A Bayesian Analysis of Debunking Arguments in Ethics

Shang Long Yeo

National University of Singapore, Australian National University

(Forthcoming in the *Philosophical Studies* – please cite the final version, thanks!)

Abstract: Debunking arguments in ethics contend that our moral beliefs have dubious evolutionary, cultural, or psychological origins – hence concluding that we should doubt such beliefs. Debates about debunking are often couched in coarse-grained terms – about whether our moral beliefs are justified or not, for instance. In this paper, I propose a more detailed Bayesian analysis of debunking arguments, which proceeds in the fine-grained framework of rational confidence. Such analysis promises several payoffs: it highlights how debunking arguments don't affect all agents, but rather only those agents who updated on their intuitions using a specific range of evidentiary weights; it underscores how the debunkers shouldn't conclude that we should reduce confidence beyond some threshold, but rather only that we should reduce confidence by some amount; and it proposes a method of integrating different kinds of evidence – about the kinds of epistemic flaws at play, about the different possible origins of our moral beliefs, about the background normative assumptions we're entitled to make – in order to arrive at a rational moral credence in light of debunking.

1. Introduction

We all have moral beliefs, and should change them upon learning relevant evidence. When we learn about the dubious origins of our moral beliefs, how, if ever, should we revise such beliefs? We might learn that our moral beliefs were strongly influenced by evolutionary, cultural, or psychological causes – and so had we evolved differently, been raised in a different culture, or been exposed to a thought experiment in a different condition, we would have very different moral intuitions and beliefs (or no such mental states at all).[1] Or we might learn that these causes strongly inculcated certain beliefs or intuitions in us – such that had the moral facts been different, we would still have judged the way we did.[2] Some philosophers

[1] See Street (2006, pp. 120–121), Joyce (2006, p. 181), Bogardus (2016, pp. 655–658), de Lazari-Radek and Singer (2012, pp. 21–28), Machery (2017, pp. 72–73, 81–83), Liao et al. (2012), Sinnott-Armstrong (2011), Horowitz (1998).

[2] See Morton (2016, pp. 242–243), Ruse and Wilson (1985, p. 52).

argue that learning these things should make us less confident in our moral beliefs – they press what's known as a *debunking argument* in the literature.[3]

Talk of learning *evidence* that then dictates a reduction in *confidence* naturally suggests adopting a Bayesian approach to analysing debunking – Bayesian epistemology being our leading theory of normative constraints on probabilistic belief. In this paper, I hope to propose a thorough Bayesian analysis that can help us better understand debunking arguments – in particular, one that will help clarify key mechanics, uncover crucial assumptions, and allow us to measure the strength of a debunking argument.[4]

My analysis has the following payoffs: First, I highlight an important but unnoticed condition for debunking to work – debunking arguments don't undermine everyone's confidence in the relevant moral beliefs; rather they only undermine confidence for agents who have previously updated on their intuitions using a certain evidentiary weight. Second, I argue that the debunkers shouldn't conclude that we should reduce confidence beyond some threshold like 0.5 (or that we should be agnostic about our moral beliefs), rather they should only conclude that we should reduce confidence by some amount. Third, I provide a method for determining how big this reduction should be, in light of different factors – concerning the different kinds of epistemic flaws that might be at play, the different possible origins of our moral beliefs, and the kinds of background normative assumptions we're entitled to make. I also argue that one variant of debunking (corresponding to sensitivity-based arguments) has a greater marginal impact on our credences than another (which corresponds to safety-based arguments). With this analysis, I hope to move the debunking debate from a coarse-grained description of epistemic impact – about whether our moral beliefs are justified or unjustified, for instance – to a finer-grained and more nuanced treatment, about when we should reduce confidence, and by how much.

This paper will proceed as follows: I introduce the Bayesian framework in section 2, and give reasons why it's well-suited for understanding debunking arguments. In section 3, I specify the possible hypotheses and evidence at play in the debunking debate. I present the Bayesian

[3] The debunking argument establishes an epistemic conclusion about what we should believe, not a metaphysical one about whether moral facts exist (Vavova, 2015, p. 105). Street (2006) thinks this epistemic conclusion only holds if we assume that the moral facts are mind-independent – if you agree with her, then you can see this paper as working out the Bayesian picture on this assumption.

[4] For other work using Bayesianism to analyse debunking, see Brosnan (2011), Goldman (2016; 2015), and O'Neill (2015), who cites Roush's (2007) truth-tracking approach. Philosophers might be reluctant to apply a probabilistic Bayesian framework to debunking because they see moral truths as necessary truths. However I believe that Bayesianism can still be profitably used to analyse our evidential (or epistemic) probabilities about morality, or just the rational degrees of belief to hold about morality, given our uncertainty about it.

model in section 4, draw out some implications and extensions in section 5, and conclude in section 6.

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

Bayesian epistemology is a normative theory of probabilistic belief – it concerns how we *should* revise our beliefs in light of the evidence. It takes agents to have doxastic attitudes called *credences* – these are modelled using numbers from 0 to 1, and indicate how confident an agent is in a proposition. For example, I might have 0.6 credence in the proposition "The patient has the disease". Bayesians propose a few normative constraints governing our credences. One core set governs how credences should behave at a time – these are Kolmogorov's probability axioms.[5] Another constraint governs how credences should evolve over time as the agent gains more evidence. Let h be the hypothesis that the agent is entertaining (for instance, that "The patient has the disease"), let e be the evidence that she gets (for instance, that "The patient returned a positive test result"), and assume (for now) that the evidence is learned for sure. Let $Cr_{new}(.)$ be the new, post-learning credence function, and $Cr_{old}(.)$ be the old credence function. Bayesians claim that the agent should update her credences according to Conditionalization, which says that

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

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

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

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

[5] For a precise statement, see Talbott (2016).

each update, she changes her credence in the evidence to 1, and then redistributes her credences proportionately to the hypothesis h and its alternatives.

