 168/2: 457–473.

Freud, S. (1927), *The Future of an Illusion*: *Civilization, Society and Religion* (London: Penguin Books).

Fumerton, R. (1988), 'Foundationalism, Conceptual Regress, and Reliabilism', *Analysis*, 48/4: 178–184.

Gardiner, G. (2017), 'Safety's Swamp. Against The Value of Modal Stability', *American Philosophical Quarterly*, 54/2: 119–129.

Geach, P. T. (1965), 'Assertion', *The Philosophical Review*, 74/4: 449.

Gelfert, A. (2011), 'Who is an Epistemic Peer?', *Logos & Episteme*, 2/4: 507–514.

Gettier, E. (1963), 'Is Justified True Belief Knowledge?', *Analysis*, 23/6: 121–123.

Gibbard, A. (1990), *Wise Choices, Apt Feelings*: *A Theory of Normative Judgment* (Cambridge, MA: Harvard University Press).

—— (2003), *Thinking How to Live* (Cambridge, MA: Harvard University Press).

Goldberg, S. C. (2013), 'Disagreement, Defeat, and Assertion', in D. Christensen and J. Lackey (eds.), *The Epistemology of Disagreement. New Essays* (Oxford: Oxford University Press), 167–89.

Goldman, A. I. (1986), *Epistemology and Cognition* (Cambridge, MA: Harvard University Press).

Golub, C. (2017), 'Expressivism and the Reliability Challenge', *Ethical Theory and Moral Practice*, 2017.

Goodwin, G. P., and Darley, J. M. (2008), 'The Psychology of Meta-ethics: Exploring Objectivism', *Cognition*, 106/3: 1339–1366.

—— (2010), 'The Perceived Objectivity of Ethical Beliefs. Psychological Findings and Implications for Public Policy', *Review of Philosophy and Psychology*, 1/2: 161–188.

—— (2012), 'Why Are Some Moral Beliefs Perceived to be More Objective Than Others?', *Journal of Experimental Social Psychology*, 48/1: 250–256.

Gowans, C. W. (2000) (ed.), *Moral Disagreements*: *Classic and Contemporary Readings* (London: Routledge).

Graber, A. (2012), 'Medusa's Gaze Reflected. A Darwinian Dilemma for Anti-Realist Theories of Value', *Ethical Theory and Moral Practice*, 15/5: 589–601.

Greco, J. (1999), 'Agent Reliabilism', *Philosophical Perspectives*, 13: 273–296.

—— (2009), 'Knowledge and Success from Ability', *Philosophical Studies*, 142/1: 17–26.

—— (2010), 'Knowledge as Credit for True Belief', in M. R. DePaul and L. Zagzebski (eds.), *Intellectual Virtue. Perspectives from Ethics and Epistemology* (Oxford: Clarendon Press), 111–34.

Greene, J. D. (2008), 'The Secret Joke of Kant's Soul', in W. Sinnott-Armstrong (ed.), *Moral psychology. The neuroscience of morality: Emotion, brain disorders, and development* (Cambridge, MA: MIT Press), 35–79.

—— (2010), 'Notes on the 'Normative Insignificance of Neuroscience' by Selim Berker' <https://static1.squarespace.com/static/54763f79e4b0c4e55ffb000c/t/54cb945ae4b001aedee69e81/1422627930781/notes-on-berker.pdf>, accessed 18 Mar 2018.

—— (2013), *Moral Tribes*: *Emotion, Reason, and the Gap Between Us and Them* (New York, NY: The Penguin Press).

Grefte, J. de (2017), 'Epistemic Justification and Epistemic Luck', *Synthese*, 5/1: 53.

Grundmann, T. (2011), 'Defeasibility Theory', in S. Bernecker and D. Pritchard (eds.), *The Routledge companion to epistemology* (New York, NY: Routledge), 156–66.

Hales, S. D. (2016), 'Why Every Theory of Luck is Wrong', *Noûs*, 50/3: 490–508.

Haley, K. J., and Fessler, D. M.T. (2005), 'Nobody's Watching?', *Evolution and Human Behavior*, 26/3: 245–256.

Hamilton, W. D. (1963), 'The Evolution of Altruistic Behavior', *The American Naturalist*, 97/896: 354–356.

Handfield, T. (2016), 'Genealogical Explanations of Chance and Morals', in U. D. Leibowitz and N. Sinclair (eds.), *Explanation in Ethics and Mathematics* (Oxford: Oxford University Press), 58–82.

Hanson, L. (2017), 'The Real Problem with Evolutionary Debunking Arguments', *The Philosophical Quarterly*, 67/268: 508-33.

Hare, R. M. (1963), *The Language of Morals* (Oxford: Oxford University Press).

Harman, G. (1975), 'Moral Relativism Defended', *The Philosophical Review*, 84/1: 3–22.

—— (1977), *The Nature of Morality*: *An Introduction to Ethics* (New York, NY: Oxford University Press).

—— (1986), 'Moral Explanations of Natural Facts. Can Moral Claims be tested against Moral Reality?', *The Southern Journal of Philosophy*, 24/S1: 57–68.

—— (1988), *Change in View*: *Principles of Reasoning* (Cambridge, MA: The MIT Press).

Harman, G., and Thomson, J. J. (1996) (eds.), *Moral Relativism and Moral Objectivity* (Cambridge, MA: Wiley-Blackwell).

Harms, W. F. (2004), *Information and Meaning in Evolutionary Processes* (Cambridge: Cambridge University Press).

Hart, H. L. A. (1948), 'The Ascription of Responsibility and Rights', *Proceedings of the Aristotelian Society*, 49: 171–194.

Hauser, M. D. (2006), *Moral Minds*: *How Nature Designed our Universal Sense of Right and Wrong* (New York, NY: Ecco).

Heiphetz, L., and Young, L. L. (2017), 'Can Only One Person be Right? The Development of Objectivism and Social Preferences Regarding Widely Shared and Controversial Moral Beliefs', *Cognition*, 167: 78–90.

Henrich, J. P. (2016), *The Secret of our Success*: *How Culture is Driving Human Evolution, Domesticating our Species, and Making Us Smarter* (Princeton, NJ: Princeton University Press).

Henrich, J. P., Boyd, R., Bowles, S. et al. (2001), 'In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies', *American Economic Association*, 91/2: 73–78.

Hill, K. (2002), 'Altruistic Cooperation During Foraging by the Ache, and the Evolved Human Predisposition to Cooperate', *Human nature*, 13/1: 105–128.

Hill, S. (2016), 'From Isolation to Skepticism', *Erkenntnis*, 81/3: 649–668.

Hills, A. (2010), *The Beloved Self*: *Morality and the Challenge from Egoism* (Oxford: Oxford University Press).

Hirvelä, J. (2017), 'Is it Safe to Disagree?', *Ratio*, 30/3: 305–321.

Horn, J. (2017), 'Moral Realism, Fundamental Moral Disagreement, and Moral Reliability', *The Journal of Value Inquiry*, 51/3: 363–381.

Horwich, P. (2016), *Probability and Evidence* (Cambridge: Cambridge University Press).

Hricko, J., and Leben, D. (2017), 'In Defense of Best-Explanation Debunking Arguments in Moral Philosophy', *Review of Philosophy and Psychology*, 2017.

Huemer, M. (2001), 'The Problem Of Defeasible Justification', *Erkenntnis*, 54/3: 375–397.

—— (2005), *Ethical Intuitionism* (Basingstoke: Palgrave Macmillan).

—— (2007), 'Moore's Paradox and the Norms of Belief', in S. Nuccetelli and G. Seay (eds.), *Themes from G.E. Moore. New Essays in Epistemology and Ethics* (Oxford: Oxford University Press), 142–57.

—— (2016), 'A Liberal Realist Answer to Debunking Skeptics. The Empirical Case for Realism', *Philosophical Studies*, 173/7: 1983–2010.

Hume, D. (1738 [2007]), *A Treatise of Human Nature*: *A Critical Edition* (Oxford: Clarendon Press).

Ichikawa, J., and Steup, M. (2017), 'The Analysis of Knowledge', in E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. *Fall 2017 Edition*.

Jackson, A. (2012), 'Two Ways to Put Knowledge First', *Australasian Journal of Philosophy*, 90/2: 353–369.

Jackson, F. (1998), *From Metaphysics to Ethics*: *A Defence of Conceptual Analysis* (Oxford: Clarendon Press).

Jackson, F., and Pettit, P. (1995), 'Moral Functionalism and Moral Motivation', *The Philosophical Quarterly*, 45/178: 20–40.

James, S. M. (2011), *An Introduction to Evolutionary Ethics* (Malden, MA: Wiley-Blackwell).

James, W. (1902 [2002]), *Varieties of Religious Experience*: *A Study in Human Nature* (London: Routledge).

Janvid, M. (2008), 'The Experiential Defeasibility and Overdetermination of A Priori Justification', *Journal of Philosophical Research*, 33: 121–128.

Jonas, S. (2017), 'Access Problems and Explanatory Overkill', *Philosophical Studies*, 174/11: 2731–2742.

Jones, S. (2001), *Almost Like A Whale*: *The Origin of Species Updated* (London: Black Swan).

Joyce, R. (2001), *The Myth of Morality* (Cambridge: Cambridge University Press).

—— (2006), *The Evolution of Morality* (Cambridge, MA: MIT Press).

—— (2013a), 'Irrealism and the Genealogy of Morals', *Ratio*, 26/4: 351–372.

—— (2013b), 'The Evolutionary Debunking of Morality', in J. Feinberg and R. Shafer-Landau (eds.), *Reason and Responsibility. Readings in Some Basic Problems of Philosophy* (Boston, MA: Wadsworth), 527–34.

—— (2016a), *Essays in Moral Skepticism* (Oxford: Oxford University Press).

—— (2016b), 'Evolution and Moral Naturalism', in K. J. Clark (ed.), *The Blackwell Companion to Naturalism* (Hoboken, NJ: Wiley-Blackwell), 369–85.

—— (2016c), 'Evolution, Truth-tracking, and Moral Scepticism', in , *Essays in Moral Skepticism* (Oxford: Oxford University Press), 142–58.

—— (2016d), 'Reply: Confessions of a Modest Debunker', in U. D. Leibowitz and N. Sinclair (eds.), *Explanation in Ethics and Mathematics* (Oxford: Oxford University Press), 124–45.

—— (2018), 'Arguments from Moral Disagreement to Moral Skepticism', in D. E. Machuca (ed.), *Moral Skepticism. New Essays* (New York, NY: Routledge).

Kahane, G. (2011), 'Evolutionary Debunking Arguments', *Noûs*, 45/1: 103–125.

—— (2012), 'Must Metaethical Realism Make a Semantic Claim?', *Journal of Moral Philosophy*, 10/2: 148–178.

Kalf, W. F. (2013), 'Moral Error Theory. A Cognitivist Realist Defence', PhD Thesis (University of Leeds).

Kelly, T. (2005), 'The Epistemic Significance of Disagreement', in T. S. Gendler and J. P. Hawthorne (eds.), *Oxford Studies in Epistemology*, Volume 1 (Oxford: Clarendon Press), 167–96.

—— (2010), 'Peer-Disagreement and Higher-Order Evidence', in R. Feldman and T. A. Warfield (eds.), *Disagreement* (Oxford: Oxford University Press), 183–217.

—— (2013), 'Disagreement and the Burdens of Judgment', in D. Christensen and J. Lackey (eds.), *The Epistemology of Disagreement. New Essays* (Oxford: Oxford University Press), 31–53.

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.

Kitcher, P. (1980), 'A Priori Knowledge', *The Philosophical Review*, 89/1: 3–23.

—— (2011), *The Ethical Project* (Cambridge, MA: Harvard University Press).

—— (2012), *Preludes to Pragmatism*: *Toward a Reconstruction of Philosophy* (Oxford: Oxford University Press).

Klein, P. D. (1971), 'A Proposed Definition of Propositional Knowledge', *The Journal of Philosophy*, 68/16: 471.

Klenk, M. (forthcoming), 'Evolutionary Ethics', in C. Hendricks (ed.), *Introduction to Philosophy* (The Rebus Foundation), 1–16.

—— (2015a), '[Review of the book The Metaethical Implications of Speaking, by Terence Cuneo]', *Ethical Perspectives*, 22/2: 345–350.

—— (2015b), '[Review of the book Why Sex Matters - A Darwinian Look at Human Behaviour, by Bobbi S. Low]', *Metapsychology*, 19/36: 1–6.

—— (2016a), '[Review of the book A Natural History of Morality, by Michael Tomasello]', *Metapsychology*, 20/20: 1–7.

—— (2016b), '[Review of the book A Remarkable Journey. The Story of Evolution, by R. Paul Thompson.]', *The Quarterly Review of Biology*, 91/3: 362.

—— (2016c), '[Review of the book Robust Ethics. The Metaphysics and Epistemology of Godless Normative Realism, by Erik Wielenberg]', *Dialectica*, 70/3: 482–488.

—— (2016d), '[Review of the book The Origins of Fairness. How Evolution Explains our Moral Nature, by Nicolas Baumard]', *Metapsychology*, 20/36: 1–6.

