This is a preprint. Please cite the published version:

Koon, J. (forthcoming). "Recalibrating Evolutionary Debunking," *Philosophy and Phenomenological Research*.

Recalibrating Evolutionary Debunking

Justis Koon

1. Introduction

Moral realism is the view that moral language should be interpreted literally, that moral claims express beliefs, that some moral claims are true, and that moral truths are mind- and language-independent.[1] One prominent challenge to moral realism comes from evolutionary debunking arguments.[2] Evolutionary debunking arguments are motivated by recent empirical work on the evolution of morality, work which suggests that our moral instincts were selected chiefly to facilitate cooperation in our prehistoric ancestors.[3] But, if our moral instincts were selected for the practical function of promoting cooperation, it's puzzling how we could have ended up with epistemic access to the realist's realm of mind-independent moral facts. There appears to be a serious disconnect, in other words, between the evolved function of human moral cognition and the epistemic powers the realist

1 There are a few meta-ethical views, like the thin forms of reductive naturalism defended by Copp (2008) and Sterelny and Fraser (2017), where it's unclear whether we should classify them as types of realism or not. I won't be addressing these sorts of views in this paper.

2 For compelling presentations of these arguments, see Joyce (2006), Street (2006), Horn (2017), and Lutz (2018).

3 For a variety of perspectives on the evolution of morality, see Alexander (1987), Richerson and Boyd (2005), Hauser (2006), Joyce (2006), Bowles and Gintis (2011), Kitcher (2011), Baumard et al. (2012), Boehm (2012), Tomasello (2016), and Sterelny (2021). Although the details of these accounts differ, my reading of the literature as a whole is that there's an emerging consensus that moral cognition is largely an adaptation for facilitating cooperation. Claims I make in later sections of the paper will take this point for granted.

attributes to it.

According to the debunker, this disconnect is grounds for thinking that, if the realist's metaphysical claims are true, our capacity for moral judgment is likely to be unreliable – after all, it seems like it would be a remarkable coincidence if moral instincts selected to make our ancestors into better cooperators also turned out to be a reliable guide to the mind-independent moral facts. And, the debunker claims, if we have grounds for thinking that our capacity for moral judgment is likely to be unreliable, it follows that none of our moral beliefs are justified, just as we would not be justified in trusting any of the readings taken by a scientific instrument if we had good reason to suspect that it was unreliable. It would be one thing if we could corroborate our moral beliefs using some other cognitive faculty whose epistemic credentials are above reproach, but (the debunker will argue) this is impossible, because ethics is autonomous from other domains of inquiry. It appears, then, that moral realists who are familiar with these facts about our evolutionary history are saddled with an unappealing form of moral skepticism, one on which we have no justified moral beliefs, no moral knowledge, not even a clear picture of what morality is about. Instead of accepting this unhappy marriage of realism and skepticism, the debunker concludes, we would be better off rejecting one or more of the realist's package of linguistic and metaphysical theses, and choosing to become anti-realists instead.

Moral realists have raised a large number of objections to evolutionary debunking arguments, challenging them on both empirical and philosophical grounds.[4] My aim in this paper is to respond, on behalf of the debunker, to a pair of these objections that seem particularly pressing. The first objection is that debunking arguments are self-undermining: they cannot be formulated without invoking epistemic principles, the objection claims, but epistemic principles are just as vulnerable to evolutionary debunking as our moral beliefs, making it impossible to defend the debunking argument's

4 Most of these objections can be sorted into three categories: third-factor responses (Enoch 2010; Wielenberg 2010; Brosnan 2011; Skarsaune 2011; Berker 2014); objections to the debunking argument's epistemic principle (White 2010; Shafer-Landau 2012; Bogardus 2016; Clarke-Doane 2016; Sinclair 2018; Clarke-Doane and Baras 2021), and objections to the debunking argument's empirical premise (Machery and Mallon 2010; FitzPatrick 2015; Levy and Levy 2020).

premises while simultaneously accepting its conclusion.[5] In Section 3, I argue that this objection suffers from several problems: it's empirically unsupported, it has the implausible consequence that the justification for our moral beliefs can never be undermined by evidence that our capacity for normative reasoning is globally impaired, and it only succeeds in locating an inconsistency in the debunker's belief set in the unlikely event that the debunker is also an epistemic realist.

The second objection, which comes to us from Katia Vavova (2014; 2021), claims that, because debunking arguments attempt to show that we're unreliable in the moral domain without making any assumptions about the nature of morality, they're doomed to fail. In Section 4, I contend that the epistemic rule this objection relies on, that we cannot know our judgments in domain *D* are unreliable without making some assumptions about what *D* is like, is vulnerable to numerous counter-examples. Indeed, we'll see that just about any type of higher-order evidence can serve as a counter-example to the rule, including the etiological higher-order evidence the debunker claims has been unearthed by scientists and philosophers studying the evolution of morality. Of particular interest are cases where we learn that the causal history of a belief is random or chancy, like a belief chosen by roulette; I will argue that the evolutionary forces that shaped human moral cognition share important formal features with roulette, supporting the debunker's contention that our moral instincts are likely to be unreliable guides to the moral truth.

If I'm right, and both objections are flawed, this substantially narrows the range of options available to the realist for resisting evolutionary debunking arguments. These are, in a sense, the two most ambitious objections that have been advanced in defense of realism: the self-undermining objection claims that debunking arguments are formally defective, while Vavova's objection purports to show that the task debunkers have set for themselves is impossible. Hence, if both objections can be disarmed, this would represent a major setback for the realist, and a major step towards the debunker's ultimate goal of showing that moral realism leads ineluctably to moral skepticism.[6]

5 Berker (2014), Vavova (2014), Korman (2019), and Levy and Levy (2020) all discuss different versions of this objection. Street (2009a) responds to a related objection.

6 While my focus will be on evolutionary debunking in what follows, the arguments I develop bear on a number of important, related issues as well – on how we should respond to evidence suggesting that our capacity for normative judgment is globally impaired, on whether it's possible to establish that we're unreliable in a given

A key theme of this paper, which will emerge more fully in what follows, is that questions about higher-order evidence (and especially the calibrationist view of higher-order evidence) are central to the debate over the success or failure of evolutionary debunking arguments. We've made enormous strides in our understanding of higher-order evidence over the past decade, which means that many of the first wave of papers on evolutionary debunking, which predate this literature, will need to be revisited, or recalibrated, in light of recent developments in epistemology. In particular, a number of early objections to evolutionary debunking arguments, which may have seemed convincing at the time, fail to hold up to scrutiny once similar cases involving other types of higher-order evidence are brought into the picture. Evolutionary debunking arguments are fundamentally concerned with our abilities as knowers, with our capacity for moral judgment, and it's impossible to meaningfully evaluate or respond to these arguments without first understanding the epistemic significance of higher-order evidence on a more general level.

Before getting into any of this, though, it will be helpful to go over some more background on debunking arguments, to better understand how they work.

2. Evolutionary Debunking Arguments

Here is one way of formalizing the debunking argument outlined in the introduction:

Empirical Premise: We have good reason to think that our moral instincts were selected principally for functions other than producing true moral beliefs.

Etiological Principle: If moral realism is true, and if we have good reason to think that our moral instincts were selected principally for functions other than producing true moral beliefs, then, unless we're able to corroborate our moral beliefs through the use of some other faculty whose etiology is not

domain without making any first-order assumptions about that domain, on the different ways that information about a belief's etiology can undermine its justification, and on the role that chance and contingency played in the evolution of morality.

subject to similar doubts, the balance of independent evidence suggests that our moral beliefs are unreliable.

Autonomy Clause: We're not able to corroborate our moral beliefs through the use of some other faculty whose etiology is not subject to similar doubts.

Epistemic Principle: If the balance of independent evidence suggests that our moral beliefs are unreliable, the justification for our moral beliefs is defeated.

Conclusion: If moral realism is true, the justification for our moral beliefs is defeated.[7]

A few notes are in order. First, although the argument targets the justification for our moral beliefs, nothing hangs on this, and it could easily be modified to target the warrant for our moral beliefs or their status as knowledge instead. Second, while the argument refers in several places to moral instincts, this is solely for ease of exposition, and is not intended to carry a commitment to any particular view on the moral nativism debate. For the purposes of this paper, we can understand "moral instincts" to refer to the evolved components of human moral cognition, whatever they might turn out to be.[8] Although there's a great deal of dispute over how much of our moral cognition is innate, and how distinct it is from other types of normative cognition, it's now widely accepted that we come into this world with a broad array of innate moral emotions, dispositions, and concepts, and that these were likely shaped by natural selection over the millennia in order to facilitate cooperation in our hunter-gatherer ancestors.[9]

Third, we will need some background on the epistemology of evolutionary debunking as well. As I understand them, evolutionary debunking arguments claim that facts about our evolutionary

7 This way of formalizing the argument is adapted from Koon (2021).
8 Similarly, we should understand phrases like "our capacity for moral cognition" as referring to whatever cognitive faculty or faculties are responsible for producing our moral intuitions, judgments, and beliefs, when those faculties are applied to the moral domain.
9 Compare Machery and Mallon (2010) on this point.

history provide us with higher-order evidence that our moral beliefs are unreliable.[10] Unlike first-order evidence, which bears directly on the truth or falsity of a belief, the distinguishing characteristic of higher-order evidence is that it gives us information about ourselves instead, about our ability to judge the evidence available to us. Evolutionary debunking arguments deal, in particular, with etiological higher-order evidence, higher-order evidence derived from facts about the causal history of our beliefs.