—— (2016e), '[Review of the book The Social Psychology of Morality, edited by Joseph Forgas et al]', *Metapsychology*, 20/48: 1–6.

—— (2017a), 'Can Moral Realists Deflect Defeat Due to Evolutionary Explanations of Morality?', *Pacific Philosophical Quarterly*, 98/S1: 227–248.

—— (2017b), 'Measuring Moral Development', *De Filosoof*, 75: 21–23.

—— (2017c), 'Old Wine in New Bottles. Evolutionary Debunking Arguments and the Benacerraf-Field Challenge', *Ethical Theory and Moral Practice*, 20/4: 781–795.

—— (2017d), '[Review of the book Essays in Moral Skepticism, by Richard Joyce]', *Ethical Perspectives*, 24/1: 158–162.

—— (2017e), '[Review of the book The Moral Economy. Why Good Incentives are no substitute for good citizens, by Samuel Bowles]' <http://marxandphilosophy.org.uk/reviewofbooks/reviews/2017/2623>.

Knobe, S., and Nichols, S. (2008) (eds.), *Experimental Philosophy* (Oxford: Oxford University Press).

—— (2014) (eds.), *Experimental Philosophy*, Volume 2 (Oxford: Oxford University Press).

Kohlberg, L., and Hersh, R. H. (1977), 'Moral Development. A Review of the Theory', *Theory Into Practice*, 16/2: 53-59.

Kölbel, M. (2004), 'Faultless Disagreement', *Proceedings of the Aristotelian Society*, 104: 53–73.

Kornblith, H. (2010), 'Belief in the Face of Controversy', in R. Feldman and T. A. Warfield (eds.), *Disagreement* (Oxford: Oxford University Press), 29–52.

Kotzen, M. (2010), *A Formal Account of Epistemic Defeat*, Manuscript.

—— (2013), 'Multiple Studies and Evidential Defeat', *Noûs*, 47/1: 154–180.

Kramer, M. H. (2009), *Moral Realism as a Moral Doctrine* (Oxford: Wiley-Blackwell).

Kripke, S. A. (1984), *Naming and Necessity* (Oxford: Wiley-Blackwell).

Kumar, V. (2015), 'Moral Judgment as a Natural Kind', *Philosophical Studies*, 172/11: 2887–2910.

Kusch, M. (2006), *Psychologism*: *A Case Study in the Sociology of Philosophical Knowledge* (London: Routledge).

Kvanvig, J. L. (1998), 'Why Should Inquiring Minds Want to Know? "Meno" Problems and Epistemological Axiology', *Monist*, 81/3: 426–451.

—— (2007), 'Two Approaches to Epistemic Defeat', in D.-P. Baker (ed.), *Alvin Plantinga* (Cambridge: Cambridge University Press), 107–24.

Lackey, J. (2008), 'What Luck is Not', *Australasian Journal of Philosophy*, 86/2: 255–267.

Ladyman, J., and Ross, D. (2010), *Every Thing Must Go*: *Metaphysics Naturalized* (Oxford: Oxford University Press).

Laland, K. N., and Brown, G. R. (2011), *Sense and Nonsense*: *Evolutionary Perspectives on Human Behaviour* (Oxford: Oxford University Press).

Lasonen-Aarnio, M. (2010a), 'Is There a Viable Account of Well-Founded Belief?', *Erkenntnis*, 72/2: 205–231.

—— (2010b), 'Unreasonable Knowledge', *Philosophical Perspectives*, 24/1: 1–21.

—— (2013), 'The Dogmatism Puzzle', *Australasian Journal of Philosophy*, 92/3: 417–432.

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

Lazari-Radek, K. d., and Singer, P. (2012), 'The Objectivity of Ethics and the Unity of Practical Reason', *Ethics*, 123/1: 9–31.

Leben, D. (2013), 'When Psychology Undermines Beliefs', *Philosophical Psychology*, 27/3: 328–350.

Lehrer, K., and Paxson, T. (1969), 'Knowledge. Undefeated Justified True Belief', *The Journal of Philosophy*, 66/8: 225–237.

Leiter, B. (2004), 'The Hermeneutics of Suspicion. Recovering Marx, Nietzsche, and Freud', in B. Leiter (ed.), *The Future of Philosophy* (Oxford: Clarendon Press), 74–105.

Levy, N. (2014), *Hard luck*: *How Luck Undermines Free Will and Moral Responsibility* (Oxford: Oxford University Press).

Lewis, D. K. (1973), *Counterfactuals* (Malden, MA: Wiley-Blackwell).

—— (1979), 'Counterfactual Dependence and Time's Arrows', *Noûs*, 13/4: 455–476.

Lillehammer, H. (2010), 'Methods of Ethics and the Descent of Man. Darwin and Sidgwick on Ethics and Evolution', *Biology and Philosophy*, 25/3: 361–378.

—— (2016), ''An Assumption of Extreme Significance'. Moore, Ross, and Spencer on Ethics and Evolution', in U. D. Leibowitz and N. Sinclair (eds.), *Explanation in Ethics and Mathematics* (Oxford: Oxford University Press), 103–23.

Linnebo, Ø. (2006), 'Epistemological Challenges to Mathematical Platonism', *Philosophical Studies*, 129/3: 545–574.

Littlejohn, C. (2013), 'XV-The Russellian Retreat', *Proceedings of the Aristotelian Society*, 113/3pt3: 293–320.

Locke, D. (2014), 'Darwinian Normative Skepticism', in M. Bergmann and P. Kain (eds.), *Challenges to Moral and Religious Belief. Disagreement and Evolution* (Oxford: Oxford University Press), 220–36.

Loeb, D. (1998), 'Moral Realism and the Argument from Disagreement', *Philosophical Studies*, 90/3: 281–303.

Low, B. S. (2015), *Why Sex Matters*: *A Darwinian Look at Human Behavior* (Princeton, NJ: Princeton University Press).

Lutz, M. (forthcoming), 'What Makes Evolution a Defeater?', *Erkenntnis*, forthcoming: 1–22.

Lycan, W. G. (1988), *Judgement and Justification* (Cambridge: Cambridge University Press).

Lynch, M. P. (2009), 'The Values of Truth and the Truth of Values', in A. Haddock, A. Millar, and D. Pritchard (eds.), *Epistemic Value* (Oxford: Oxford University Press), 225–42.

Maagt, S. de (2017), 'Constructing Morality. Transcendental Arguments in Ethics', PhD Thesis (Utrecht, Utrecht University).

Machery, E., and Mallon, R. (2010), 'The Evolution of Morality', in J. M. Doris (ed.), *The Moral Psychology Handbook* (Oxford: Oxford University Press), 4–40.

Machuca, D. E. (2018) (ed.), *Moral Skepticism*: *New Essays* (New York, NY: Routledge).

Mackie, J. L. (1977), *Ethics*: *Inventing Right and Wrong* (London: Penguin Books).

Mason, K. (2010), 'Debunking Arguments and the Genealogy of Religion and Morality', *Philosophy Compass*, 5/9: 770–778.

Matheson, J. (2015), *The Epistemic Significance of Disagreement* (Basingstoke: Palgrave Macmillan).

Mayr, E. (1977), *Populations, Species, and Evolution*: *An Abridgment of 'Animal Species and Evolution'* (Cambridge, MA: Harvard University Press).

—— (2003), 'Introduction', in C. Darwin (ed.), *On the Origin of Species by Means of Natural Selection, or, The Preservation of Favoured Races in the Struggle for Life* (Cambridge, MA: Harvard University Press), xii–xxvii.

McGrath, S. (2008), 'Moral Disagreement and Moral Expertise', in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics*. *Volume 3* (Oxford: Oxford University Press), 87–108.

—— (2011), 'Moral Knowledge and Experience', in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics*. *Volume 6* (Oxford: Oxford University Press), 107–27.

Melis, G. (2014), 'Understanding Undermining Defeat', *Philosophical Studies*, 170/3: 433–442.

—— (2016), 'Undermining Defeat and Propositional Justification', *Argumenta*, 1/2: 271–280.

—— (2017), 'The Intertwinement of Propositional and Doxastic Justification', *Australasian Journal of Philosophy*, 2017: 1–13.

Millar, A. (2009), 'What is it that Cognitive Abilities are Abilities to Do?', *Acta Analytica*, 24/4: 223.

Mogensen, A. L. (2014), 'Evolutionary Debunking Arguments in Ethics', PhD Thesis (Oxford, Oxford University).

—— (2016a), 'Contingency Anxiety and the Epistemology of Disagreement', *Pacific Philosophical Quarterly*, 97/4: 590-611.

—— (2016b), 'Do Evolutionary Debunking Arguments Rest on a Mistake about Evolutionary Explanations?', *Philosophical Studies*, 173/7: 1799–1817.

—— (2017), 'Disagreements in Moral Intuition as Defeaters', *The Philosophical Quarterly*, 67/267: 282–302.

—— (2018), 'Ethics, Evolution, and the Coincidence Problem. A Skeptical Appraisal' <http://andreasmogensen.com/wp-content/uploads/2014/10/Coincidence-Problem-v5.pdf>, accessed 18 Mar 2018.

Moon, A. (2012), 'Three Forms of Internalism and the New Evil Demon Problem', *Episteme*, 9/04: 345–360.

—— (2017), 'Debunking Morality. Lessons from the EAAN Literature', *Pacific Philosophical Quarterly*, 98/S1: 208–226.

Moore, G. E. (1903 [1988]), *Principia Ethica* (Amherst, NY: Prometheus Books).

—— (1939), 'Proof of an External World', *Proceedings of the British Academy*, 23/5: 273–300.

Morillo, C. R. (1984), 'Epistemic Luck, Naturalistic Epistemology and the Ecology of Knowledge or What the Frog Should Have Told Dretske', *Philosophical Studies*, 46/1: 109–129.

Nagel, T. (2012), *Mind and Cosmos*: *Why the Materialist Neo-Darwinian Conception of Nature is Almost Certainly False* (Oxford: Oxford University Press).

Neta, R. (2009), 'Defeating the Dogma of Defeasibility', in P. Greenough and D. Pritchard (eds.), *Williamson on Knowledge* (Oxford: Oxford University Press), 161–82.

—— (2011), 'The Basing Relation', in S. Bernecker and D. Pritchard (eds.), *The Routledge companion to epistemology* (New York, NY: Routledge), 110–8.

Nichols, S. (2014), 'Process Debunking and Ethics', *Ethics*, 124/4: 727–749.

Nickerson, R. S. (1998), 'Confirmation Bias. A Ubiquitous Phenomenon in Many Guises', *Review of General Psychology*, 2/2: 175–220.

Nietzsche, F. W. (1887 [2013]), *On the Genealogy of Morals*: *A Polemic*, translated by Michael A. Scarpitti; with an introduction and notes by Robert C. Holub (London: Penguin Books).

Nisbett, R. E., and Cohen, D. (1996), *Culture of Honor*: *The Psychology of Violence in the South* (Boulder, CO: Westview Press).

Nozick, R. (1981), *Philosophical Explanations* (Cambridge, MA: Harvard University Press).

Oddie, G. (2009), *Value, Reality, and Desire* (Oxford: Oxford University Press).

Olson, J. (2014), *Moral Error Theory*: *History, Critique, Defence* (Oxford: Oxford University Press).

Palermos, S. O. (2011), 'Belief-Forming Processes, Extended', *Review of Philosophy and Psychology*, 2/4: 741–765.

Palmer, A. (2018), 'Thought Experiments', *The Economist,* 2018: 1–12.

Parfit, D. (2011a), *On Mhat Matters*: *Volume Two* (Oxford: Oxford University Press).

—— (2011b), *On What Matters*: *Volume One* (Oxford: Oxford University Press).

Paul, D. (2006), 'Darwin, Social Darwinism and Eugenics', in J. Hodge and G. Radick (eds.), *The Cambridge Companion to Darwin* (Cambridge: Cambridge University Press), 214–39.

Peacocke, C. (1999), *Being Known* (Oxford: Oxford University Press).

Pettigrew, R. (2016), *Accuracy and the Laws of Credence* (Oxford: Oxford University Press).

Pettit, P. (1991), 'Realism and Response-dependence', *Mind*, 100/4: 587–626.

Philipse, H. (2015), *Global EDA's in Ethics Revisited*, Manuscript.

Pichot, A. (2009), *The Pure Society*: *From Darwin to Hitler* (London: Verso).

Plantinga, A. (1993), *Warrant and Proper Function* (Oxford: Oxford University Press).

—— (1994), 'Naturalism Defeated' <https://www.calvin.edu/academic/philosophy/virtual_library/articles/plantinga_alvin/naturalism_defeated.pdf>, accessed 31 Oct 2017.

—— (2000), *Warranted Christian Belief* (Oxford: Oxford University Press).

—— (2002), 'Reply to Beilby's Cohorts', in J. K. Beilby (ed.), *Naturalism defeated? Essays on Plantinga's evolutionary argument against naturalism* (Ithaca, NY: Cornell University Press), 204–75.

Pollock, J. L. (1970), 'The Structure of Epistemic Justification', American Philosophical Quarterly: 62–78.

—— (1974), *Knowledge and Justification* (Princeton, NJ: Princeton University Press).

—— (1987), 'Defeasible Reasoning', *Cognitive Science*, 11/4: 481–518.

—— (1995), *Cognitive Carpentry*: *A Blueprint for How to Build a Person* (Cambridge, MA: MIT Press).

—— (2001), 'Defeasible Reasoning with Variable Degrees of Justification', *Artificial Intelligence*, 133: 233–282.

Pollock, J. L., and Cruz, J. (1999), *Contemporary Theories of Knowledge* (Lanham, MD: Rowman & Littlefield Publishers).

Pollock, J. L., and Gillies, A. S. (2000), 'Belief Revision and Epistemology', *Synthese*, 122/1/2: 69–92.

Pölzler, T. (2017), 'Are Moral Judgements Adaptations? Three Reasons Why It Is So Difficult to Tell', *South African Journal of Philosophy*, 36/3: 425–439.

Prinz, J. J. (2007), *The Emotional Construction of Morals* (Oxford: Oxford University Press).

—— (2009), 'Against Moral Nativism', in D. Murphy and M. A. Bishop (eds.), *Stich and His Critics* (Oxford: Wiley-Blackwell), 167–89.

Prior, A. N. (1960), 'The Autonomy of Ethics', *Australasian Journal of Philosophy*, 38/3: 199–206.

Pritchard, D. (2002), 'Recent Work on Radical Skepticism', *American Philosophical Quarterly*, 39/3: 215–257.

—— (2005), *Epistemic Luck* (Oxford: Oxford University Press).

—— (2009), 'Safety-based Epistemology. Wither Now?', *Journal of Philosophical Research*, 34: 33–45.

—— (2010), 'Cognitive Ability and the Extended Cognition Thesis', *Synthese*, 175/1: 133–151.

—— (2012), 'Anti-Luck Virtue Epistemology', *Journal of Philosophy*, 109/3: 247–279.

—— (2014), 'The Modal Account of Luck', *Metaphilosophy*, 45/4-5: 594–619.

Pritchard, D., Millar, A., and Haddock, A. (2012), *The Nature and Value of Knowledge*: *Three Investigations* (Oxford: Oxford University Press).

Pritchard, D., and Whittington, L. J. (2015) (eds.), *The Philosophy of Luck* (Hoboken, NJ: Wiley-Blackwell).

Pryor, J. (2000), 'The Skeptic and the Dogmatist', *Noûs*, 34/4: 517–549.

—— (2013), 'Problems for Credulism', in C. Tucker (ed.), *Seemings and Justification. New Essays on Dogmatism and Phenomenal Conservatism* (Oxford: Oxford University Press), 89–132.

Putnam, H. (2013), *Meaning and the Moral Sciences* (New York, NY: Routledge).

Quine, W. V. (1969), *Ontological Relativity and Other Essays* (New York, NY: Columbia University Press).

—— (1980), 'On What There Is', in W. V. Quine (ed.), *From a Logical Point of View. 9 Logico-Philosophical Essays* (Cambridge, MA: Harvard University Press), 1–19.

Rachels, J. (1998), 'Introduction', in J. Rachels (ed.), *Ethical Theory* (Oxford: Oxford University Press), 1–18.

Rai, T. S., and Holyoak, K. J. (2013), 'Exposure to Moral Relativism Compromises Moral Behavior', *Journal of Experimental Social Psychology*, 49/6: 995–1001.

Railton, P. (1986), 'Moral Realism', *The Philosophical Review*, 95/2: 163–207.

Richards, R. J. (1986), 'A Defense of Evolutionary Ethics', *Biology and Philosophy*, 1/3: 265–293.

—— (2017), 'Evolutionary Naturalism and Valuation', in M. Ruse and R. J. Richards (eds.), *Cambridge Handbook to Evolutionary Ethics* (Cambridge: Cambridge University Press), 129–42.

Richerson, P. J., and Boyd, R. (2006), *Not by Genes Alone*: *How Culture Transformed Human Evolution* (Chicago, IL: University of Chicago Press).

Riggs, W. (2007), 'Why Epistemologists Are So Down on Their Luck', *Synthese*, 158/3: 329–344.

Roeser, S. (2011), *Moral Emotions and Intuitions* (Basingstoke: Palgrave Macmillan).

Roland, J., and Cogburn, J. (2011), 'Anti-Luck Epistemologies and Necessary Truths', *Philosophia*, 39/3: 547–561.

Rorty, R. (1979), *Philosophy and The Mirror of Nature* (Princeton, NJ: Princeton University Press).

—— (2008), *Consequences of Pragmatism* (Minneapolis, MN: University of Minnesota Press).

Rosen, G. A. (2018), 'Metaphysical Relations in Metaethics', in T. McPherson and D. Plunkett (eds.), *The Routledge Handbook of Metaethics* (New York, NY: Routledge), 151–69.

Ross, W. D. (1930 [2007]), *The Right and the Good*, P. Stratton-Lake (Oxford: Clarendon Press).

Rovane, C. A. (2013), *The Metaphysics and Ethics of Relativism* (Cambridge, MA: Harvard University Press).

Ruse, M. (1995a), 'Evolutionary Ethics. A Phoenix Arisen', in R. P. Thompson (ed.), *Issues in Evolutionary Ethics* (Albany, NY: State University of New York Press), 225–48.

—— (1995b), *Evolutionary Naturalism*: *Selected Essays* (New York, NY: Routledge).

—— (1998), *Taking Darwin Seriously*: *A Naturalistic Approach to Philosophy* (Amherst, NY: Prometheus Books).

—— (2006), 'Is Darwinian Metaethics Possible (And If It Is, Is It Well Taken)?', in G. Boniolo and G. de Anna (eds.), *Evolutionary Ethics and Contemporary Biology* (Cambridge: Cambridge University Press), 13–26.

—— (2009), 'Evolution and Ethics. The Sociobiological Approach', in M. Ruse (ed.), *Philosophy after Darwin. Classic and Contemporary Readings* (Princeton, NJ: Princeton University Press), 489–511.

Ruse, M., and Richards, R. J. (2017) (eds.), *Cambridge Handbook to Evolutionary Ethics* (Cambridge: Cambridge University Press).

Ruse, M., and Wilson, E. O. (1986), 'Moral Philosophy as Applied Science', *Philosophy*, 61/236: 173–192.

—— (2006), 'Moral Philosophy as Applied Science', in E. Sober (ed.), *Conceptual issues in Evolutionary Biology* (Cambridge, MA: MIT Press), 555–74.

Russett, C. E. (1976), *Darwin in America*: *The Intellectual Response, 1865-1912* (San Francisco, CA: W.H. Freeman).

Sauer, H. (2012), 'Psychopaths and Filthy Desks', *Ethical Theory and Moral Practice*, 15/1: 95–115.

—— (2017), *Moral Judgments as Educated Intuitions* (Cambridge, MA: The MIT Press).

Sayre-McCord, G. (1986), 'The Many Moral Realisms', *The Southern Journal of Philosophy*, 24/S1: 1–22.

Scanlon, T. M. (2014), *Being Realistic About Reasons* (Oxford: Oxford University Press).

Schacter, D. L., Guerin, S. A., and St Jacques, P. L. (2011), 'Memory Distortion. An Adaptive Perspective', *Trends in Cognitive Sciences*, 15/10: 467–474.

Schafer, K. (2010), 'Evolution and Normative Scepticism', *Australasian Journal of Philosophy*, 88/3: 471–488.

—— (2014), 'Knowledge and Two Forms of Non-Accidental Truth', *Philosophy and Phenomenological Research*, 89/2: 373–393.

Schechter, J. (2010), 'The Reliability Challenge and the Epistemology of Logic', *Philosophical Perspectives*, 24/1: 437–464.

—— (2013), 'Could Evolution Explain Our Reliability about Logic?', in T. S. Gendler and J. P. Hawthorne (eds.), *Oxford Studies in Epistemology*, Volume 4 (Oxford: Oxford University Press), 214–50.

—— (2018), 'Explanatory Challenges in Metaethics', in T. McPherson and D. Plunkett (eds.), *The Routledge Handbook of Metaethics* (New York, NY: Routledge), 443–58.

Schlesinger, G. N. (1991), *The Sweep of Probability* (New Brunswick, NJ: University of Notre Dame Press).

Schloss, J. P. (2014), 'Darwinian Explanations of Morality. Accounting for the Normal but not the Normative', in H. Putnam, S. Neiman, and J. P. Schloss (eds.), *Understanding Moral Sentiments. Darwinian Perspective?* (New Brunswick, NJ: Transaction Publishers), 81–121.

Schneewind, J. B. (1998), *The Invention of Autonomy*: *A History of Modern Moral Philosophy* (Cambridge: Cambridge University Press).

Schroeder, M. A. (2008), 'What is the Frege-Geach Problem?', *Philosophy Compass*, 3/4: 703–720.

Secord, J. A. (2003), *Victorian Sensation*: *The Extraordinary Publication, Eeception, and Secret Authorship of Vestiges of the Natural History of Creation* (Chicago, IL: University of Chicago Press).

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

Severini, E., and Sterpetti, F. (2017), 'Darwinism in Metaethics. What If the Universal Acid Cannot be Contained?', *History and philosophy of the life sciences*, 39/3: 27.

Shafer-Landau, R. (2003), *Moral Realism*: *A Defence* (Oxford: Clarendon Press).

—— (2004), *Whatever Happened to Good and Evil?* (Oxford: Oxford University Press).

—— (2007), 'Moral and Theological Realism. The Explanatory Argument', *Journal of Moral Philosophy*, 4/3: 311–329.

—— (2012), 'Evolutionary Debunking, Moral Realism and Moral Knowledge', *Journal of Ethics and Social Philosophy*, 7/1: 1–37.

Sher, G. (2001), 'But I Could Be Wrong', *Social Philosophy and Policy*, 18/02: 64.

Sidgwick, H. (1981 [1874]), *The Methods of Ethics* (Indianapolis, IN: Hackett Publishing).

Silins, N. (2014), 'The Agony of Defeat?', *Philosophy and Phenomenological Research*, 88/3: 505–532.

Silva, P. (2016), 'Etiological Information and Diminishing Justification', *Inquiry*, 55/3: 1–25.

Sinclair, N. (forthcoming), 'Belief-Pills and the Possibility of Moral Epistemology', in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics* (Oxford: Oxford University Press).

Singer, P. (2005), 'Ethics and Intuitions', *The Journal of Ethics*, 9/3-4: 331–352.

Sinnott-Armstrong, W. (2006a), 'Moral Intuitionism Meets Empirical Psychology', in T. Horgan and M. Timmons (eds.), *Metaethics after Moore* (Oxford: Oxford University Press), 340–66.

—— (2006b), *Moral Skepticisms* (Oxford: Oxford University Press).

—— (2011), 'An Empirical Challenge to Moral Intuitionism', in J. Hernandez (ed.), *The New Intuitionism* (London: Continuum), 11–28.

—— (2014), 'Moral Disagreements with Psychopaths', in M. Bergmann and P. Kain (eds.), *Challenges to Moral and Religious Belief. Disagreement and Evolution* (Oxford: Oxford University Press), 40–60.

Sinnott-Armstrong, W., and Wheatley, T. (2013), 'Are Moral Judgments Unified?', *Philosophical Psychology*, 27/4: 451–474.

Skarsaune, K. O. (2011), 'Darwin and Moral Realism. Survival of the Iffiest', *Philosophical Studies*, 152/2: 229–243.

Skorupski, J. (1999), 'Irrealist Cognitivism', *Ratio*, 12/4: 436–459.

Slater, G. (2014), 'A Peircean Response to the Evolutionary Debunking of Moral Knowledge', *Zygon*, 49/3: 593–611.

Smyth, N. (2017), 'The Function of Morality', *Philosophical Studies*, 174/5: 1127–1144.

Sober, E. (1984a), 'Common Cause Explanation', *Philosophy of Science*, 51/2: 212–241.

—— (1984b), *The Nature of Selection*: *Evolutionary Theory in Philosophical Focus* (Chicago, IL: University of Chicago Press).

—— (1994), 'Prospects for an Evolutionary Ethics', in E. Sober (ed.), *From a Biological Point of View* (Cambridge: Cambridge University Press), 93–113.

—— (2009), 'Parsimony Arguments in Science and Philosophy. A Test Case for Naturalism', *Proceedings and Addresses of the American Philosophical Association*, 83/2: 117–155.

—— (2016), *Ockham's Razors*: *A User's Manual* (Cambridge: Cambridge University Press).

Sober, E., and Wilson, D. S. (1998), *Unto Others*: *The Evolution and Psychology of Unselfish Behavior* (Cambridge, MA: Harvard University Press).

—— (2011), 'Adaptation and Natural Selection Revisited', *Journal of Evolutionary Biology*, 24/2: 462–468.

Sosa, E. (1999), 'How to Defeat Opposition to Moore', *Philosophical Perspectives*, 13: 141–153.

—— (2005), 'Tracking, Competence, and Knowledge', in P. K. Moser (ed.), *The Oxford Handbook of Epistemology* (Oxford: Oxford University Press), 264–86.

—— (2007), *Apt Belief and Reflective Knowledge* (Oxford: Clarendon Press).