The debunking argument's epistemic principle is a consequence of calibrationism, the most prominent account of how we should revise our beliefs in response to higher-order evidence.[11] Calibrationists hold that higher-order evidence concerning the reliability of our beliefs has the power to undermine their justification – if you have good reason to think that one of your belief-forming mechanisms is unreliable, calibrationists maintain, you're no longer justified in retaining any of the beliefs that it produces. Hence, if your evidence indicates, on balance, that your capacity for moral judgment is likely to be unreliable, the right response, according to the calibrationist, would be to abandon all of your moral beliefs and retreat to a position of agnosticism about morality instead.

The debunking argument's epistemic principle contains an important restriction: it says that we shouldn't take all of the evidence into account when determining how reliable our moral beliefs are, only the independent evidence. This independence requirement also comes to us from calibrationism, which tells us that, when estimating how reliable we are in a given domain, we should always do so on the basis of the independent evidence alone.[12] What makes a piece of evidence independent, in the sense we're interested in here? When our ability to judge that piece of evidence has not been called into question by the higher-order evidence. For example, if you were to discover that a drug you were taking had impaired your ability to think and reason about mathematics, the independence

10 Christensen (2010) gives a nice overview of higher-order evidence.

11 White (2009), Sliwa and Horowitz (2015), Christensen (2016), Schoenfield (2018), Vavova (2018), and Kappel (2019) discuss different versions of calibrationism, while Schoenfield (2015) and Isaacs (2021) develop objections to the view. Calibrationism is typically formulated in terms of credences, but for the sake of simplicity, I'll stick to all-or-nothing beliefs here.

12 Most discussion of the independence requirement has been in the context of peer disagreement. Elga (2007) and Christensen (2007; 2009; 2010; 2011; 2018; 2019) defend the requirement, while Arsenault and Irving (2012), Kelly (2013), and Lord (2014) raise objections.

requirement would instruct you to bracket off all of your mathematical intuitions and beliefs when figuring out how to respond to this discovery. But it would still be fine to rely on your background knowledge of human psychology, pharmacology, and so on, because you would have no reason to think that your judgment about those subjects had also been impaired.

This independence requirement is needed to block a number of quick objections to the debunking argument.[13] For instance, the realist might try to respond to the argument as follows: "You claim that facts about our evolutionary history give me strong evidence, on balance, for thinking that my capacity for moral judgment is unreliable. But your assessment doesn't give any weight to my moral intuitions, which are also a form of evidence. And, when I consult my moral intuitions, they confirm that every single one of my moral beliefs is true. So, if we take all of the evidence into account, rather than arbitrarily restricting our attention to a subset of the available evidence, it turns out that I'm extremely reliable in the moral domain. As a result, your debunking argument poses no threat to the justification for my moral beliefs."

To many, this response sounds blatantly question-begging. Surely it's impermissible to use the outputs of a cognitive faculty whose reliability has been called into question as evidence that that very same faculty is reliable. The realist's reasoning seems comparable to that of an astrologer, who, when confronted with evidence that there's no conceivable mechanism by which heavenly bodies could influence our actions here on Earth, replies that she knows her horoscopes are reliable because she's seen it written in the stars. The independence requirement serves to rule out this type of question-begging response.

Third-factor explanations present a more sophisticated version of the same objection.[14] Here is how one such explanation might go: "You claim that facts about our evolutionary history give us strong evidence, on balance, for thinking that our capacity for moral judgment is unreliable. But, in

13 The importance of the independence requirement to the evolutionary debunking project was first noted by White (2010) and Vavova (2014).

14 Third-factor explanations are so-called because they seek to explain why our capacity for moral judgment is reliable by positing that some third factor – well-being, for instance – influenced the evolution of our moral instincts, while also serving to partly ground the moral facts. Enoch (2010), Wielenberg (2010), Brosnan (2011) and Skarsaune (2011) all defend slightly different third-factor explanations, while Berker (2014) gives a nice overview. Also see Lutz (2018) for compelling criticism of this type of response to debunking arguments.

responding to this challenge, it's only fair that realists be allowed to cite some platitudes about what morality is like. In particular, it seems obvious that moral goodness is in some way connected to well-being, so we should at least be granted the very weak and plausible assumption that the goodness facts are partly grounded in the well-being facts. And, once we're granted this assumption, it's no longer clear that the evidence supports your claim. To the contrary, it looks like the evidence would then support the opposite conclusion, that evolution would likely have made us into reliable judges in the moral domain. After all, natural selection can be expected to favor the belief that well-being is good, since seeking to improve one's own well-being, and the well-being of one's tribesmen, would typically have improved our ancestors' prospects of survival and reproduction. Hence, if we assume that the goodness facts are grounded in the well-being facts, it stands to reason that natural selection would push us towards holding moral beliefs that are largely true. So it looks like your debunking argument does not, in the end, succeed at undermining the justification for our moral beliefs."[15]

Here, too, the realist violates the independence requirement, in helping herself to the assumption that moral goodness is in some way connected to well-being. Debunkers will insist that, when evaluating higher-order evidence that calls our reliability in the moral domain into question, we're not allowed to make any assumptions about what morality is like, including assumptions about how moral facts are grounded in natural facts. That's because there's a serious worry that our grounding judgments are being illicitly influenced by our first-order moral beliefs and intuitions, that the underlying reason why we find it plausible that the goodness facts are grounded in the well-being facts is just that, intuitively, well-being seems morally good to us. And, the debunker claims, if we bracket off all of our beliefs and presuppositions about morality and attend to the independent evidence alone, we'll then be forced to conclude that our capacity for moral judgment is likely to be unreliable, since it was selected for a biological function other than producing true beliefs.

That's all the stage-setting we'll need. Let's turn, now, to the charge that evolutionary

15 Some readers might be concerned that the realist is cheating a bit by claiming that moral goodness is grounded in well-being, since well-being is often considered a normative property, rather than a purely natural property. This won't matter much for our purposes, but, if necessary, the realist could substitute "human health, knowledge, and happiness" for "well-being" throughout the objection, without detracting from the point she's making.

debunking arguments are self-undermining.

3. The Self-Undermining Objection

The self-undermining objection has been advanced by a number of authors, including Vavova (2014), Selim Berker (2014), and Arnon and Yair Levy (2020). Although these authors frame the objection in different ways, the basic idea is this: evolutionary debunking arguments purport to show that, if moral realism is true, our moral beliefs are systematically unjustified. But evolutionary debunking arguments, if sound, would undermine the justification for our epistemic beliefs (realistically construed) just as effectively as they undermine the justification for our beliefs about morality (realistically construed). If it's mysterious how our moral instincts could have evolved to serve as a reliable guide to a realm of mind-independent moral facts, it's equally mysterious how our epistemic instincts could have evolved to serve as a reliable guide to a realm of mind-independent epistemic facts. Hence, if we accept that our moral beliefs have been successfully debunked, we should think that our normative epistemic beliefs have been debunked as well. However, as we saw in Section 2, evolutionary debunking arguments invoke epistemic principles as premises, and can't get by without them. This is bad news for the debunker, since it means that evolutionary debunking arguments end up debunking themselves, too – they're self-undermining. And, since they're self-undermining, that means they've failed to show that our moral beliefs are unjustified after all.

The objection can be formalized as a *reductio*:

(1) Assume, for purposes of *reductio*, that evolutionary debunking arguments succeed at undermining the justification for our moral beliefs, realistically construed.

(2) If evolutionary debunking arguments succeed at undermining the justification for our moral beliefs, realistically construed, they also undermine the justification for our epistemic beliefs, realistically construed.

(3) If evolutionary debunking arguments undermine the justification for our epistemic beliefs,

realistically construed, then we shouldn't accept the epistemic principles invoked by evolutionary debunking arguments.

(4) If we shouldn't accept the epistemic principles invoked by evolutionary debunking arguments, evolutionary debunking arguments don't succeed at undermining the justification for our moral beliefs, realistically construed.

(5) Evolutionary debunking arguments don't succeed at undermining the justification for our moral beliefs, realistically construed. (1-4, MP)

(6) $\perp$ (1, 5)

Conclusion: By *reductio*, evolutionary debunking arguments don't succeed at undermining the justification for our moral beliefs, realistically construed.