—— (2011), *Knowing Full Well* (Princeton, NJ: Princeton University Press).

Spencer, H. (1879), *The Data of Ethics* (London: Williams and Norgate).

—— (1893), *Social Statics* (New York, NY: Appleton).

Spohn, W. (2012), *The Laws of Belief*: *Ranking Theory and Its Philosophical Applications* (Oxford: Oxford University Press).

Srinivasan, A. (2015), 'The Archimedean Urge', *Philosophical Perspectives*, 29/1: 325–362.

Sripada, C., and Stich, S. P. (2006), 'A Framework for the Psychology of Norms', in P. Carruthers, S. Laurence, and S. P. Stich (eds.), *The Innate mind. Culture and Cognition*, Volume 2 (Oxford: Oxford University Press), 280–301.

Sterelny, K. (2012), *The Evolved Apprentice*: *How Evolution Made Humans Unique* (Cambridge, MA: MIT Press).

Sterelny, K., and Fraser, B. (2017), 'Evolution and Moral Realism', *The British Journal for the Philosophy of Science*, 68/4: 981-1006.

Steup, M. (2017), 'Epistemology', in E. N. Zalta (ed.), *Stanford Encyclopedia of Philosophy*. *Fall 2017 Edition*.

Stevenson, C. L. (1937), 'The Emotive Meaning of Ethical Terms', *Mind*, 46/181: 14–31.

—— (1963), *Facts and Values*: *Studies in Ethical Analysis* (New Haven, CT: Yale University Press).

Street, S. (2006), 'A Darwinian Dilemma for Realist Theories of Value', *Philosophical Studies*, 127/1: 109–166.

—— (2008a), 'Constructivism about Reasons', in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics. Volume 3* (Oxford: Oxford University Press), 207–45.

—— (2008b), 'Reply to Copp. Naturalism, Normativity, and the Varieties of Realism Worth Worrying About.', *Philosophical Issues*, 18/1: 207–228.

—— (2010), 'What is Constructivism in Ethics and Metaethics?', *Philosophy Compass*, 5/5: 363–384.

—— (2011), 'Mind-Independence Without the Mystery. Why Quasi-Realists Can't Have it Both Ways', in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics. Volume 6* (Oxford: Oxford University Press), 1–32.

—— (2012), 'Coming to Terms with Contingency. Humean Constructivism about Practical Reason', in J. Lenman and Y. Shemmer (eds.), *Constructivism in Practical Philosophy* (Oxford: Oxford University Press), 40–59.

—— (2016), 'Objectivity and Truth. You'd Better Rethink It', in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics. Volume 11* (Oxford: Oxford University Press), 293–334.

Streumer, B. (2017), *Unbelievable Errors*: *An Error Theory about all Normative Judgements* (Oxford: Oxford University Press).

Stroud, B. (1984), *The Significance of Philosophical Scepticism* (Oxford: Clarendon Press).

Sturgeon, N. L. (1986), 'Harman on Moral Explanations of Natural Facts', *The Southern Journal of Philosophy*, 24/S1: 69–78.

—— (1988 [1995]), 'Moral Explanations', in G. Sayre-McCord (ed.), *Essays on Moral Realism* (Ithaca, NY: Cornell University Press), 229–55.

—— (1992), 'Nonmoral Explanations', *Philosophical Perspectives*, 6: 97.

Sturgeon, S. (2014), 'Pollock on Defeasible Reasons', *Philosophical Studies*, 169/1: 105–118.

Sudduth, M. C. (2017), 'Defeaters in Epistemology' <http://www.iep.utm.edu/ep-defea/>, accessed 26 Sep 2017.

Swinburne, R. (2001), *Epistemic Justification* (Oxford: Clarendon Press).

Talbott, W. J. (forthcoming), 'A New Reliability Defeater for Evolutionary Naturalism', *Philosophy and Phenomenological Research*, forthcoming.

—— (2015), 'How Could a 'Blind' Evolutionary Process have Made Human Moral Beliefs Sensitive to Strongly Universal, Objective Moral Standards?', *Biology and Philosophy*, 30/5: 691–708.

Tersman, F. (2006), *Moral Disagreement* (Cambridge: Cambridge University Press).

—— (2013), 'Moral Disagreement. Actual vs. Possible', in D. E. Machuca (ed.), *Disagreement and Skepticism* (New York, NY: Routledge), 90–108.

—— (2014), 'Disagreement. Ethics and Elsewhere', *Erkenntnis*, 79/S1: 55–72.

—— (2016), 'Explaining the Reliability of our Moral Beliefs', in U. D. Leibowitz and N. Sinclair (eds.), *Explanation in Ethics and Mathematics* (Oxford: Oxford University Press), 37–58.

—— (2017), 'Debunking and Disagreement', *Noûs*, 51/4: 754-774.

Thompson, R. P. (2015), *A Remarkable Journey*: *The Story of Evolution* (London: Reaktion Books).

Thurow, J. C. (2006), 'Experientially Defeasible A Priori Justification', *The Philosophical Quarterly*, 56/225: 596–602.

Tomasello, M. (2016), *A Natural History of Human Morality* (Cambridge, MA: Harvard University Press).

Trivers, R. L. (1971), 'The Evolution of Reciprocal Altruism', *The Quarterly Review of Biology*, 46/1: 35–57.

Tropman, E. (2014), 'Evolutionary Debunking Arguments. Moral Realism, Constructivism, and Explaining Moral Knowledge', *Philosophical Explorations*, 17/2: 126–140.

Tudge, C. (2006), *The Variety of Life*: *A Survey And a Celebration of All The Creatures That Have Ever Lived* (Oxford: Oxford University Press).

Turri, J. (2010), 'On the Relationship between Propositional and Doxastic Justification', *Philosophy and Phenomenological Research*, 80/2: 312–326.

Unger, P. (1968), 'An Analysis of Factual Knowledge', *The Journal of Philosophy*, 65/6: 157–170.

—— (1978), *Ignorance*: *A Case for Scepticism* (Oxford: Oxford University Press).

van Inwagen, P. (2010), 'We're Right, They're Wrong', in R. Feldman and T. A. Warfield (eds.), *Disagreement* (Oxford: Oxford University Press), 10–28.

Vavova, K. (2014a), 'Debunking Evolutionary Debunking', in R. Shafer-Landau (ed.), *Oxford Studies in Metaethics*. *Volume 9* (Oxford: Oxford University Press), 76–101.

—— (2014b), 'Moral Disagreement and Moral Skepticism', *Philosophical Perspectives*, 28/1: 302–333.

—— (2015), 'Evolutionary Debunking of Moral Realism', *Philosophy Compass*, 10/2: 104–116.

—— (2018), 'Irrelevant Influences', *Philosophy and Phenomenological Research*, 96/1: 134–152.

Velleman, J. D. (2013), *Foundations for Moral Relativism* (Cambridge: Open Book Publishers).

Verplaetse, J., Braeckman, J., Schrijver, J. et al. (2009) (eds.), *The Moral Brain*: *Essays on the Evolutionary and Neuroscientific Aspects of Morality* (Dordrecht: Springer).

Warren, J. (2017), 'Epistemology Versus Non-causal Realism', *Synthese*, 194/5: 1643–1662.

Weatherson, B. (2004), 'Luminous Margins', *Australasian Journal of Philosophy*, 83/3: 373–383.

—— (2013), 'Disagreements, Philosophical and Otherwise', in D. Christensen and J. Lackey (eds.), *The Epistemology of Disagreement*. *New Essays* (Oxford: Oxford University Press), 1–17.

Wedgwood, R. (2007), *The Nature of Normativity* (Oxford: Clarendon Press).

—— (2010), 'The Moral Evil Demons', in R. Feldman and T. A. Warfield (eds.), *Disagreement* (Oxford: Oxford University Press), 216–46.

Weikart, R. (2006), *From Darwin to Hitler*: *Evolutionary Ethics, Eugenics, and Racism in Germany* (Basingstoke: Palgrave Macmillan).

West-Eberhard, M. J. (1999), 'Adaptation. Current Usage', in E. F. Keller and E. A. Lloyd (eds.), *Keywords in Evolutionary Biology* (Cambridge, MA: Harvard University Press), 13–8.

White, R. (2010), 'You Just Believe That Because…', *Philosophical Perspectives*, 24/1: 573–615.

Wielenberg, E. J. (2010), 'On the Evolutionary Debunking of Morality', *Ethics*, 120/3: 441–464.

—— (2014), *Robust Ethics*: *The Metaphysics and Epistemology of Godless Normative Realism* (Oxford: Oxford University Press).

—— (2016a), 'Ethics and Evolutionary Theory', *Analysis*, 76/4: 502-515.

—— (2016b), 'Evolutionary Debunking Arguments in Religion and Morality', in U. D. Leibowitz and N. Sinclair (eds.), *Explanation in Ethics and Mathematics* (Oxford: Oxford University Press), 83–102.

Wilkins, J. S., and Griffiths, P. E. (2013), 'Evolutionary Debunking Arguments in Three Domains. Fact, Value, and Religion', in G. W. Dawes and J. Maclaurin (eds.), *A New Science of Religion* (New York, NY: Routledge), 133–46.

Williams, B. (1973), 'Deciding to Believe', in , *Problems of the Self*. *Philosophical Papers 1956-1972* (Cambridge: Cambridge University Press), 136–51.

Williams, G. C. (1988), 'Huxley's Evolution and Ethics in Sociobiological Perspective', *Zygon*, 23/4: 383–407.

Williams, M. (2001), 'Contextualism, Externalism and Epistemic Standards', *Philosophical Studies*, 103/1: 1–23.

Williamson, T. (2000), *Knowledge and Its Limits* (Oxford: Oxford University Press).

Wilson, E. O. (1975 [2002]), *Sociobiology*: *The new synthesis* (Cambridge, MA: Harvard University Press).

Wittgenstein, L., Pears, D., and McGuinness, B. (1921 [2006]), *Tractatus Logico-Philosophicus* (New York, NY: Routledge).

Wong, D. B. (2006), *Natural Moralities*: *A Defense of Pluralistic Relativism* (Oxford: Oxford University Press).

Woods, J. (2016), 'Mathematics, Morality, and Self-Effacement', *Noûs*, 51/2: 161.

Wright, C. (1991), 'Scepticism and Dreaming. Imploding the Demon', *Mind*, C/397: 87–116.

—— (1992), *Truth and Objectivity* (Cambridge, MA: Harvard University Press).

Wright, J. C., Grandjean, P. T., and McWhite, C. B. (2013), 'The Meta-Ethical Grounding of Our Moral Beliefs. Evidence for Meta-ethical Pluralism', *Philosophical Psychology*, 26/3: 336–361.

Wright, R. (1995), *The Moral Animal*: *Evolutionary Psychology and Everyday Life* (New York, NY: Vintage Books).

Yamada, M. (2011), 'Getting It Right By Accident', *Philosophy and Phenomenological Research*, 83/1: 72–105.

Young, L. L., and Durwin, A. J. (2013), 'Moral Realism as Moral Motivation. The Impact of Meta-ethics on Everyday Decision-making', *Journal of Experimental Social Psychology*, 49/2: 302–306.

Young, R. W. (2003), 'Evolution of the Human Hand. The Role of Throwing and Clubbing', *Journal of Anatomy*, 202/1: 165–174.

Zagzebski, L. (2002), *Virtues of the Mind*: *An Inquiry Into the Nature of Virtue and the Ethical Foundations of Knowledge* (Cambridge: Cambridge University Press).

# List of Figures

# Samenvatting

Geeft de evolutionaire oorsprong van onze moraliteit reden om te denken dat we het moreel goede en slechte niet meer adequaat kunnen onderscheiden? Volgens sommige ethici (waaronder filosofen, psychologen, en biologen) beïnvloedt de evolutie onze morele overtuigingen, maar deze evolutionaire invloeden houden geen rekening met de waarheid van die overtuigingen. Evolutie werkt dus als een *verstoring* van menselijke opvattingen over moraal en wij kunnen niet meer volhouden dat we goed en kwaad adequaat kunnen onderscheiden. In mijn proefschrift onderzoek ik deze stelling grondig en leg ik de epistemologische voorwaarden bloot, waaronder evolutie onze morele overtuigingen inderdaad ontkracht.

In deze samenvatting introduceer ik het probleem en mijn oplossing vanuit een vogelperspectief. Die lezers die al bekend zijn met evolutionaire verklaringen van moraliteit, de metaethische positie van het morele objectivisme, en de epistemologische vragen die daaruit volgen, verwijs ik naar de introductie en hoofdstuk 1 van dit proefschrift, waar ik mijn onderzoeksprobleem iets dieper en in relatie tot recente literatuur introduceer. Om het onderzoeksprobleem vanuit een vogelperspectief te begrijpen zijn twee kwesties van belang: hoe evolutie de menselijke moraliteit beïnvloedt en waarom evolutie een verstorend middel zou kunnen zijn voor onze morele overtuigingen.

Laten we allereerst *overtuigingen* aangaande moraliteit wat nader bekijken, dat wil zeggen, wat mensen geloven over wat moreel verboden of toegestaan is. Alle menselijke culturen hechten bijvoorbeeld belang aan speciale verantwoordelijkheden van ouders tegenover kinderen; hieruit volgt het wijdverspreide idee dat ouders voor hun kinderen moeten zorgen. Waarom deze eenheid in morele waarden? Evolutie biedt een antwoord. Vanuit evolutionair oogpunt is te verwachten dat mensen zorgzaamheid ontwikkelen omdat onze voorouders door deze zorgzaamheid een grotere kans hadden om gezonde nakomelingen op te voeden. De meeste mensen ervaren een dergelijke zorgplicht nog altijd als een moreel verplichting en dit kan verklaard worden door middel van de evolutionaire psychologie. We kunnen dus beargumenteren dat sommige morele aannames evolutionair verklaard kunnen worden.

Charles Darwin, de grondlegger van de theorie van de evolutie speculeerde in 1870 dat onze morele overtuigingen fundamenteel zouden verschillen indien het proces van evolutie anders was gelopen.

Een enigszins extreem gedachte-experiment van Darwin illustreert zijn idee. Stel je voor dat ons evolutionaire pad ons naar een soort van 'menselijke bij' had gevoerd, vergelijkbaar in gedrag en samenleving met de hedendaagse bijenkolonies. In zulke bijenkolonies is er bijvoorbeeld maar één moeder, de koningin. De rest van de vrouwelijke bijen zijn steriele werkbijen. Als 'menselijke bijen' zouden we dus leven in een maatschappij met slechts één vruchtbare vrouw waar verder alleen de steriele dochters in leven zouden worden gelaten. Darwin vermoedde dat we het als menselijke bij als onze plicht zouden beschouwen om de vruchtbare vrouwelijke nakomelingen te doden. Volgens Darwin zou een dergelijke morele regel de evolutionaire aanpassing van de menselijke bij immers vergroten. Onze morele aannames zouden dus heel anders zijn geweest dan ze nu zijn, enkel omdat we toevallig op een ander evolutionair pad zitten. Volgens dit gedachte-experiment zijn enkele van de diepste menselijke morele overtuigingen beïnvloed door de evolutie en daarom contingent: een andere evolutionaire geschiedenis zou tot andere morele overtuigingen hebben geleid.

Beschouw nu de morele *waarheden*. Volgens het *Moreel Objectivisme* zijn er objectieve morele waarheden, welke onafhankelijk zijn van, onder andere, ons evolutionaire verleden. Het moreel objectivisme is een historisch gezien populaire positie binnen de metaethiek, die na een daling tijdens de jaren zeventig en tachtig recentelijk weer populairder aan het worden is. Wat moreel verboden of toegestaan is, staat los van de vraag of we het daadwerkelijk erkennen als moreel verboden of toegestaan. Het vermoorden van dochters en zonen is dus, vervolgens het moreel objectivisme, altijd onrechtvaardig, zelfs als we het als 'menselijke bijen' legitiem zouden vinden. Een ander voorbeeld: slavernij werd in de achttiende eeuw veelal nog steeds als moreel aanvaardbaar beschouwd, en toch was het verkeerd volgens het morele objectivisme. Hetzelfde geldt voor de verwerpelijkheid van foltering en vrouwenbesnijdenis, en voor de goedheid van rechtvaardigheid, om maar een paar voorbeelden te noemen.

Het probleem is dus het volgende. Ten eerste, lijkt de inhoud van onze diepste morele aannames beïnvloed te zijn door evolutie en dus hadden we ook andere

morele overtuigingen kunnen hebben, maar, ten tweede, volgens het moreel objectivisme lijkt er slechts *één* morele waarheid te bestaan. Het uitgangspunt van mijn onderzoek is de vraag of contingente, evolutionaire invloeden op onze morele aannames aan de ene kant, en vaste objectieve morele waarheden aan de andere kant verenigbaar zijn. Stel dat we het morele objectivisme accepteren. Hoe kunnen we dan overtuigt zijn dat onze morele vooronderstellingen kloppen?

Het eerder genoemde antwoord van veel ethici is dus dat we ofwel moreel objectivisme opgeven, ofwel onze morele overtuigingen als weerlegd beschouwen: we kunnen niet beweren dat we adequaat goed van kwaad kunnen onderscheiden. Het volgt dat we of een alternatief voor het moreel objectivisme moeten zoeken (zoals constructivisme) of, als we overtuigd zijn dat moreel objectivisme juist is, een positie van moreel scepticisme, het idee dat we niet kunnen weten wat moreel goed of slecht is, moeten aanvaarden. Ik verlaat de mogelijkheid om moreel objectivisme op te geven, omdat er veel argumenten zijn onafhankelijk van evolutionaire overwegingen voor of tegen moreel objectivisme, die ik niet kan behandelen binnen dit proefschrift. Mijn vraag is daarom of evolutie een probleem is als we ervan overtuigd zijn dat er objectieve morele waarheden zijn en niet of we moeten aannemen dat er objectieve morele waarheden zijn.

Maar deze sceptische conclusie klopt niet helemaal, omdat we gegronde epistemologische regels nodig hebben om ons te overtuigen dat we onze morele overtuigingen moeten opgeven op basis van een argument dat gestoeld is op evolutie. Om deze vraag te kunnen beantwoorden gebruik ik verschillende benaderingen uit de metaethiek en epistemologie om de omstandigheden bloot te leggen waar onze morele veronderstellingen evolutionair kunnen worden weerlegd. Vervolgens is het doel te onderzoeken of er geldige epistemologische regels zijn voor het opgeven van onze morele overtuigingen in het licht van evolutionaire verklaringen van moraliteit.[1]

[1] Het onderzoeksgebied van *epistemologie* verwijst onder anderen naar de voorwaarden van *gerechtvaardigde aannames* en de voorwaarden van *kennis*. Gissen is bijvoorbeeld in de meeste gevallen niet gerechtvaardigd en vormt daarom geen kennis. Aan de andere kant is het waarschijnlijker dat een aanname op basis van betrouwbaar bewijsmateriaal gerechtvaardigd is. Voor zover de epistemologie ons regels biedt voor de rationele, valide vorming en afwijzing van aannames, moet worden nagegaan of er geldige regels zijn die leiden tot een verwerping van onze morele veronderstellingen in het licht van de evolutie.

Bovendien beschouw ik evolutie beschouw ik als een potentiële *weerlegging* van onze morele overtuigingen. Weerleggende informatie, of weerleggers, komen we regelmatig tegen. Stel je bijvoorbeeld voor dat we een fabriek bezoeken en schijnbaar rode componenten op een transportband zien. Het is geaccepteerd, en vanuit een epistemologisch oogpunt gerechtvaardigd, om aan te nemen dat de componenten daadwerkelijk rood zijn. De voorman legt nu echter uit dat de componenten om technische redenen worden verlicht met rood licht. Deze informatie weerlegt onze oorspronkelijke overtuiging: we kunnen niet langer zeker zijn dat de rode schijn van de componenten een bewijs is van hun roodheid. Het interessante aan het fenomeen van het weerleggen van overtuigingen is dat de overtuigingen in eerste instantie epistemologisch verantwoord zijn, maar vervolgens hun rechtvaardiging verliezen door de nieuwe informatie. Zo zou het ook kunnen zijn met objectieve morele opvattingen: in eerste instantie zijn ze gerechtvaardigd maar verliezen deze rechtvaardiging daarna. We kunnen bijvoorbeeld met recht en reden beweren dat marteling objectief verwerpelijk is. Echter, zodra we de evolutionaire oorsprong ervan herkennen, moeten we de overtuiging over marteling opgeven – we weten niet meer of we dit beweerden omdat marteling *objectief* slecht is of dat we het daarom beweren, of dat wij dankzij evolutionaire redenen zo denken. Het vernieuwende van mijn onderzoek is dus dat ik het fenomeen van weerlegging in detail bestudeer en daardoor verschillende open vragen over de vermeende evolutionaire weerlegging beantwoord.

In **hoofdstuk 1** beschouw ik allereerst de huidige staat van onderzoek over de evolutionaire oorsprong van moraliteit. Het is belangrijk om onderscheid te maken tussen verschillende soorten evolutionaire verklaringen. Het is bijvoorbeeld nog niet duidelijk te bepalen of morele veronderstellingen zijn *geselecteerd* of slechts *neveneffecten* van evolutionaire selectieprocessen zijn. Het is echter aannemelijk dat interacties tussen natuurlijke en culturele selectie onze morele veronderstellingen hebben gevormd. Vervolgens situeer ik mijn onderzoek in een historische context, en beperk ik deze door te kijken naar misplaatste pogingen om morele aannames door evolutie moreel te rechtvaardigen. Denk hierbij aan het grove en valse idee dat de sterkere belangrijker is dan de zwakkere *omdat* evolutie het 'zo wil'. Ten slotte laat ik in detail zien waarom evolutionaire weerlegging met 'uitsterven bedreigd' lijkt: terwijl velen een meta-ethisch

probleem in de evolutie zien, zijn er genoeg critici die dit op basis van goede argumenten ontkennen. Ik beargumenteer dat de enige hoopvolle bestaande interpretaties van evolutionaire weerlegging de volgende zijn:

> Het *probleem van de evolutionaire betrouwbaarheid*: in het licht van de evolutie kan de betrouwbaarheid van onze morele aannames niet worden verklaard. Aannames waarvan de betrouwbaarheid niet kan worden verklaard, moeten worden opgegeven.
>
> Het *evolutionaire onenigheidsprobleem*: in het licht van de evolutie hadden we gemakkelijk morele veronderstellingen kunnen hebben die in tegenspraak zijn met onze huidige veronderstellingen. Het is niet mogelijk om deze (hypothetische) onenigheid op te lossen en daarom zijn onze morele veronderstellingen ongeldig.

Beide interpretaties komen opnieuw voorbij in de hoofdstukken 3, 4, 5, 6 en 7. Maar eerst gaat het om een beter begrip van het fenomeen van evolutionaire weerlegging zelf.

In **hoofdstuk 2** behandel ik de epistemologische voorwaarden voor het weerleggen van overtuigingen. Subjectivisten beweren dat iemands overtuiging ongeldig wordt zodra we het ongeldig achten. Stel ik zou er stellig van overtuigd zijn dat mijn collega een beruchte leugenaar is, dan zouden de aannames die ik heb gedaan op basis van mijn uitwisseling met deze collega ongeldig zijn, hoewel mijn verdenking zelf volkomen ongerechtvaardigd kan zijn volgens subjectivisten. De Objectivisten (die niet noodzakelijkerwijs overeenkomen met de morele objectivisten) beweren dat ontkrachters bepaalde regels moeten volgen en niet herleidbaar zijn tot hun waarneming. Ik laat zien dat verschillende varianten van subjectivisme over ontkrachters epistemologisch verwerpelijk zijn en dat daarom objectieve voorwaarden voor ontkrachters moeten worden gevonden.

Maar wat voor soort informatie kan onze morele overtuigingen ontkrachten? In **hoofdstuk 3** betoog ik dat evolutie slechts een essentiële rol *lijkt* te spelen bij het ontkrachten van morele overtuigingen. Het enige wat ertoe doet, is dat er feitelijk geen goede verklaring is waarom de inhoud van onze morele aannames adequaat zou moeten correleren met morele waarheid. Dus het fundamentele probleem is meer het *ontbreken* van een verklaring van de betrouwbaarheid van

onze morele veronderstellingen, dan het *bestaan* van een verklaring van de ontwrichtende factor van evolutie. Sterker nog, evolutie duidt slechts op één mogelijke factor waar onze morele veronderstellingen door beïnvloed worden. Andere factoren zouden bijvoorbeeld cultureel of psychologisch zijn. We weten nu dat er objectieve condities moeten zijn voor het ontkrachten van morele veronderstellingen, en dat evolutie alleen een essentiële rol *lijkt* te spelen. We weten nog niet, welke condities er voor het opgeven van aannames bestaan noch of evolutie deze mogelijk kan vervullen.

In **hoofdstuk 4** veronderstel ik dat evolutionaire weerlegging mogelijk is. Vervolgens laat ik zien dat de epistemologische rechtvaardiging van onze ooit verlaagde morele overtuigingen onmogelijk te herstellen is. Ik behandel en weerleg verschillende bestaande argumenten die vóór deze mogelijkheid pleiten. Uit mijn bespreking zal duidelijk worden dat een succesvolle evolutionaire weerlegging fataal zou zijn voor onze morele overtuigingen, wat de urgentie van het probleem alleen maar bevestigt.

In **hoofdstuk 5** weerleg ik het eerdergenoemde evolutionaire onenigheidsprobleem. Het evolutionaire onenigheidsprobleem beschrijft waarom evolutie onze morele overtuigingen ongeldig maakt: we hadden gemakkelijk andere morele overtuigingen kunnen hebben, wat verklaard wordt door evolutie. Een welbekend epistemologisch principe stelt dat men overtuigingen waarover onoplosbare onenigheid bestaat moet opgeven. Strikt genomen worden echter alleen onoplosbare meningsverschillen tussen *gelijken*, zoals experts met hetzelfde niveau van kennis over het onderwerp, als ontkrachtend beschouwd. Ik beargumenteer dat er geen totale onenigheid kan zijn over moraliteit tussen gelijken in termen van moraliteit. In geen geval worden onze morele overtuigingen helemaal ontkracht. Het evolutionaire onenigheidsprobleem als interpretatie van evolutionaire weerlegging faalt.