A first problem for the objection is that (2) lacks adequate support. We have a reasonably good handle, at this point, on how and why morality evolved in human beings – as I noted in the introduction, the prevailing view in the scientific literature is that morality was selected chiefly to facilitate cooperation in our hunter-gatherer ancestors, because of the enormous boost to reproductive fitness that our ancestors would have gained through mutual cooperation with other members of their tribe. As a result, we can say with some confidence that our moral instincts were selected to promote cooperation, not to acquire true beliefs about a mind-independent domain of moral facts, and it's this mismatch between the evolved function of morality and the purposes the realist wishes to put it to that creates an opening for the debunking argument. We know next to nothing, in contrast, about how or why epistemic norms arose in our lineage, which means that we can't be confident there's a similar disconnect between their evolved function (if they have one) and the purposes epistemic realists wish to put them to. Hence, the empirical basis for an evolutionary debunking argument targeting our epistemic beliefs is in much worse shape than the empirical basis for an evolutionary debunking argument targeting our moral beliefs, so the parity the realist lays claim to in (2) is not really present.

To confirm this, let's take another look at the empirical premise of the argument from Section 2:

Empirical Premise: We have good reason to think that our moral instincts were selected principally for functions other than producing true moral beliefs.

This empirical premise is widely endorsed when it comes to morality, but little to no research has been conducted on its epistemic equivalent. We just don't know enough about the evolutionary history of our epistemic instincts at present to affirm that they were selected for functions other than producing true epistemic beliefs. Thus, since an evolutionary debunking argument against epistemic realism will fail even if a similar argument targeting moral realism succeeds, the debunker can make a strong case that we should reject (2).

Still, I don't think debunkers should be content with this response. To see why, let's consider a pair of plausible hypotheses about how our epistemic instincts may have evolved. First, our epistemic instincts might have been selected to encourage the adoption of epistemic norms that are instrumentally useful in helping us to acquire true beliefs about the world around us. Consider epistemic norms like "if all past instances of *F* were *G*, you should expect future instances of *F* to be *G* as well" and "in the absence of specific reasons for skepticism, it's reasonable to trust the testimony of others." Conceivably, a biological predisposition to endorse norms like these may have assisted our ancestors in the task of acquiring true beliefs about their surroundings, and, as many authors have observed, holding accurate beliefs about your surroundings will (by and large) tend to improve your chances of survival and reproduction. Second, it's also possible that our epistemic instincts, like morality, are chiefly a social adaptation; perhaps they were selected to assist our hunter-gatherer ancestors in coordinating their beliefs with their fellow tribesmen, regardless of whether those beliefs happened to be true. It's not hard to see how this might go when we consider the large number of cultural, moral, and religious beliefs that play a central role in structuring human relations, but which

are not themselves susceptible to simple empirical confirmation or refutation.[16]

Notice, though, that both of these hypotheses suggest that our epistemic instincts were selected for functions other than guiding us to true beliefs about a mind-independent domain of epistemic facts – the first hypothesis claims that our epistemic instincts were selected to aid us in acquiring true beliefs about the world, not about epistemology, while the second hypothesis claims that our epistemic instincts were selected to facilitate cooperation by helping us coordinate our beliefs with other members of our tribe. Consequently, if either hypothesis ends up being vindicated by subsequent research into human evolution, the empirical premise of a debunking argument targeting epistemic realism would then come out true. Thus, while debunking arguments against epistemic realism can't succeed at present, given the current state of our scientific knowledge, there's a fair chance that they will become viable in the future. So I think it would be unwise for debunkers to gamble that (2) will remain false in the long run.[17]

A second and more serious problem for the self-undermining objection is that it generalizes too broadly. As we'll see, if we apply the objection consistently across similar cases, it effectively suggests that we should never revise our moral beliefs in response to higher-order evidence that our capacity for normative judgment is globally unreliable, no matter how compelling that evidence may be. Suppose, for instance, that Eric has just drawn conclusions about a variety of difficult moral topics he'd never previously put much thought into, but then learns he's been dosed with Stopsucarin, a drug which systematically distorts your ability to reason about all normative matters (including epistemic matters) without giving you any conscious indication that your judgment is impaired. To ensure that the thought experiment isn't confounded by the effects of memory, let's also assume that Eric has no pre-existing opinion about how he should revise his beliefs in response to this type of higher-order

16 See Tomasello (2014; 2019), Mercier and Sperber (2017), and Dethier (2023) for some discussion of these issues.

17 Some authors, including Berker (2014), have wondered whether evolutionary debunking arguments really depend on the finer details of how we evolved, or if they can instead be constructed for any causal history our normative beliefs might turn out to have. So, is there any hypothesis about how epistemic cognition evolved that would not lend itself to evolutionary debunking arguments? Yes, at least one – if we have no epistemic instincts at all, and our epistemic beliefs are generated by a domain-neutral information-processing mechanism in the brain, whose biological function is just to produce true beliefs about whatever subject it's applied to.

evidence.[18] After careful deliberation, Eric decides that, because he has strong evidence that his capacity for normative judgment is unreliable, he's not justified in retaining any of the moral beliefs he adopted while under the influence of Stopsucarin, and he elects to suspend judgment about the moral issues he'd been considering instead.[19]

Intuitively speaking, this response seems quite reasonable: in retreating to a position of agnosticism about the moral issues he'd been considering, Eric has reacted to the news that his judgment is impaired in exactly the way he ought to (indeed, he should probably be commended for keeping such a level head while under the effects of the drug). But that's not the conclusion we reach if we try to adapt the self-undermining objection to Eric's case. Call this the self-undermining objection*:

(1*) Assume, for purposes of *reductio*, that Eric learning that he's been dosed with Stopsucarin succeeds at undermining the justification for his recently-acquired moral beliefs.[20]

(2*) If Eric learning that he's been dosed with Stopsucarin succeeds at undermining the justification for his recently-acquired moral beliefs, it also undermines the justification for his recently-acquired epistemic beliefs.

(3*) If Eric learning he's been dosed with Stopsucarin undermines the justification for his recently-acquired epistemic beliefs, he shouldn't accept the epistemic principle which led him to think that his recently-acquired moral beliefs were unjustified.

(4*) If Eric shouldn't accept the epistemic principle which led him to think that his recently-acquired

18 There's an immediate worry with this example about whether it's even possible for Eric to learn that he's been dosed with Stopsucarin, since, if Eric's capacity for epistemic reasoning has been compromised, he can no longer trust that he's responding appropriately to the evidence that he's been drugged, or his evidence about the drug's effects. We can finesse this problem by stipulating that Stopsucarin has been designed so that it doesn't interfere with our ability to learn or remember facts about the drug itself, and that Eric has access to the bottle of Stopsucarin and can verify this by reading the label. I concede that there are still some residual questions about whether Eric can truly know that he's interpreting the label correctly under these conditions, but the claims I make in what follows don't require Eric to be certain that his capacity for normative reasoning is impaired, only that he has good reason to think that it's impaired.

19 Note that Eric's reasoning here invokes a calibrationist-style epistemic principle.

20 This version of the objection also requires “realistically construed” qualifiers in the same locations as the original objection. In the interests of clarity, I've suppressed these in the text.

moral beliefs were unjustified, Eric learning that he's been dosed with Stopsucarin doesn't succeed at undermining the justification for his recently-acquired moral beliefs.

(5*) Eric learning that he's been dosed with Stopsucarin doesn't succeed at undermining the justification for his recently-acquired moral beliefs. (1*-4*, MP)

(6*) $\perp$ (1*, 5*)

Conclusion*: By *reductio*, Eric learning that he's been dosed with Stopsucarin doesn't succeed at undermining the justification for his recently-acquired moral beliefs.

The self-undermining objection* suggests that Eric should ignore the evidence that he's been drugged, and retain all of his moral beliefs instead. And this version of the objection seems like it must be sound if the first one is; if anything, the self-undermining objection* looks to be on better footing than the original, since (2*) follows straightforwardly from the fact that Stopsucarin interferes with all types of normative judgment, while (2) in the original objection depends on a highly speculative account of how our epistemic instincts evolved. What's more, there's nothing distinctive about Eric that's making it possible to adapt the self-undermining objection to this case, which means the objection can be further extended to any case where we learn that our capacity for normative reasoning has been globally impaired. So, suitably generalized, the self-undermining objection ends up saddling us with the unpalatable consequence that we should never revise our moral beliefs when faced with evidence that our normative judgments are globally unreliable. Clearly, something has gone wrong here.