In **hoofdstuk 6** laat ik zien dat het probleem van onenigheid en het betrouwbaarheidsprobleem niet kunnen worden gecombineerd om evolutionair ontkrachten te verklaren. Het probleem van betrouwbaarheid vereist een verklaring van onze morele betrouwbaarheid. Stel dat er een duidelijke verklaring is, maar er is onenigheid over deze uitleg. Bijvoorbeeld zal je kunnen beweren dat overleven moreel goed is, en als dat zo is, dan is er een verklaring waarom wij

zowel overtuigt zijn, dat overleven goed is, maar ook waarom enkele van onze morele overtuigingen kloppen. Maar wat gebeurt er als een ander expert beweert dat overleven *niet* moreel goed is? In dat geval zou de verklaring epistemologisch onacceptabel zijn volgens het probleem van de onenigheid. Maar ik laat zien dat er geen rationele onenigheid kan bestaan als we de premissen van het betrouwbaarheidsprobleem hebben aanvaard en dus is een combinatie van de twee problemen uitgesloten.

Tot slot ga ik in **hoofdstuk 7** dieper in op het betrouwbaarheidsprobleem, dat na het uitsluitingsproces van de voorgaande hoofdstukken de beste verklaring lijkt te zijn voor evolutionair ontkrachten. Er kan echter gesteld worden dat wijdverspreide epistemologische veronderstellingen over de aard van moraliteit en de beste beschikbare benadering van de objectieve omstandigheden van ontkrachters, tot de conclusie leiden dat de betrouwbaarheid van onze morele veronderstellingen kan worden verklaard. Niet ondanks, maar juist *dankzij* de evolutie. Ik kan echter aantonen dat dit argument onjuist is. Gedeeltelijk zijn er ontkrachters die ons laten zien dat een veronderstelling geen kennis is, hoewel de ontkrachter niet impliceert dat de aanname ongerechtvaardigd of onnauwkeurig is. Dit argument is gebaseerd op de aanname dat wij naar kennis streven, en niet alleen naar waarheid. De kennisvoorwaarden gaan verder dan de vereisten van waarheid en rechtvaardiging. Bijvoorbeeld, je kunt door toeval een juiste veronderstelling maken, en door gelukkige omstandigheden kun je ook betrouwbaar zijn. Desondanks tellen de op deze manier gemaakte aannames niet als kennis. Maar als het ons doel is om kennis te bereiken en we leren dat onze aannames dit niet zijn, dan hebben we redenen om deze aannames op te geven. Deze mogelijkheid is, zo laat ik zien, de enige mogelijkheid voor evolutionair ontkrachten: evolutie moet aantonen dat we geen morele kennis hebben, en wanneer we ons daarvan bewust worden, worden onze morele veronderstellingen ongeldig. Dus er is hoop voor het 'overleven van weerleggen'.

Samengevat: Dit proefschrift illustreert het verband tussen evolutie en ontkrachters en dit biedt een epistemologisch argument dat de morele relevantie van de menselijke evolutie begrijpelijker maakt. In de loop van dit argument wordt het duidelijker wat ontkrachters werkelijk zijn, en wat de belangrijkste punten zijn voor het 'overleven van weerlegging'. Omdat morele psychologie altijd meer

duidelijkheid schept over de oorzakelijke oorsprong van onze morele veronderstellingen, wordt de vraag steeds dringender om te kijken in hoeverre dergelijke inzichten de geldigheid van onze bestaande morele veronderstellingen beïnvloeden. Mijn werk biedt duidelijke richtlijnen in dit verband om deze vaak geclaimde implicaties op hun geldigheid te controleren.

# Zusammenfassung

Zeigen evolutionäre Erklärungen der menschlichen Moral, dass wir das moralisch Gute nicht mehr vom moralisch Schlechten unterscheiden können? Diese Frage wird von vielen Moralwissenschaftlern (darunter Philosophen, Psychologen, und Biologen) mit Ja beantwortet. Sie behaupten, dass die Evolution menschliche Moralannahmen ohne Rücksicht auf deren Wahrheit beeinflusst und somit wie ein Störmittel auf unsere Moralannahmen wirkt. In meiner Dissertation stelle ich diese Behauptung gründlich auf die Probe zeige, warum bestehende Argumente in dieser Hinsicht fehlschlagen und was die erkenntnistheoretischen Bedingungen sind, unter denen die Evolution unsere moralischen Annahmen tatsächlich entkräften kann.

In dieser Zusammenfassung beschaue ich das Problem und meine Lösung aus der Vogelperspektive. Diejenigen Leser, die bereits mit der metaethisch-epistemologischen Problematik sowie den relevanten evolutionspsychologischen Fakten vertraut sind, verweise ich auf den detaillierteren Überblick, der in der Einleitung und Kapitel 1 dieser Arbeit gegeben wird. Um das Problem aus der Vogelperspektive zu verstehen muss man sich zwei Dinge klar machen: inwiefern Evolution menschliche Moralannahmen beeinflusst und warum dieser Einfluss wie ein Störmittel wirken könnte.

Betrachten wir zunächst unsere *Annahmen* über die Moral, also das, was wir für moralisch geboten oder erlaubt halten. Zum Beispiel kennen alle menschlichen Kulturen eine besondere Fürsorgepflicht der Eltern für ihre Kinder und so glaubt eine Vielzahl von Menschen, dass es für Eltern moralisch geboten ist, sich um ihre Kinder zu kümmern. Warum diese Einigkeit? Die Evolution liefert eine Antwort. Aus evolutionärer Sicht macht es Sinn für uns Menschen ein Fürsorgegefühl zu entwickeln und als moralische Regel zu akzeptieren, denn so hatten unsere Vorfahren größere Chancen gesunden Nachwuchs großzuziehen. Dass die meisten Menschen noch heute eine solche Fürsorgepflicht als moralisch geboten wahrnehmen, kann also durch die evolutionäre Psychologie erklärt werden.

Charles Darwin, der Begründer der Evolutionstheorie, hat daher schon 1870 spekuliert, dass unsere moralischen Vorstellungen grundsätzlich anders wären, wenn wir nur einen anderen evolutionären Ursprung hätten. Ein

Gedankenexperiment Darwins verdeutlicht die Relevanz des evolutionären Einflusses auf unsere Moralannahmen.

Stellen wir uns vor, unser evolutionärer Pfad hätte uns den Honigbienen ähnlich werden lassen – einer Art ‚Mensch-Biene'. In Honigbienenvölkern gibt es nur eine Mutter, die Königin. Der Rest der weiblichen Bienen sind sterile Arbeiterbienen. Als ‚Mensch-Bienen' würden also auch wir in einer Gesellschaft leben, in der es nur eine geschlechtsreife Frau gibt, die nur sterile Töchter leben lässt. In diesem Fall, mutmaßte Darwin, würden wir es als moralisch geboten erachten, fruchtbare weibliche Nachkommen zu töten. Schließlich, so Darwin, würde eine solche Regel die evolutionäre Fitness der ‚Mensch-Biene' steigern. Unsere Moralannahmen wären also deutlich anders als sie es heute tatsächlich sind und das nur, weil wir zufällig einen anderen evolutionären Pfad genommen haben. Laut dieser Hypothese sind zumindest einige der grundlegendsten menschlichen Moralannahmen von der Evolution beeinflusst und damit kontingent: eine andere evolutionäre Geschichte hätte zu anderen Moralannahmen geführt.

Betrachten wir nun die moralischen *Wahrheiten*. Laut dem *Moralischen Objektivismus* sind moralische Wahrheiten objektiv und damit unabhängig von unserer evolutionären Vergangenheit. Der moralische Objektivismus ist eine historisch betrachtete populäre metaethische Position, die nach einer Abkehr in den 70er und 80er nun wieder verbreiteter vertreten wird. Bemerkenswert ist, dass laut dem moralischen Objektivismus das moralisch Gebotene oder Erlaubte unabhängig davon ist, ob wir es tatsächlich als moralisch geboten erkennen. Es wäre also falsch seine fruchtbaren Töchter zu ermorden, selbst wenn wir es als ‚Mensch-Biene' als berechtigt ansehen würden. Die Sklaverei, um ein realitätsnäheres Beispiel zu nennen, wurde noch im 18. Jahrhundert für moralisch akzeptabel befunden, und dennoch war sie, laut dem Moralischen Objektivismus, auch damals schon falsch. Genauso verhält es sich mit der Verwerflichkeit der Folter und der weiblichen Beschneidung oder der Gutheit von Gerechtigkeit, um nur einige Beispiele zu nennen.

Das Problem ist nun folgendes. Auf der einen Seite beeinflusst die Evolution unsere tiefsten moralischen Annahmen, sodass wir tendenziell das für gut halten, was evolutionär ‚nützlich' ist. Auf der anderen Seite ist aber die Wahrheit unserer

moralischen Annahmen laut dem Moralischen Objektivismus völlig unabhängig davon, was evolutionär ‚nützlich' ist. Das ist in etwa so, als ob wir bei einer Segelfahrt ein bestimmtes Ziel ansteuern, aber unseren Kurs beliebig vom Wind bestimmen lassen. Es scheint unwahrscheinlich, dass wir so je ans Ziel gelangen. Wenn es sich mit der Evolution und der Moral ähnlich verhält wie mit dem Wind und unserem angesteuerten Ziel beim Segeln, können wir nicht mehr behaupten, dass unsere moralischen Annahmen wahr sind.

Die eingangs erwähnte Lösung vieler Moralwissenschaftler ist daher, dass wir entweder den moralischen Objektivismus aufgeben oder aber unsere moralischen Annahmen als entkräftet betrachten: wir wissen nicht, ob sie wahr oder gerechtfertigt sind. Selbstverständlich gibt es Alternativen zum moralischen Objektivismus, z.B. den moralischen Konstruktivismus. Die Möglichkeit den Moralischen Objektivismus aufzugeben lasse ich in dieser Arbeit außen vor, da es viele von evolutionären Betrachtungen unabhängige Argumente für oder gegen den Moralischen Objektivismus gibt, die ich nicht allesamt betrachten kann. Meine Frage ist daher, ob die Evolution ein Problem darstellt *wenn* wir überzeugt sind, dass es objektive moralische Wahrheiten gibt und nicht *ob* wir annehmen sollten, dass es objektive moralische Wahrheiten gibt. Wenn es aber keine *adäquate* Alternative zum moralischen Objektivismus gibt, dann scheint die Konsequenz ein moralischer Skeptizismus zu sein: in Anbetracht von evolutionären Erklärungen der Moral können wir nicht wissen, was moralisch gut oder schlecht ist.

Der Schluss von einer evolutionären Erklärung der Moral zum moralischen Skeptizismus ist aber verfrüht, selbst wenn wir annehmen, dass es keine adäquate Alternative zum moralischen Objektivismus gäbe. Mein Ansatz ist es, die erkenntnistheoretischen Regeln zu betrachten, die uns zum Aufgeben einer Annahme bewegen und untersuchen, ob es valide erkenntnistheoretische Regeln gibt, nach denen wir unsere moralischen Annahmen im Lichte von evolutionären Erklärungen der Moral aufzugeben haben.[1]

[1] Die Erkenntnistheorie beschäftigt sich unter anderem mit den Bedingungen von gerechtfertigten Annahmen und den Bedingungen von Wissen. Beispielsweiße ist eine durch Raten getroffene Annahme in den meisten Fällen nicht gerechtfertigt und stellt

Dabei behandele ich die Evolution als einen *Widerleger* unserer moralischen Annahmen. Widerlegende Informationen begegnen uns ständig. Stellen wir uns beispielsweise einen Fabrikbesuch vor, bei dem wir scheinbar rote Bauteile auf einem Förderband sehen. Es ist normal und aus erkenntnistheoretischer Sicht *berechtigt*, anzunehmen, dass die Bauteile rot sind. Wenn aber der Vorarbeiter erläutert, dass die Bauteile aus technischen Gründen mit rotem Licht bestrahlt werden, dann widerlegt diese Information unsere ursprüngliche Annahme: wir können nicht mehr sicher sein, ob der rote Schein der Bauteile für deren rot-sein spricht. Das Interessante am Phänomen des Widerlegens von Annahmen ist, dass die Annahmen erst erkenntnistheoretisch *gerechtfertigt* sind und dann durch eine neue Information an Rechtfertigung einbüßen.[2] Genauso könnte es sich mit objektiven moralischen Annahmen verhalten: erst sind sie gerechtfertigt, und wir können mit Fug und Recht behaupten, dass wir eine moralische Pflicht haben für unseren Nachwuchs zu sorgen. Doch sobald wir deren evolutionären Ursprung erkennen, sollten wir sie aufgeben. Meine Untersuchung ist neuartig, indem sie das Phänomen des Entkräftens genau unter die Lupe nimmt und verschiedene offene Fragen zum angeblichen evolutionären Entkräften beantwortet. Gibt es ein valides Bild vom entkräftenden Widerlegen, welches erklären kann, warum und wie die Evolution unsere Moralannahmen entkräftet?