I suspect that the problem lies with (4) and (4*). Let's begin with Eric's case (speaking from the perspective, now, of a normative realist). The self-undermining objection* is right that Eric learning he's been dosed with Stopsucarin undermines the justification for his recently-acquired epistemic beliefs, and that this should lead him to be skeptical of the epistemic principle that he relied on when he initially concluded that his moral beliefs were unjustified. But this does nothing to rehabilitate the

epistemic status of his moral beliefs, as (4*) suggests. That's because, when you learn that your capacity for normative judgment is globally impaired, the correct response is to suspend judgment across the board about all normative matters, not to ditch the epistemic beliefs while retaining all of the others. Here's what someone in Eric's position ought to think: "I have evidence that the moral beliefs I formed while under the influence of the drug are unreliable, and it seems to me that this means I should abandon those beliefs. Of course, I can't trust the epistemic principle I'm relying on now, either, since the same evidence calls my epistemic judgments into question as well. And, while we're at it, I also can't be confident that this evidence really undermines my epistemic judgments, since drawing this conclusion also involves making an epistemic judgment… Oh dear. This is turning into a bit of a debacle. I suppose the only safe thing to do under these circumstances is to suspend judgment about all of these issues."[21]

In other words, if you not only have evidence that your moral beliefs are unreliable, but also can't trust yourself to figure out how to respond rationally to this evidence, this should make you all the more inclined to throw up your hands and retreat to a position of agnosticism about morality. The undermining of your epistemic beliefs doesn't somehow cancel out the undermining of your moral beliefs, restoring the latter to their full epistemic status. Hence, (4*) is false, and Eric was right to abandon his moral beliefs after learning he'd been dosed with Stopsucarin.

By parity of reasoning, the original self-undermining objection suffers from the same defect. Assuming normative realism is true for the moment, (4) is false: even if the objection does show that we shouldn't accept the debunking argument's epistemic principle, that does nothing to restore justification to our moral beliefs. It would only mean that, since we're suffering from global normative confusion, we're no longer justified in retaining any of our normative beliefs at all, moral or epistemic. Just like in Eric's case, the right response would be to suspend judgment across the board. I conclude, therefore, that the self-undermining objection fails.[22]

21 This is not to say that Eric should endorse the epistemic principle that the appropriate thing to do in his circumstances is to suspend judgment; he can't trust that he's right about that, either. In other words, Eric ought to suspend judgment, but he's not in a position where he can rationally be confident that he ought to suspend judgment.

22 How can it be that evolutionary debunking arguments succeed at undermining the justification for our moral

But perhaps I've been interpreting the self-undermining objection the wrong way – perhaps, rather than the formal argument laid out above, we should instead understand it as a *tu quoque* objection pointing to an inconsistency in the belief set of the anti-realist who is pushing the debunking argument. If (2) is true, evolutionary debunking arguments threaten the justification for our epistemic beliefs just as surely as they threaten the justification for our moral beliefs. So how can anti-realists insist that the debunking argument's conclusion is correct, while simultaneously asserting that we should believe the argument's epistemic principle is true? Doesn't this make them guilty of some kind of inconsistency?

Generally speaking, the answer is no, because most debunkers aren't epistemic realists,[23] and debunking arguments will succeed at undermining the justification for our epistemic beliefs only on the assumption that epistemic realism is true. The source of the debunking argument's plausibility is that it seems like it would be a remarkable coincidence if epistemic instincts selected for some function other than acquiring true epistemic beliefs – to help our ancestors coordinate their cultural and religious mores with other members of their tribe, for example – also served as a reliable guide to a mind-independent domain of epistemic facts. But this problem only arises for a realist metaphysics; it won't affect constructivists and relativists about epistemic norms, who locate epistemic facts in our minds, institutions, and social practices. There's just no evolutionary mystery about how we can gain

beliefs, if we're unable to accept the epistemic principles they take as premises? There are two ways of understanding how this might work. First, it may be that, in the unusual epistemic conditions created by evidence of global normative unreliability, the debunking argument is able to succeed even if we adopt attitudes weaker than acceptance towards its epistemic premises. Perhaps it's sufficient, under these conditions, that it appears to us that there's a true epistemic principle which entails that our moral beliefs are unjustified, and that the reason why things appear this way to us is because there actually is a true epistemic principle which entails that our moral beliefs are unjustified. Alternatively, it may be that, strictly speaking, it's not the debunking argument itself which undermines the justification for our moral beliefs, but the higher-order evidence about human evolution cited by the argument. On this view, the debunking argument's role is to indicate the epistemic significance of a batch of higher-order evidence, and the justification for our moral beliefs is not really undermined by our acceptance of the debunking argument's premises and conclusion, but by the evidence it adverts to. If this interpretation is correct, the self-undermining objection is technically sound, but it fails at accomplishing its goal, because even if the debunking argument itself doesn't undermine the justification for our moral beliefs, the evidence still does. Note, however, that there's no important difference between evolutionary debunking and Eric's case on this score, since Eric's reasoning could easily be represented as a formal argument as well (and the same goes for any other case involving evidence of global normative unreliability). Hence, whatever we say about this question, it won't affect the conclusion that the self-undermining objection overgeneralizes.

23 For instance, of the two most well-known debunkers, Sharon Street (2009a) is a constructivist about epistemic norms, while Richard Joyce (2019) is uncommitted.

knowledge of these things; we learn about them on an individual level through introspection, observation, and testimony, and study them more systematically as part of anthropology and the other social sciences. The same goes for expressivists, who, inasmuch as they recognize the existence of epistemic knowledge at all, will tell a similar story about how we acquire it as constructivists and relativists.[24] Accordingly, epistemic constructivists, relativists, and expressivists have an easy way of escaping from the *tu quoque* objection – they can maintain that the debunking argument succeeds at undermining moral realism, without causing any collateral damage to their own anti-realist epistemic views in the process.[25]

The situation is more complicated for epistemic error theorists, who agree with the realist that normative claims can only be true if there are mind- and language-independent facts to answer to them, but who depart from the realist in thinking that no such facts exists and hence that all normative claims are false. This does create a puzzle – what business does someone who thinks that all epistemic claims are false have advancing an argument that explicitly invokes an epistemic principle? More generally, if you're a universal error theorist, how can you make sense of debunking arguments in the first place, and how should you respond to the self-undermining objection?

Here's how I think an error theorist should answer these questions: she should say that both the moral norms and the epistemic norms we accept are, to a large extent, illusions built into our minds by natural selection, in conjunction with other evolutionary forces.[26] If our moral and epistemic norms are illusions, however, nothing guarantees that they'll be consistent, internally or with each

24 As I understand the view, this will also be true for quasi-realist forms of epistemic expressivism, modeled on the moral quasi-realism defended by Blackburn (1996) and others. Quasi-realism is notoriously difficult to interpret, however, so I concede that this point is not altogether clear.

25 Berker (2014) argues that the constructivist's meta-ethical beliefs about how moral facts are grounded in our attitudes are vulnerable to a parallel evolutionary debunking argument. I don't have space to respond to Berker at length here, but I will say that it's difficult to see how the empirical premise of this parallel debunking argument could be substantiated, given how poorly we understand the psychological origins of the constructivist's meta-ethical beliefs. What psychological faculties produce the constructivist's beliefs, what are their innate components, and what, if anything, were these innate components selected for? No one has any idea how to answer these questions at present, given the current state of our knowledge of the brain. We can't even say that there are plausible answers to these questions that would be congenial to debunking arguments, as we could in the case of our epistemic norms. This contrasts with the realist's beliefs about the grounding of moral facts – "the badness facts are grounded in the pain facts," and so on – which are pretty obviously a product of the same moral instincts which influence our first-order moral judgments, and so will be vulnerable to the same debunking argument.

26 This view originates with Ruse (1986).

other, particularly when we begin to bring scientific facts about our nature as knowers and valuers into the picture as well. Evolutionary debunking arguments reveal one such conflict: when combined with contemporary scientific research into humanity's evolutionary history, our epistemic norms suggest that our confidence in our moral beliefs is misplaced, and we would be better off abandoning them.[27] What the self-undermining objection is pointing out is that these same epistemic norms will likely end up undermining their own justification as well – our epistemic norms entail that we're not justified in trusting our epistemic norms, either. But this does nothing to damage the error theorist's case; indeed, if anything, it strengthens it. From the error theorist's perspective, the more internal conflicts in our norms that can be located, the more plausible the thesis that they're illusions built into our minds by natural selection, since selection does not share our human preoccupation with consistency.

Of course, the error theorist will say, the epistemic principle invoked by the debunking argument isn't literally true. But the epistemic principle doesn't need to be true for the debunking argument to accomplish its goal, because the purpose of the debunking argument isn't to show that it's a fact about our moral beliefs that they're unjustified (naturally, as there are no such facts), it's to demonstrate that the system of norms we accept, taken as a whole, is replete with inconsistencies. These inconsistencies are inexplicable if we understand our moral and epistemic beliefs to correspond to a mind-independent domain of normative facts – how could facts be inconsistent? – but they're far less surprising on the hypothesis that all of our norms are an illusion that natural selection has cobbled together from random mutations in order to make us into better cooperators.

In the final accounting, then, every type of epistemic anti-realist has a viable response to the *tu quoque* version of the self-undermining objection, which means the objection will succeed only against epistemic realists – epistemic realists, that is, who also advance evolutionary debunking argument against moral realism. Anyone fitting this description can rightfully be convicted of inconsistency, of propounding an argument which can readily be extended to undermine its own premises. But the

27 This isn't a strict logical inconsistency, but an instance of epistemic akrasia – our moral norms present themselves to us as being true, and to some extent are probably psychologically incorrigible, while at the same time the debunking argument suggests that all of our moral beliefs are epistemically unjustified. For background on epistemic akrasia, see Horowitz (2014).

scope of the objection ends up being far more limited than it might have seemed at first glance.

There is an important point in the vicinity of the self-undermining objection, namely, that the fortunes of moral realism and epistemic realism are closely linked. If you're willing to believe that there's a mind-independent domain of moral facts out there, there's little additional cost to positing a mind-independent domain of epistemic facts to go along side it; conversely, if you reject moral realism, either because you don't see how moral facts can fit into a world that science tells us is made out of particles and fields of force, or because you don't see how creatures like us could have epistemic access to the realist's mind-independent moral facts, even if they did exist, then you should probably have the same misgivings about epistemic realism as well. So critics of evolutionary debunking arguments are entitled to point out that epistemic realism and moral realism are companions in guilt, or companions in innocence. In other words, if we accept that debunking arguments targeting moral realism are sound, this will likely end up forcing us to sacrifice epistemic realism, too.[28] My view is that this companions-in-guilt strategy gives the realist only a tiny bit of extra traction in resisting debunking arguments, and mainly serves to put epistemic realism on the chopping block as well. Unfortunately, defending this claim would take us too far beyond the scope of this paper, so we'll have to let matters rest there.

4. Vavova's Objection

On, now, to the second objection. Recall, from Section 2, that the calibrationist independence requirement ends up playing a crucial role in defending evolutionary debunking arguments, because it's needed to block a variety of quick objections, all of which start out by making realist-friendly assumptions about the nature of morality, and reason from there to the conclusion that our moral beliefs are largely true. The debunker claims that, if we heed the independence requirement and set aside all of our beliefs and presuppositions about morality, we'll then be left with strong evidence, on balance, for thinking that our moral beliefs are unreliable. At this juncture, Vavova (2014: 92) raises

28 See Cuneo (2007), Cowie (2014; 2016), and Case (2019) for some discussion of companions-in-guilt arguments linking moral realism and epistemic realism, outside of the context of evolutionary debunking.

an important objection. She writes:

> [W]e cannot determine if we are likely to be mistaken about morality if we can make no assumptions at all about what morality is like… [T]he debunker's challenge threatens anyone who holds that the attitude-independent moral truths do not, in any helpful way, coincide with the evolutionarily advantageous beliefs… But even to make this crucial judgment, that these two sets do not have the same contents, we need to know something about the contents of those sets—what they are or what they are like. Compare: I cannot demonstrate that I am not hopeless at interacting with external objects in my manifest surroundings without knowing something about what those objects and surroundings are like. Likewise, I cannot show that I am not hopeless at understanding right and wrong without being allowed to make some assumptions about what is right and wrong.[29]

This line of thinking can be summarized in a principle I'll call Background:

Background: It's impossible to establish that your beliefs in a given domain, *D*, are likely to be unreliable, without either having some background knowledge of *D,* or making some assumptions about what *D* is like.

At first glance, Background seems quite plausible. If you have no background knowledge of *D*, and make no assumptions about what *D* is like, how can you be confident that you're an unreliable judge of the *D*-related facts? Surely, as Vavova suggests, to determine that you're unreliable about *D*, you'll need to be able to compare your beliefs about *D* to a list of truths about *D*, and see how well they correspond. But there's no way for you to come into possession of a list like this unless you have some

29 Vavova (2021) reiterates this point.

background knowledge on the subject, or you're allowed to make some assumptions about what it's like. The debunker, in other words, has set an impossible task for herself: she intends to show that our moral beliefs are likely to be unreliable, even though the independence requirement precludes her from making any assumptions about the nature of morality. This cannot be done. Consequently, the debunker's project is doomed to fail.

Vavova's point may be correct if we restrict our attention to methods that seek to establish that you're unreliable about *D* on the basis of your track record of beliefs concerning *D*. This procedure is liable to be fruitless unless you're allowed to make some assumptions about what *D* is like.[30] But one of the virtues of higher-order evidence is that it allows us to extend our knowledge of our own reliability or unreliability to realms where we don't have track record evidence of this sort. And, once we start thinking about different types of higher-order evidence, it soon becomes apparent that we can use them to construct a variety of counter-examples to Background.

To illustrate, here's a counter-example to Background involving higher-order evidence acquired by testimony: suppose that Madeleine is a child who has just heard the term "morphic resonance field" for the first time on a popular science-fiction television show, but is unsure what it means.[31] She wonders if morphic resonance fields are the sort of thing you can sense or feel, so she asks her father, "Am I a reliable judge of the presence or absence of morphic resonance fields?" "No," her father replies (if it makes a difference, we can assume that the father is an expert on the topic). Madeleine now has good reason to think that any judgments she makes about the presence or absence of morphic resonance fields are unreliable, despite the fact that she still has no idea what they are like, or whether they even exist. Hence, Background is false in Madeleine's case.[32]

30 In fact, even this conclusion is too strong, because it will sometimes be possible to establish that you're unreliable about *D* from the formal features of your *D*-related judgments alone, if, for instance, those judgments are logically inconsistent. This method will not require you to know or assume anything about *D*.

31 Here and in what follows, I use "morphic resonance fields" as a generic example of a metaphysical domain that agents are unfamiliar with. Nothing significant hangs on this choice.

32 If you think this case only works if we assume that the father is an expert on morphic resonance fields, you might wonder if we can modify Background to avoid the counter-example as follows:

Background*: It's impossible to establish that your beliefs in a given domain, *D*, are likely to be unreliable, without either making some assumptions about what *D* is like, having some background knowledge of *D*, or being connected by a chain of testimony to someone who either makes some

We can construct a similar counter-example involving moral judgments instead. Suppose Luke is a child who was born with a congenital disease that makes him completely insensible to moral considerations. He's just heard the term "morality" for the first time while watching a popular crime drama, but is unsure what it means, so he asks his mother, "Am I a reliable judge of what's moral or immoral?" "No," his mother responds. Luke now has strong evidence that his moral judgments are unreliable, despite the fact that he lacks even a rudimentary grasp of what morality is about. Luke's case demonstrates that Background is not only false in general, but also false specifically when it comes to morality.

Other types of higher-order evidence supply us with yet more counter-examples. Suppose that you – having no background knowledge of morphic resonance fields – are told that you've been dosed with a highly-specialized psychotropic drug, one that distorts any judgments you make about resonance fields, but has no other effects on your cognitive abilities. Or suppose that a predictor, like the one featured in Newcomb's paradox, informs you that it has studied your constituent atoms and conducted 14,000,605 simulations, and you turned out to be an incompetent judge of facts related to morphic resonance fields in all of them. Again, in both of these cases, you would have good reason to think that your judgments about morphic resonance fields are unreliable, without having any background knowledge about what they are, or making any assumptions about what they're like.

Most importantly for our purposes, information about the etiology of your *D*-related judgments can also demonstrate that those judgments are likely to be unreliable, while simultaneously pushing you to renounce all of your current beliefs and assumptions about *D*. There are two types of cases like

---

assumptions about what *D* is like or has background knowledge of *D*.

Background* will get the right result about the version of Madeleine's case where her father is an expert, and can still be used to block evolutionary debunking arguments, since debunking arguments are supposed to succeed even if no one anywhere has any moral knowledge. Intuitively, however, it seems clear to me that Madeleine should heed her father's testimony so long as she reasonably believes that her father is an expert, whether or not he actually is an expert. If Madeleine knows her father is generally a trustworthy source of information, she should conclude that her judgments about morphic resonance fields are likely to be unreliable even if, as it turns out, her father has no actual knowledge of the subject and is just yanking her chain (compare Lackey [1999] on this point). In the end, though, it won't matter what we say about this case, because later on I'll present counter-examples to Background that definitely don't require anyone to have relevant background knowledge, and Background* will be no help with those.

this. In the first type, you learn that your judgments have a deviant etiology, in the sense that the function of the cognitive mechanism generating those beliefs is radically disconnected from the pursuit of truth. In the second type of case, you learn that your judgments have a random etiology, in the sense that the belief-forming mechanism responsible for those beliefs has been shaped in ways that are arbitrary or chancy, rather than truth-guided, and could have easily led you to hold different beliefs on the subject instead.