In **Kapitel 1** zeige ich zunächst den aktuellen Stand der Forschung zum evolutionären Ursprung der Moral auf. Dabei gilt es verschiedene Arten der ‚evolutionären Erklärung' zu unterscheiden und es ist noch nicht sicher zu sagen, ob moralische Annahmen *selektiert* wurden oder bloße *Seiteneffekte* von evolutionären Selektionsprozessen sind. Es ist aber plausibel, dass Wechselwirkungen zwischen natürlicher und kultureller Selektion unsere moralischen Annahmen entscheidend geprägt haben. Dann situiere ich meinen Forschungsansatz historisch und grenze ihn ab von fehlgeleiteten Versuchen, die

auch kein Wissen dar. Eine Annahme, die auf verlässlichen Beweisen basiert, ist dagegen mit größerer Wahrscheinlichkeit auch gerechtfertigt. Insofern als die Erkenntnistheorie uns Regeln zum rationalen, validen Bilden und Verwerfen von Annahmen bietet, sollte sich feststellen lassen, ob es valide Regeln gibt, die besagen, dass wir unsere moralischen Annahmen im Lichte der Evolution zu verwerfen haben.

[2] Nicht jeder würde also im Beispiel mit der Fabrik tatsächlich die Annahme aufgeben, dass die Bauteile rot sind, aber jeder *sollte* es tun. Die Erkenntnistheorie liefert teilweise normative Aussagen.

Evolution zur Rechtfertigung moralischer Aussagen zu gebrauchen. Man denke hier an die krude und falsche Idee, dass der Stärkere mehr gilt als der Schwächere, *weil* es die Evolution so will, dass der Stärkere überlebt. Zuletzt zeige ich im Detail, warum das evolutionäre Widerlegen ‚vom Aussterben bedroht' zu sein scheint: zwar sehen viele ein metaethisches Problem in der Evolution, aber es gibt genug Kritiker, die keine gute erkenntnistheoretische Grundlage für diese Annahme sehen. Die einzig hoffnungsvollen bestehenden Interpretationen des evolutionären Widerlegens, so zeige ich, sind die folgenden:

> Das *Evolutionäre Verlässlichkeitsproblem*: Im Lichte der Evolution ist die Verlässlichkeit unserer moralischen Annahmen nicht zu erklären. Annahmen, deren Verlässlichkeit nicht erklärt werden kann, müssen aufgegeben werden.
>
> Das *Evolutionäre Uneinigkeitsproblem*: Im Lichte der Evolution zeigt sich, dass wir leicht moralische Annahmen hätten treffen können, die unseren jetzigen Annahmen widersprechen. Es ist nicht möglich, diese (hypothetische) Uneinigkeit aufzulösen und daher sind unsere moralischen Annahmen entkräftet.

Beide Interpretationen begegnen uns wieder in den Kapiteln 3, 4, 5, 6 und 7. Zunächst geht es jedoch darum, das Phänomen des evolutionären Widerlegens selbst besser zu durchleuchten.

In **Kapitel 2** betrachte ich die erkenntnistheoretischen Bedingungen für das Entkräften von Annahmen. Subjektivisten behaupten, dass eine Annahme entkräftet ist, sobald wir diese für entkräftet erachten. Beispielsweiße könnte ich der felsenfesten Überzeugung sein, dass mein Kollege ein notorischer Lügner ist und diese Überzeugung würde die Annahmen entkräften, die ich auf Basis der Aussagen meines Kollegen getroffen habe, obwohl die Überzeugung selbst vollkommen ungerechtfertigt sein kann. Objektivisten (die nicht zwingend mit den *Moralischen* Objektivisten gleichzusetzen sind) behaupten dagegen, dass Entkräfter bestimmten Regeln zu folgen haben und nicht auf deren Wahrnehmung reduzierbar sind. Ich zeige, dass verschiedene Varianten des Subjektivismus über Entkräfter erkenntnistheoretisch untragbar sind, und dass daher objektive Bedingungen für Entkräfter gefunden werden müssen.

Was aber wäre die entkräftende Information im Falle der moralischen Annahmen? In **Kapitel 3** argumentiere ich, dass die Evolution nur scheinbar eine essentielle Rolle im Entkräften von moralischen Annahmen spielt. Tatsächlich geht es lediglich darum, dass es keine gute Erklärung dafür gibt, warum der Inhalt unserer moralischen Annahmen adäquat mit der moralischen Wahrheit korrelieren sollte. Das grundlegende Problem ist also eher der *Mangel* einer Erklärung für die Verlässlichkeit unserer Moralannahmen als das *Bestehen* einer Erklärung durch den Störfaktor Evolution. Die Evolution zeigt lediglich einen möglichen Einflussfaktor auf unsere moralischen Annahmen auf. Alternative Erklärungen wären beispielsweise kultureller oder psychologischer Natur. Wir wissen nun, dass es objektive Bedingungen für Entkräfter geben muss und dass die Evolution nur eine scheinbar essentielle Rolle einnimmt, aber noch nicht, welche Bedingungen dies sind und ob sie durch die Evolution erfüllt werden.

In **Kapitel 4** nehme ich an, dass evolutionäres Widerlegen möglich ist und zeige dann, dass es keine Möglichkeit gäbe die erkenntnistheoretische Rechtfertigung unserer einmal entkräfteten moralischen Annahmen wieder herzustellen. Dabei behandle und widerlege ich verschiedene aktuelle Vorschläge, wie das doch möglich wäre. Damit ist klar, dass ein erfolgreiches evolutionäres Widerlegen fatal wäre für unsere moralischen Annahmen, was die Dringlichkeit des Problems nur erhöht.

In **Kapitel 5** gehe ich auf das zuvor genannte evolutionäre Uneinigkeitsproblem ein und widerlege es. Das evolutionäre Uneinigkeitsproblem beschreibt, warum die Evolution unsere moralischen Annahmen entkräftet: wir hätten leicht andere moralische Annahmen treffen können (was durch die Evolution erklärt wird) und ein vielbeachtetes erkenntnistheoretisches Prinzip besagt, dass man unauflösbar umstrittene Annahmen aufgeben muss. Streng genommen gelten allerdings nur logisch unauflösbare Uneinigkeiten zwischen *Gleichgestellten* (etwa Experten mit gleichem Wissensstand zum Thema) als entkräftend und ich zeige, dass es über die Moral entweder keine totale Uneinigkeit geben kann oder, wenn doch, dann nicht zwischen Gleichgestellten in Bezug auf die Moral. Das bedeutet, dass wir bei einer evolutionär implizierten Uneinigkeit bezüglich unserer fundamentalen Moralannahmen entweder eine der beiden Streitparteien einen Fehler gemacht hat, oder aber nicht als Gleichgestellt

betrachtet werden muss. In keinem Fall sind unsere moralischen Annahmen allesamt entkräftet. Das evolutionäre Uneinigkeitsproblem als Interpretation des evolutionären Widerlegens schlägt fehl.

In **Kapitel 6** zeige ich, dass sich das Uneinigkeitsproblem und das Verlässlichkeitsproblem nicht miteinander verbinden lassen, um evolutionäres Entkräften zu erklären. Das Verlässlichkeitsproblem verlangt nach einer Erklärung unserer moralischen Verlässlichkeit. Nehmen wir an, es gäbe eine solche, aber es bestünde Uneinigkeit über diese Erklärung. Beispielsweiße könnte man annehmen, dass die elterliche Fürsorge moralisch gut ist und sodann (evolutionär) erklären warum wir entsprechende Annahmen formen und warum diese zugleich wahr sind. Aber was passiert wenn ein anderer Experte behauptet, dass die Fürsorgepflicht nicht moralisch gut ist? In diesem Fall wäre die Erklärung, gemäß des Uneinigkeitsproblems, erkenntnistheoretisch nicht akzeptabel. Allerdings zeige ich, dass es keine rationale Uneinigkeit geben kann, wenn wir die Prämissen des Verlässlichkeitsproblems einmal akzeptiert haben, und so ist eine Verkettung der beiden Probleme ausgeschlossen.

In **Kapitel 7** gehe ich schließlich genauer auf das Verlässlichkeitsproblem ein, welches nach dem Ausschlussverfahren der vorangegangenen Kapitel die beste Erklärung für evolutionäres Entkräften zu sein scheint. Allerdings lässt sich argumentieren, dass weitverbreitete erkenntnistheoretische Annahmen über die Natur der Moral und den besten verfügbaren Ansatz über die objektiven Bedingungen von Entkräftern dazu führen, dass die Verlässlichkeit unserer moralischen Annahmen erklärbar ist, nicht *trotz* sondern gerade *wegen* der Evolution. Ich kann allerdings zeigen, dass diese Argumentation falsch ist. Teilweise gibt es Entkräfter, die uns zeigen, dass eine Annahme kein Wissen darstellt, obwohl der Entkräfter nicht impliziert, dass die Annahme ungerechtfertigt oder inakkurat ist. Das ist so, wenn man annimmt, dass wir Annahmen treffen, um Wissen zu erlangen. Die Bedingungen von Wissen gehen aber über die Voraussetzungen von Wahrheit und Rechtfertigung hinaus. Beispielsweise kann man durch Glück eine wahre Annahme treffen, und durch glückliche Umstände sogar sehr verlässlich dabei sein. Dennoch zählen die so getroffenen Annahmen nicht als Wissen. Wenn aber unsere Annahmen Wissen darstellen sollen und wir erfahren, dass sie das nicht tun, dann haben wir Gründe

sie aufzugeben. Diese Möglichkeit ist, so zeige ich, die einzige Möglichkeit für evolutionäres Entkräften: evolutionäre Erklärungen der Moral müssen zeigen, dass wir kein moralisches Wissen haben, und wenn wir dieser Implikation gewahr werden, sind unsere moralischen Annahmen entkräftet. Es besteht also Hoffnung für das ‚Überleben von Widerlegen'.

Zusammengefasst verdeutlicht meine Arbeit den Zusammenhang von Evolution und Entkräftern und bietet ein erkenntnistheoretisches Argument, welches die moralische Relevanz der menschlichen Evolution besser begreifbar macht. Im Zuge dieser Argumentation wird deutlicher, was Entkräfter eigentlich sind und was die kritischen Punkte sind für das ‚Überleben von Widerlegen'. Da beispielsweise die Moralpsychologie stets mehr Deutlichkeit über den kausalen Ursprung unserer Moralannahmen schafft, wird die Frage stets dringlicher, inwiefern solche Einsichten die Gültigkeit unserer bestehenden Moralannahmen beeinflusst. Meine Arbeit bietet in dieser Hinsicht klare Richtlinien, um diese oft behaupteten Implikationen auf ihre Gültigkeit zu prüfen.

## Curriculum Vitae

Michael Klenk was born on November 10th 1989 in Heilbronn-Neckargartach, Germany. In 2011 he graduated from Cooperative State University Baden Wuerttemberg in Stuttgart with a bachelor's degree in business administration. The degree was financed by Andretta, an international trade and logistics company, where he worked as a trader during the duration of his degree. After his first contact with academic philosophy during a semester abroad at Bond University, Australia, he obtained a master's degree in philosophy at University College London, where he graduated in 2012 with merit. From 2012 to 2014 he worked as a management consultant for Atos Consulting in Munich. During that time, he enrolled in an extramural bachelor's degree in psychology at the University of Hagen, which he will complete in 2018. In October 2014, he started as a PhD candidate at Utrecht University in the NWO-funded research programme 'Evolutionary Ethics? The (Meta-)Ethical Implications of Evolutionary Explanations of Morality' under the supervision of prof.dr.mr. Herman Philipse. During his time as a PhD candidate, Michael was a visiting fellow at Oxford University, Columbia University, and Harvard University. He has taught courses on metaethics, logic, and metaphysics, chaired the PhD Council of the Dutch Research School of Philosophy for two years, and (co-)organised several academic conferences. His main reseach interests are epistemology, metaethics, and moral psychology, and he presented on these topics at more than twenty conferences in places such as Athens, Bogota, Cambridge, MA, Munich, St. Andrews, and Oxford.

This page intentionally contains only this sentence.

# Quaestiones Infinitae

## PUBLICATIONS OF THE DEPARTMENT OF PHILOSOPHY AND RELIGIOUS STUDIES

VOLUME 21. D. VAN DALEN, *Torens en Fundamenten* (valedictory lecture), 1997.

VOLUME 22. J.A. BERGSTRA, W.J. FOKKINK, W.M.T. MENNEN, S.F.M. VAN VLIJMEN, *Spoorweglogica via EURIS*, 1997.

VOLUME 23. I.M. CROESE, *Simplicius on Continuous and Instantaneous Change* (dissertation), 1998.

VOLUME 24. M.J. HOLLENBERG, *Logic and Bisimulation* (dissertation), 1998.

VOLUME 25. C.H. LEIJENHORST, *Hobbes and the Aristotelians* (dissertation), 1998.

VOLUME 26. S.F.M. VAN VLIJMEN, *Algebraic Specification in Action* (dissertation), 1998.