For an example of a deviant etiology, suppose that Marcus has a variety of beliefs about morphic resonance fields, but subsequently learns that all of those beliefs were produced by a microchip that was surgically implanted in his brain by venal neuroscientists looking to swindle him out of his inheritance.[33] Under these circumstances, it seems like it would be rational for Marcus to retreat to a position of total agnosticism about morphic resonance fields, while at the same time considering himself an unreliable judge on the subject. Similarly, we can also imagine a case where we discover that all of our moral instincts were implanted in our brains at birth by sinister aliens, with the aim of making us into pliable slaves and consumers, not unlike the plot of the classic John Carpenter movie *They Live*. As with Marcus, it seems like the right response to this discovery would be a combination of agnosticism about the nature of morality, along with skepticism about your ability to tell right from wrong. If you find out that everything you believe about morality is the product of deceit and manipulation, it's difficult to see how you could continue to trust that your moral judgments are reliable, or that you have any real insight into what morality is like.

A defender of Background might wonder whether we're really making no assumptions at all about *D* in these cases. For instance, for the second example to work, don't we have to assume that moral goodness doesn't consist in being a servile conformist to alien interests? If morality were connected to conformism in this way, the aliens' intervention might have the effect of making our moral beliefs more reliable overall, rather than less reliable. Of course, the proposal that conformism is the highest moral good seems a bit far-fetched, but don't we still need to make some implicit

33 Compare the Napoleon pill case from Joyce (2006: 179).

assumptions about the nature of morality in order to rule it out? The answer to this question is no; the question is confusing making the assumption that $\neg p$ with not giving undue weight to the hypothesis that p. If we retreat to a position of agnosticism after discovering that our moral beliefs are the product of brain manipulation, we won't be assuming that there's no connection between morality and conformism, we'll be suspending judgment on the matter. An attitude of suspension of judgment will leave open the possibility that conformism is the highest good, and hence allow for a chance that our moral judgments will turn out to be reliable, but won't give that possibility significantly more weight than any other specific hypothesis about the nature of morality. Thus, the intuitive response to finding out that your beliefs in domain *D* have a deviant etiology – that you should retreat to a position of agnosticism about *D*, while also thinking that you're likely to be an unreliable judge on the subject – doesn't require you to make any assumptions about the nature of *D*.[34]

Cases where you learn that your beliefs have a random etiology can serve as counter-examples to Background as well. For instance, we can imagine that Marcus instead finds out that his beliefs about morphic resonance fields are an inadvertent side effect of the microchip, which the neurosurgeons installed in his brain merely to monitor him. Or, we can imagine that the meddling aliens chose which moral instincts to implant in our brains by some chance process, as part of a twisted science experiment, rather than with the aim of enslaving or exploiting us. This discovery, I

34 One caveat: in order for an agent to understand that her moral judgments are likely to be unreliable, it may be necessary that she, or at least someone she's linked to by testimony, be familiar with basic normative concepts like permission and obligation. It's difficult to imagine how we could determine that our moral judgments are likely to be unreliable if no one anywhere had any idea what a rule or a norm was, or what it would mean for an action to be permissible or impermissible. Essentially, we need a rich enough conceptual vocabulary to be able to affirm that, if there are norms governing how human beings can permissibly treat one another, we have no way of knowing what they are. It's less clear that we would need any specifically moral concepts, even "moral" itself – it seems possible to recognize that I'm unreliable in a given domain without knowing what the domain is called (for instance, I knew my intuitions about how light and matter interact on the subatomic level were unreliable long before I learned that this field was called quantum electrodynamics). Fortunately for the debunker, understanding the meaning of basic normative concepts doesn't require us to make any assumptions about the nature of morality. The proof of this is that these concepts are widely shared across different views in ethics and metaethics, which suggests that they don't carry any substantive commitments in either of these domains. If Ross, Ayer, Mackie, and Boyd can all comfortably discuss norms and obligations, the concepts themselves can't be constraining our theorizing to any significant degree. Additionally, it's important to remember that Background is based on the debunking argument's independence requirement, which instructs us to bracket off our moral beliefs and intuitions when figuring out how to respond to higher-order evidence suggesting that our capacity for moral judgment is unreliable. The independence requirement was never intended to prohibit us from making use of basic normative or moral concepts, so if Background is interpreted this broadly, it no longer works as a response to the debunking argument.

take it, would still give us reason to think that our beliefs about morality are likely to be unreliable, while at the same time pushing us to suspend judgment concerning the truth or falsity of all moral claims. There are many situations, then, in which information about the etiology of your beliefs can establish that you're an unreliable judge in domain *D* (including situations where *D* is, specifically, morality), while simultaneously making it reasonable to adopt an attitude of total agnosticism towards *D*, making no assumptions about what *D* is like. So Background must be false.

As I noted earlier, evolutionary debunking arguments are etiological: they claim that facts about the causal history of our moral instincts suggest that our moral judgments are likely to be unreliable. This gives debunkers a clear path to rejecting Background, modeled on the cases outlined above. But which of these two types of reliability-undermining etiology do debunkers claim for morality – a deviant etiology or a random etiology? I suggest that the right answer to this question for the debunker is: both. Our moral instincts have a deviant etiology, because they were built into our minds by natural selection to achieve the practical goal of facilitating cooperation among our ancestors, rather than to guide us to a mind-independent domain of moral truths.[35] But that's not all; our moral instincts also have a random etiology, because what sort of cooperative instincts we ended up with is a result of arbitrary features of our ancestors' genetic constitution and environment, together with chancy evolutionary processes like mutation and genetic drift.

The first of these two theses about the evolution of morality has been defended extensively in other works, so I will focus on the latter here, which is far less familiar.[36] Why think that the evolution

35 See Koon (2021: 12164-12168) on this point.

36 The closest precedent I know of is Wilson and Ruse (1985). Note that the debunking argument presented in Sections 1 and 2 is the deviant-etiology version of the argument. To adapt this into the random-etiology version of the debunking argument, the following changes are needed:

> Empirical Premise*: We have good reason to think that our moral instincts were shaped by evolutionary forces in ways that were random or chancy, and could easily have led us to adopt a substantially different collection of moral beliefs instead.
>
> Etiological Principle*: If moral realism is true, and if we have good reason to think that our moral instincts were shaped by evolutionary forces in ways that were random or chancy, and could easily have led us to adopt a substantially different collection of moral beliefs instead, then, unless we're able to corroborate our moral beliefs through the use of some other faculty whose etiology is not subject to similar doubts, the balance of independent evidence suggests that our moral beliefs are unreliable.

of our moral instincts was random, in the same sense that (say) the outcome of a roulette spin is random? The key feature of a roulette spin is that it exhibits sensitive dependence on initial conditions: small changes to the position and angular velocity of the wheel, or the speed and trajectory at which the ball is thrown, will lead to different outcomes. My suggestion is that the evolution of morality likely exhibited this same sensitive dependence on initial conditions: small changes to our ancestors' genomes, to the environment in which they evolved, or to the mutations and other chance evolutionary processes which occurred along the way would have produced large corresponding changes in our moral instincts, which in turn would have led us to adopt a comprehensively different collection of moral beliefs than we in fact hold.[37]

To get a sense of just how small a change in initial conditions would be needed, consider the evolutionary history of our closest living relatives, the chimpanzee and the bonobo. Although the two animals share 99.6% of the same genes, they behave in very different ways – chimpanzees are patriarchal and brutally violent, while bonobo troops are dominated by coalitions of females and far more peaceful.[38] The most prominent explanation of how the two species diverged comes from the primatologist Richard Wrangham, who believes the split was driven by the absence of gorillas on the south bank of the Congo River.[39] Chimpanzees, who inhabit the north bank of the Congo, were forced to compete with gorillas for scarce food supplies, and grew more territorial and aggressive in response to this scarcity. Bonobos, meanwhile, had the good fortune of evolving on the south side of the Congo, and the conditions of relative abundance they enjoyed there, with no gorillas around as competitors, allowed them to shed many of their violent instincts and adopt a more irenic lifestyle instead. If Wrangham's explanation is correct, this small environmental difference, the location of a few thousand gorillas, was sufficient to cause the speciation event that separated the chimpanzee and the bonobo, together with the major behavioral changes that accompanied it.

Given the close connections between moral instincts and behavior, it stands to reason that

37 There is now a large literature in biology and the philosophy of biology discussing evolution's sensitivity to initial conditions. See, for starters, Gould (1989) and Beatty (2006).
38 I take this from Prufer et al. (2012).
39 See Wrangham and Peterson (1996), Yamakoshi (2004), and Hare et al. (2012).

equally small shifts in our evolutionary history would have been sufficient to dramatically reshape the range of moral beliefs that human beings today find plausible.[40] In other words, if environmental conditions in the Pleistocene had been slightly different, if a few mutations had occurred or failed to occur at the appropriate time, or if genetic drift walked us in a different direction when our species's population size was at a low ebb, we could easily have ended up with radically different moral instincts than we actually have. There's no way of knowing, of course, exactly what instincts we would have ended up with under what conditions. If we were lucky, we might have evolved into a race of peaceful communitarians like the bonobos; unlucky, and we might have instead turned out more like chimpanzees, living in dystopian societies ruled by warlords and governed by the precept that might makes right. Or perhaps we would have ended up as cold-blooded utilitarians, or Stalinists, or proponents of some stranger moral code still.

The key point is that our moral instincts are the product of evolutionary processes which exhibit sensitive dependence on initial conditions, and so, like a roulette wheel, could easily have steered us in a different direction instead.[41] Hence, in a large proportion of the nearby possible worlds where our species followed a slightly different evolutionary trajectory, our moral beliefs would end up being massively in error. This means we can be confident, without making any assumptions about what morality is like, that moral instincts with an etiology similar to ours have a high chance of being an unreliable guide to the moral truth.

Vavova (2021: 725-726) considers a comparison between moral beliefs shaped by evolution and beliefs about US history chosen by a roulette-like game. While Vavova concedes that knowing your

40 Of course, nothing in this paper depends on Wrangham's account of how the chimpanzee and bonobo diverged being exactly correct. This is just an example illustrating the more general point that small changes in environmental conditions can have major effects on the subsequent evolutionary trajectory of a species, including on the behavioral traits which determine how members of the species treat one another.

41 I should note that there are a handful of moral instincts, chiefly concerning oneself and one's first-order relatives, whose evolution was probably not sensitive to initial conditions, because they would have been favored by selection in almost all ecological conditions remotely similar to those our ancestors actually encountered. I have in mind instincts like the disposition to see one's own pain as a bad thing, or the feelings of warmth and affection that mothers bear for their children. Natural selection will favor instincts like these under almost all circumstances, because their adaptive value doesn't depend on the finer details of our species's social organization. Apes differ immensely in how they treat conspecifics other than their children, but pain avoidance and maternal care behaviors are universal (see, for instance, Allman et al. [1998]).

beliefs about US history were chosen by roulette suggests that they're likely to be unreliable, she rejects the analogy to the evolution of morality. In the case of roulette, Vavova claims, we know going in that there's no connection between where the ball lands and (say) who was president in 1880, but she doesn't see how we can know that there's no connection between the beliefs evolution pushes us toward and the moral truth, unless we make substantive assumptions about what morality is like. "We cannot," she writes, "get reason to think that a given process is random or unreliable with respect to some matter from a standpoint that is agnostic about that matter… [A]bsent such assumptions, [we] can't get reason to think that the evolutionary process is random in the same way that [roulette] is random."

I believe, however, that Vavova is mistaken on this score, and that it's often possible to determine that a process is unreliable based on its formal features alone. To illustrate, suppose that each pocket of a roulette wheel corresponds to a different collection of beliefs about *D*, and the different collections of beliefs about *D* are all, to a large degree, mutually inconsistent. If these conditions are met, then a substantial proportion of the beliefs in most of the collections are guaranteed to be false, just by virtue of the formal structure of the game. And, if a substantial proportion of the beliefs in most of the collections are false, it follows that a set of beliefs about *D* chosen on the basis of a fair roulette spin has a high chance of being unreliable. What's more, we don't need to know anything else about *D* to recognize that this is the case – that's why we're able to characterize roulette as a defective method of belief formation in the abstract, without specifying what sort of beliefs we're talking about.[42]

Since it exhibits sensitive dependence on initial conditions, the evolution of morality shares the formal structure of roulette. Evolution could easily have endowed us with any of a variety of

42 There is one circumstance where a roulette game satisfying these conditions would not be a defective method of belief formation: if the game is set up so that the beliefs corresponding to each pocket will turn out to be true just in case the ball lands in that pocket (for instance, if each pocket *n* was associated with the belief "the ball landed in *n*"). This is analogous, in the evolutionary case, to a metaethical view which says that a moral belief is true just in case we evolved to find it plausible, that if we had instead evolved into a race of natural Stalinists, Stalinism would then be morally right for us. A view like this may be immune to debunking arguments, although most will find it unacceptable for other reasons. Additionally, it seems more appropriate to describe this as a form of biological relativism than as a type of genuine moral realism. Compare Street (2009b) on this point.

different sets of moral instincts, which in turn would have led us to adopt a variety of different and (to a large degree) mutually inconsistent collections of moral beliefs. Because these collections of moral beliefs are (to a large degree) mutually inconsistent, it follows, just as a matter of logic, that a substantial proportion of the beliefs in most of the collections must be false. Hence, just like in the case of beliefs chosen by roulette, learning that our moral beliefs are the product of evolutionary forces gives us evidence that they're likely to be unreliable, and we can recognize this without making any assumptions about the nature of morality.

Let's recap. Evolutionary debunking arguments claim that, if we bracket off all of our beliefs and presuppositions about morality, we'll then be left with good independent reason for thinking that our moral judgments are unreliable. Vavova rejects this claim: it's not possible, she argues, to establish that we're likely to be unreliable in a given domain without making some assumptions about what that domain is like. As it turns out, however, the principle Vavova is invoking here, which I have called Background, is vulnerable to a large number of counter-examples. The most important of these counter-examples involve higher-order evidence derived from the etiology of our beliefs, the same type of higher-order evidence cited by debunking arguments. Cases of this sort suggest that if you learn your beliefs in a given domain have either a deviant etiology or a random etiology, this gives you reason for thinking that your judgments in that domain are unreliable, while at the same time pushing you to retreat to a position of agnosticism about the domain in question.

Since the evolutionary history of human moral cognition counts as both a deviant etiology and as a random etiology, this means we have good reason, twice over, to think that our capacity for moral judgment is likely to be unreliable. Indeed, the evolution of morality turns out to be comparable, in important respects, to chance processes like roulette, and, clearly, learning that your moral instincts were chosen by roulette would make it difficult to maintain that your moral beliefs are reliable. Ultimately, then, careful reflection on how information about a belief's etiology can affect its epistemic status serves to vindicate debunking arguments against Vavova's objection.

5. Conclusion

As we've seen, both the self-undermining objection and Vavova's objection face serious challenges, leaving realists with substantially fewer options for resisting evolutionary debunking arguments than they were previously thought to enjoy. The self-undermining objection generalizes too broadly, and ends up committing its proponents to the implausible view that the justification for our moral beliefs can never be defeated by evidence that our capacity for normative judgment has been globally impaired. The principle Vavova invokes in her objection, meanwhile, succumbs to numerous counter-examples, just as soon as we begin to consider cases involving various types of higher-order evidence. Realists must look elsewhere if they hope to avoid the threat to their view posed by evolutionary debunking arguments – although many other objections to evolutionary debunking that once seemed promising will also need to be reevaluated in light of what we now know about higher-order evidence.

Acknowledgments

I would like to thank David Christensen, Rachel Dichter, Thomas Dougherty, Sophie Horowitz, Hilary Kornblith, and several anonymous referees, along with an audience at the 2022 Eastern Division Meeting of the APA, for their helpful comments on previous drafts of this paper.

References

Alexander, R. (1987). *The Evolution of Moral Systems*. New Brunswick: AldineTransaction.

Allman, J., Rosin, A., Kumar, R., and Hasenstaub, A. (1998). "Parenting and Surviving in Anthropoid Primates: Caretakers Live Longer," *Proceedings of the National Academy of Sciences* 95(12): 6866-6869.

Arsenault, M. and Irving, Z. (2012). "Aha! Trick Questions, Independence, and the Epistemology of Disagreement," *Thought: A Journal of Philosophy* 1(3): 185-194.

Baumard, N., Andre, J., and Sperber, D. (2012). "A Mutualistic Approach to Morality: The Evolution of Morality by Partner Choice," *Behavioral and Brain Sciences* 36(1): 59-78.

Beatty, J. (2006). "Replaying Life's Tape," *The Journal of Philosophy* 103(7): 336-362.

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: Essays on the New Science of Ethics*. Oxford: Oxford University Press.

Blackburn, S. (1996). "Securing the Nots: Moral Epistemology for the Quasi-Realist," in W. Sinnott-Armstrong and M. Timmons (eds.) *Moral Knowledge? New Readings in Epistemology*. Oxford: Oxford University Press.

Boehm, C. (2012). *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. New York: Basic Books.

Bogardus, T. (2016). "Only All Naturalists Should Worry About Only One Evolutionary Debunking Argument," *Ethics* 126(3): 636-661.

Bowles, S. and Gintis, H. (2011). *A Cooperative Species: Human Reciprocity and Its Evolution*. Princeton: Princeton University Press.

Brosnan, K. (2011). "Do the Evolutionary Origins of Our Moral Beliefs Undermine Moral Knowledge?" *Biology and Philosophy* 26(1): 51-64.

Case, S. (2018). "From Epistemic to Moral Realism," *Journal of Moral Philosophy* 16(5): 541-562.

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

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

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

Christensen, D. (2011). "Disagreement, Question-Begging, and Epistemic Self-Criticism," *Philosopher's Imprint* 11(6): 1-22.