VOLUME 27. M.F. VERWEIJ, *Preventive Medicine Between Obligation and Aspiration* (dissertation), 1998.

VOLUME 28. J.A. BERGSTRA, S.F.M. VAN VLIJMEN, *Theoretische Software-Engineering: kenmerken, faseringen en classificaties*, 1998.

VOLUME 29. A.G. WOUTERS, *Explanation Without A Cause* (dissertation), 1999.

VOLUME 30. M.M.S.K. SIE, *Responsibility, Blameworthy Action & Normative Disagreements* (dissertation), 1999.

VOLUME 31. M.S.P.R. VAN ATTEN, *Phenomenology of choice sequences* (dissertation), 1999.

VOLUME 32. V.N. STEBLETSOVA, *Algebras, Relations and Geometries (an equational perspective)* (dissertation), 2000.

VOLUME 33. A. VISSER, *Het Tekst Continuüm* (inaugural lecture), 2000.

VOLUME 34. H. ISHIGURO, *Can we speak about what cannot be said?* (public lecture), 2000.

VOLUME 35. W. HAAS, *Haltlosigkeit; Zwischen Sprache und Erfahrung* (dissertation), 2001.

VOLUME 36. R. POLI, *ALWIS: Ontology for knowledge engineers* (dissertation), 2001.

VOLUME 37. J. MANSFELD, *Platonische Briefschrijverij* (valedictory lecture), 2001.

VOLUME 37A. E.J. BOS, *The Correspondence between Descartes and Henricus Regius* (dissertation), 2002.

VOLUME 38. M. VAN OTEGEM, *A Bibliography of the Works of Descartes (1637-1704)* (dissertation), 2002.

VOLUME 39. B.E.K.J. GOOSSENS, *Edmund Husserl: Einleitung in die Philosophie: Vorlesungen 1922/23* (dissertation), 2003.

VOLUME 40. H.J.M. BROEKHUIJSE, *Het einde van de sociaaldemocratie* (dissertation), 2002.

VOLUME 41. P. RAVALLI, *Husserls Phänomenologie der Intersubjektivität in den Göttinger Jahren: Eine kritisch-historische Darstellung* (dissertation), 2003.

VOLUME 42. B. ALMOND, *The Midas Touch: Ethics, Science and our Human Future* (inaugural lecture), 2003.

VOLUME 43. M. DÜWELL, *Morele kennis: over de mogelijkheden van toegepaste ethiek* (inaugural lecture), 2003.

VOLUME 44. R.D.A. HENDRIKS, *Metamathematics in Coq* (dissertation), 2003.

VOLUME 45. TH. VERBEEK, E.J. BOS, J.M.M. VAN DE VEN, *The Correspondence of René Descartes: 1643*, 2003.

VOLUME 46. J.J.C. KUIPER, *Ideas and Explorations: Brouwer's Road to Intuitionism* (dissertation), 2004.

VOLUME 47. C.M. BEKKER, *Rechtvaardigheid, Onpartijdigheid, Gender en Sociale Diversiteit; Feministische filosofen over recht doen aan vrouwen en hun onderlinge verschillen* (dissertation), 2004.

VOLUME 48. A.A. LONG, *Epictetus on understanding and managing emotions* (public lecture), 2004.

VOLUME 49. J.J. JOOSTEN, *Interpretability formalized* (dissertation), 2004.

VOLUME 50. J.G. SIJMONS, *Phänomenologie und Idealismus: Analyse der Struktur und Methode der Philosophie Rudolf Steiners* (dissertation), 2005.

VOLUME 51. J.H. HOOGSTAD, *Time tracks* (dissertation), 2005.

VOLUME 52. M.A. VAN DEN HOVEN, *A Claim for Reasonable Morality* (dissertation), 2006.

VOLUME 53. C. VERMEULEN, *René Descartes, Specimina philosophiae: Introduction and Critical Edition* (dissertation), 2007.

VOLUME 54. R.G. MILLIKAN, *Learning Language without having a theory of mind* (inaugural lecture), 2007.

VOLUME 55. R.J.G. CLAASSEN, *The Market's Place in the Provision of Goods* (dissertation), 2008.

VOLUME 56. H.J.S. BRUGGINK, *Equivalence of Reductions in Higher-Order Rewriting* (dissertation), 2008.

VOLUME 57. A. KALIS, *Failures of agency* (dissertation), 2009.

VOLUME 58. S. GRAUMANN, *Assistierte Freiheit* (dissertation), 2009.

VOLUME 59. M. AALDERINK, *Philosophy, Scientific Knowledge, and Concept Formation in Geulincx and Descartes* (dissertation), 2010.

VOLUME 60. I.M. CONRADIE, *Seneca in his cultural and literary context: Selected moral*

*letters on the body* (dissertation), 2010.

VOLUME 61. C. VAN SIJL, *Stoic Philosophy and the Exegesis of Myth* (dissertation), 2010.

VOLUME 62. J.M.I.M. LEO, *The Logical Structure of Relations* (dissertation), 2010.

VOLUME 63 M.S.A. VAN HOUTE, *Seneca's theology in its philosophical context* (dissertation), 2010.

VOLUME 64 F.A. BAKKER, *Three Studies in Epicurean Cosmology* (dissertation), 2010.

VOLUME 65 T. FOSSEN, *Political legitimacy and the pragmatic turn* (dissertation), 2011.

VOLUME 66 T. VISAK, *Killing happy animals. Explorations in utilitarian ethics.* (dissertation), 2011.

VOLUME 67 A. JOOSSE, *Why we need others: Platonic and Stoic models of friendship and self-understanding* (dissertation), 2011.

VOLUME 68 N. M. NIJSINGH, *Expanding newborn screening programmes and strengthening informed consent* (dissertation), 2012.

VOLUME 69 R. PEELS, *Believing Responsibly: Intellectual Obligations and Doxastic Excuses* (dissertation), 2012.

VOLUME 70 S. LUTZ, *Criteria of Empirical Significance* (dissertation), 2012

VOLUME 70A G.H. BOS, *Agential Self-consciousness, beyond conscious agency* (dissertation), 2013.

VOLUME 71 F.E. KALDEWAIJ, *The animal in morality: Justifying duties to animals in Kantian moral philosophy* (dissertation), 2013.

VOLUME 72 R.O. BUNING, *Henricus Reneri (1593-1639): Descartes' Quartermaster in Aristotelian Territory* (dissertation), 2013.

VOLUME 73 I.S. LÖWISCH, *Genealogy Composition in Response to Trauma: Gender and Memory in 1 Chronicles 1-9 and the Documentary Film 'My Life Part 2'* (dissertation), 2013.

VOLUME 74 A. EL KHAIRAT, *Contesting Boundaries: Satire in Contemporary Morocco* (dissertation), 2013.

VOLUME 75 A. KROM, *Not to be sneezed at. On the possibility of justifying infectious disease control by appealing to a mid-level harm principle* (dissertation), 2014.

VOLUME 76 Z. PALL, *Salafism in Lebanon: local and transnational resources* (dissertation), 2014.

VOLUME 77 D. WAHID, *Nurturing the Salafi Manhaj: A Study of Salafi Pesantrens in Contemporary Indonesia* (dissertation), 2014.

VOLUME 78 B.W.P VAN DEN BERG, *Speelruimte voor dialoog en verbeelding. Basisschoolleerlingen maken kennis met religieuze verhalen* (dissertation), 2014.

VOLUME 79 J.T. BERGHUIJS, *New Spirituality and Social Engagement* (dissertation), 2014.

VOLUME 80 A. WETTER, *Judging By Her. Reconfiguring Israel in Ruth, Esther and Judith* (dissertation), 2014.

VOLUME 81 J.M. MULDER, *Conceptual Realism. The Structure of Metaphysical Thought* (dissertation), 2014.

VOLUME 82 L.W.C. VAN LIT, *Eschatology and the World of Image in Suhrawardī and His Commentators* (dissertation), 2014.

VOLUME 83 P.L. LAMBERTZ, *Divisive matters. Aesthetic difference and authority in a Congolese spiritual movement 'from Japan'* (dissertation), 2015.

VOLUME 84 J.P. GOUDSMIT, *Intuitionistic Rules: Admissible Rules of Intermediate Logics* (dissertation), 2015.

VOLUME 85 E.T. FEIKEMA, *Still not at Ease: Corruption and Conflict of Interest in Hybrid Political Orders* (dissertation), 2015.

VOLUME 86 N. VAN MILTENBURG, *Freedom in Action* (dissertation), 2015.

VOLUME 86A P. COPPENS, *Seeing God in This world and the Otherworld: Crossing Boundaries in Sufi Commentaries on the Qurʾān* (dissertation), 2015.

VOLUME 87 D.H.J. JETHRO, *Aesthetics of Power: Heritage Formation and the Senses in Post-apartheid South Africa* (dissertation), 2015.

VOLUME 88 C.E. HARNACKE, *From Human Nature to Moral Judgement: Reframing Debates about Disability and Enhancement* (dissertation), 2015.

VOLUME 89 X. WANG, *Human Rights and Internet Access: A Philosophical Investigation* (dissertation), 2016.

VOLUME 90 R. VAN BROEKHOVEN, *De Bewakers Bewaakt: Journalistiek en leiderschap in een gemediatiseerde democratie* (dissertation), 2016.

VOLUME 91 A. SCHLATMANN, *Shi'i Muslim youth in the Netherlands: Negotiating Shi'i fatwas and rituals in the Dutch context* (dissertation), 2016.

VOLUME 92 M.L. VAN WIJNGAARDEN, *Schitterende getuigen. Nederlands luthers avondmaalsgerei als indenteitsdrager van een godsdienstige minderheid* (dissertation), 2016.

VOLUME 93 S. COENRADIE, *Vicarious substitution in the literary work of Shūsaku Endō. On fools, animals, objects and doubles* (dissertation), 2016.

VOLUME 94 J. RAJAIAH, *Dalit Humanization. A quest based on M.M. Thomas' theology of salvation and humanization* (dissertation), 2016.

VOLUME 95 D.L.A. OMETTO, *Freedom & Self-knowledge* (dissertation), 2016.

VOLUME 96 Y. YALDIZ, *The Afterlife in Mind: Piety and Renunciatory Practice in the 2nd/8th- and early 3rd/9th-Century Books of Renunciation (Kutub al-Zuhd)* (dissertation), 2016.

VOLUME 97 M.F. BYSKOV, *Between experts and locals. Towards an inclusive framework for a development agenda* (dissertation), 2016.

VOLUME 98 A. RUMBERG, *Transitions toward a Semantics for Real Possibility* (dissertation), 2016.

VOLUME 99 S. DE MAAGT, *Constructing Morality: Transcendental Arguments in Ethics* (dissertation), 2017.

VOLUME 100 S. BINDER, *Total Atheism* (dissertation), 2017.

VOLUME 101 T. GIESBERS, *The Wall or the Door: German Realism around 1800,* (dissertation), 2017.

VOLUME 102 P. SPERBER, *Kantian Psychologism* (dissertation), 2017.

VOLUME 103 J.M. HAMER, *Agential Pluralism: A Philosophy of Fundamental Rights* (dissertation), 2017.

VOLUME 104 M. IBRAHIM, *Sensational Piety: Practices of Mediation in Christ Embassy and Nasfat* (dissertation), 2017.

VOLUME 105 R.A.J. MEES, *Sustainable Action, Perspectives for Individuals, Institutions, and Humanity* (dissertation), 2017.

VOLUME 106 A.A.J. POST, *The Journey of a Taymiyyan Sufi: Sufism Through the Eyes of ʿImād al-Dīn Aḥmad al-Wāsiṭī (d. 711/1311)* (dissertation), 2017.

VOLUME 107 F.A. FOGUE KUATE, *Médias et coexistence entre Musulmans et Chrétiens au Nord-Cameroun: de la période coloniale Française au début du XXIème siècle* (dissertation), 2017.

VOLUME 108 J. KROESBERGEN-KAMPS, *Speaking of Satan in Zambia. The persuasiveness of contemporary narratives about Satanism* (dissertation), 2018.

VOLUME 109 F. TENG, *Moral Responsibilities to Future Generations. A Comparative Study on Human Rights Theory and Confucianism* (dissertation), 2018.

VOLUME 110 H.W.A. DUIJF, *Let's Do It! Collective Responsibility, Joint Action, and Participation* (dissertation), 2018.

VOLUME 111 R.A. CALVERT, *Pilgrims in the port. Migrant Christian communities in*

*Rotterdam* (dissertation), 2018.

VOLUME 112 W.P.J.L. VAN SAANE, *Protestant Mission Partnerships: The Concept of Partnership in the History of the Netherlands Missionary Council in the Twentieth Century* (dissertation), 2018.

VOLUME 113 D.K. DÜRING, *Of Dragons and Owls. Rethinking Chinese and Western narratives of modernity* (dissertation), 2018.

VOLUME 114 H. ARENTSHORST, *Perspectives on freedom. Normative and political views on the preconditions of a free democratic society* (dissertation), 2018.

VOLUME 115 M.B.O.T. KLENK, *Survival of Defeat – Evolution, Moral Objectivity, and Undercutting* (dissertation), 2018.

Quaestiones Infinitae