Christensen, D. (2016). "Disagreement, Drugs, Etc.: From Accuracy to Akrasia," *Episteme* 13(4): 397-422.

Christensen, D. (2018). "On Acting as Judge in One's Own Epistemic Case," *Proceedings and Addresses of the American Philosophical Association* 93(1): 207-235.

Christensen, D. (2019). "Formulating Independence," in M. Rasmussen and A. Steglich-Petersen (eds.) *Higher-Order Evidence: New Essays*. Oxford: Oxford University Press.

Clarke-Doane, J. (2016). "Debunking and Dispensability," in U. Leibowitz and N. Sinclair (eds.) *Explanation in Ethics and Mathematics: Debunking and Dispensability*. Oxford: Oxford University Press.

Clarke-Doane, J. and Baras, D. (2021). "Modal Security," *Philosophy and Phenomenological Research* 102(1): 162-183.

Copp, D. (2008). "Darwinian Skepticism About Moral Realism," *Philosophical Perspectives* 18(1): 186-206.

Cowie, C. (2014). "Why Companions in Guilt Arguments Won't Work," *The Philosophical Quarterly* 64(256): 407-422.

Cowie, C. (2016). "Good News for Moral Error Theorists: A Master Argument Against Companions in Guilt Strategies," *Australasian Journal of Philosophy* 95(1): 115-130.

Cuneo, T. (2007). *The Normative Web: An Argument for Moral Realism*. Oxford: Oxford University

Press.

Dethier, C. (2023). "The Cooperative Origins of Epistemic Rationality?" *Erkenntnis* 88(3): 1269-1288.

Elga, A. (2007). "Reflection and Disagreement," *Nous* 41(3): 478-502.

Enoch, D. (2010). "The Epistemological Challenge to Metanormative Realism: How Best to Understand It, and How to Cope With It," *Philosophical Studies* 148(3): 413-438.

FitzPatrick, W. (2015). "Debunking Evolutionary Debunking of Ethical Realism," *Philosophical Studies* 172(4): 883-904.

Gould, S. J. (1989). *Wonderful Life: The Burgess Shale and the Nature of History*. New York: W. W. Norton & Company.

Hare, B., Wobber, V., and Wrangham, R. (2012). "The Self-Domestication Hypothesis: Evolution of Bonobo Psychology is Due to Selection Against Aggression," *Animal Behaviour* 83(3): 573-585.

Hauser, M. (2006). *Moral Minds: The Nature of Right and Wrong*. New York: HarperCollins.

Horn, J. (2017). "Evolution and the Epistemological Challenge to Moral Realism," in M. Ruse and R. Richards (eds.) *The Cambridge Handbook of Evolutionary Ethics*. Cambridge: Cambridge University Press.

Horowitz, S. (2014). "Epistemic Akrasia," *Nous* 48(4): 718-744.

Isaacs, Y. (2021). "The Fallacy of Calibrationism," *Philosophy and Phenomenological Research* 102(2): 247-260.

Joyce, R. (2006). *The Evolution of Morality*. Cambridge: MIT Press.

Joyce, R. (2019). "Moral and Epistemic Normativity: The Guilty and the Innocent," in C. Cowie and R. Rowland (eds.) *Companions in Guilt Arguments in Metaethics*. London: Routledge.

Kappel, K. (2019). "Escaping the Akratic Trilemma," in M. Skipper and A. Steglich-Petersen (eds.) *Higher-Order Evidence: New Essays*. Oxford: Oxford University Press.

Kelly, T. (2013). "Disagreement and the Burdens of Judgment," in D. Christensen and J. Lackey (eds.) *The Epistemology of Disagreement: New Essays*. Oxford: Oxford University Press.

Kitcher, P. (2011). *The Ethical Project*. Cambridge: Harvard University Press.

Koon, J. (2021). "The Epistemology of Evolutionary Debunking," *Synthese* 199(5-6): 12155-12176.

Korman, D. (2019). "Debunking Arguments," *Philosophy Compass* 14(12).

Lackey, J. (1999). "Testimonial Knowledge and Transmission," *Philosophical Quarterly* 49(197): 471-490.

Levy, A. and Levy, Y. (2020). "Evolutionary Debunking Arguments Meet Evolutionary Science," *Philosophy and Phenomenological Research* 100(3): 491-509.

Lord, E. (2014). "From Independence to Conciliationism: An Obituary," *Australasian Journal of Philosophy* 92(2): 365-377.

Lutz, M. (2018). "What Makes Evolution a Defeater?" *Erkenntnis* 83(6): 1105-1126.

Machery, E. and Mallon, R. (2010). "Evolution of Morality," in J. Doris (ed.) *The Moral Psychology Handbook*. Oxford: Oxford University Press.

Mercier, H. and Sperber, D. (2017). *The Enigma of Reason*. Cambridge: Harvard University Press.

Prufer, K., Munch, K., Helleman, I., Akagi, K., Miller, J., Walenz, B., Koren, S., Sutton, G., Kodira, C., Winer, R., Knight, J., Mullikin, J., Meader, S., Ponting, C., Lunter, G., Higashino, S., Hobolth, A., Marques-Bonet, T., Eichler, E., Andre, C., Atencia, R., Mugisha, L., Junhold, J., Patterson, N., Siebauer, M., Good, J., Fischer, A., Ptak, S., Lachmann, M., Symer, D., Mailund, T., Schierup, M., Andres, A., Kelso, J., and Paabo, S. (2012). "The Bonobo Genome Compared With the Chimpanzee and Human Genomes," *Nature* 486(7404): 527-531.

Richerson, P. and Boyd, R. (2005). *Not By Genes Alone: How Culture Transformed Human Evolution*. Chicago: The University of Chicago Press.

Ruse. M. (1986). "Evolutionary Ethics: A Phoenix Arisen," *Zygon* 21(1): 95-112.

Schoenfield, M. (2015). "A Dilemma for Calibrationism," *Philosophy and Phenomenological Research* 91(2): 425-455.

Schoenfield, M. (2018). "An Accuracy-Based Approach to Higher-Order Evidence," *Philosophy and Phenomenological Research* 96(3): 690-715.

Shafer-Landau, R. (2012). "Evolutionary Debunking, Moral Realism, and Moral Knowledge," *Journal*

of Ethics and Social Philosophy 7(1): 1-37.

Sinclair, N. (2018). "Belief Pills and the Possibility of Moral Epistemology," in R. Shafer-Landau (ed.) *Oxford Studies in Metaethics, Volume 13*. Oxford: Oxford University Press.

Skarsaune, K. (2011). "Darwin and Moral Realism: Survival of the Iffiest," *Philosophical Studies* 152(2):229-243.

Sliwa, P. and Horowitz, S. (2015). "Respecting *All* the Evidence," *Philosophical Studies* 172(11): 2835-2858.

Sterelny, K. (2021). *The Pleistocene Social Contract: Culture and Cooperation in Human Evolution*. Oxford: Oxford University Press.

Sterelny, K. and Fraser, B. (2017). "Evolution and Moral Realism," *British Journal for the Philosophy of Science* 68(4): 981-1006.

Street, S. (2006). "A Darwinian Dilemma for Realist Theories of Value," *Philosophical Studies* 127(1): 109-166.

Street, S. (2009a). "Evolution and the Normativity of Epistemic Reasons," *Canadian Journal of Philosophy* 39(S1): 213–248.

Street, S. (2009b). "In Defense of Future Tuesday Indifference: Ideally Coherent Eccentrics and the Contingency of What Matters," *Philosophical Issues* 19(1): 273-298.

Tomasello, M. (2014). *A Natural History of Human Thinking*. Cambridge: Harvard University Press.

Tomasello, M. (2016). *A Natural History of Human Morality*. Cambridge: Harvard University Press.

Tomasello, M. (2019). *Becoming Human: A Theory of Ontogeny*. Cambridge: Belknap Press.

Vavova, E. (2014). "Debunking Evolutionary Debunking," in R. Shafer-Landau (ed.) *Oxford Studies in Meta-Ethics, Volume 9*. Oxford: Oxford University Press.

Vavova. E. (2018). "Irrelevant Influences," *Philosophy and Phenomenological Research* 96(1): 134-152.

Vavova, E. (2021). "The Limits of Rational Belief Revision: A Dilemma for the Darwinian Debunker," *Nous* 55(3): 717-734.

Wielenberg, E. (2010). "On the Evolutionary Debunking of Morality," *Ethics* 120(3): 441-464.

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

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

Wilson, E. and Ruse, M. (1985). "The Evolution of Ethics," *New Scientist* 108(1478): 50-52.

Wrangham, R. and Peterson, D. (1996). *Demonic Males: Apes and the Origin of Human Violence*. New York: Houghton Mifflin Harcourt.

Yamakoshi, G. (2004). "Food Seasonality and Socioecology in *Pan*: Are West African Chimpanzees Another Bonobo?" *African Study Monographs* 25(1): 45-60.