To get clearer on how this works, let's look at an example of a doctor trying to decide whether a patient has a disease. (This is a stock example of Bayesian epistemology, but I believe it's quite analogous to debunking – and so is worth going through in detail.) We start off in time period 1 (denoted by a credence function with subscript 1), where the doctor, informed by frequency data from the population, has some prior credences in hypotheses about the patient's condition. Let Cr_1(*disease*) be the doctor's credence in the hypothesis that "The patient has the disease", and Cr_1(~*disease*) be the doctor's credence in the alternative hypothesis. Let's say that disease is quite rare, so the doctor thinks it's more likely that the patient doesn't have the disease. We might assign the values for their credences as

$$Cr_1(disease) = 0.2$$

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

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

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

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

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

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

Substituting our values for Cr$_1$(*disease*), Cr$_1$(*positive*|*disease*), and Cr$_1$(*positive*|*~disease*), we have

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

After receiving the positive test result, the doctor should adopt a credence of 0.43 in the hypothesis that the patient has the disease. This is an increase from Cr$_1$(*disease*), which was 0.2.

This example shows how Bayesians might prescribe changes in credences, given certain values for the priors and likelihoods. But Bayesians might impose further constraints on these values too. For instance, Sober (2008, pp. 24–30) argues that scientists can sometimes use frequency data to inform their likelihoods too. In our example, the likelihood Cr$_1$(*positive*|*disease*) could be informed by giving the test to people whom we already know have the disease, and seeing how many positive results are obtained.

These Bayesian constraints are often interpreted as governing how an agent's credences should change over time. But there's another interpretation that fits better with my project. Lange (1999) argues that Bayesian norms can also govern the steps in the arguments by which our current credences are justified. He gives the example of a forecaster who wants to argue that her 0.8 credence in 'It will rain and not snow tomorrow' is justified. She starts with a prior probability distribution over this hypothesis and its alternatives – a distribution that others will grant her as an argumentative 'free move' that needs no further justification. She then introduces the first piece of evidence – say, that 'Today's barometric pressure is 29 torr' – and 'updates' her initial distribution to an intermediate conclusion (another probability distribution) via Conditionalization. She continues introducing further evidence she has, 'updating' in this way until she has considered all relevant evidence. If her justificatory argument works, then she should arrive at a distribution assigning 0.8 credence to 'It will rain and not snow tomorrow'. If not, then something has gone wrong, and she's not justified in assigning 0.8 credence to the hypothesis (Lange, 1999, pp. 302–306).

On this interpretation, her actual credences don't change as she considers each new piece of evidence. Rather, each intermediate probability distribution represents an intermediate conclusion on the way to justifying her actual credence. Furthermore, the priors represent an argumentative 'free move' that needs no further justification, while the likelihoods represent the evidentiary weight she believes should be accorded to each piece of evidence. Lange used this story to account for the confirmation of scientific theories, but it can be usefully applied to philosophical argumentation too. The debunkers' opponents cite some intuitions, judgments, or beliefs as evidence to justify their credence in moral hypotheses – much like how the forecaster cites the barometer reading as evidence for their prediction. In response, the debunkers argue that given some debunking story, these mental states shouldn't be accorded much evidential weight, if any – so their opponents are unjustified in assigning the credence they do. This is analogous to a forecaster who realizes that a barometer is faulty – arguing that we shouldn't give much evidentiary weight to its readings, and that anyone who did so would end up with an unjustified credence. The Bayesian picture thus seems well-equipped to model the debate between the debunkers and their opponents, and I hope to demonstrate its benefits below. In the next section, I start by interpreting the debate in Bayesian terms.

3. Debunking in Bayesian Terms

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

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

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

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

Here I cannot remain neutral. First, treating the evidence as everyone's believing that h shortchanges the debunking argument. The debunkers don't just want to undermine the support rendered by widespread agreement of others – they also want to undermine the support provided by an individual's own judgments, beliefs, or intuitions. After all, they still want to debunk our moral beliefs in cases where people disagree about moral matters.

Secondly, we might take the evidence to be your believing that h. There are issues with this. Normatively speaking, we generally shouldn't treat our beliefs as evidence, since that seems like double-counting. Moreover, as a descriptive matter, we don't usually treat our beliefs as evidence either. As White (2010, p. 585) argues, we don't take our belief that p as further evidence that p, and increase our confidence as a result. But perhaps things are different in the moral case: we might have been handed down certain beliefs, and just accepted that there was some good argument supporting these beliefs, even if we didn't know what that argument was. While I have some sympathy for this line, I think modelling the situation using our intuitions as evidence (see below) is the least controversial route. I believe, however, that the same Bayesian machinery I propose can be fitted to using beliefs as evidence too (see n.8).

Third, we can take our intuitions to be the evidence.[6] We often make the following inference: "I have the intuition that h. This is evidence that h, and I should increase my credence in h." We think we should update our credence in this way – and, as a descriptive matter, we do sometimes update like this. So I believe the Bayesian model should take your having certain intuitions as the evidence. We can be neutral about what exactly intuitions are – they could be dispositions to believe, sui generis mental states, etc. What's crucial here is that these intuitions are separate from the final credence in h – in the sense that our final credence in h could be low, even when we have the intuition that h.

4. The Bayesian Model of Debunking

I'll now present the Bayesian model, which involves three stages. Stage 1 defines the prior credence in the moral hypothesis and its alternatives, and some of the likelihoods. In stage 2, the debunker's opponents introduce intuitions as evidence. Updating on these intuitions increases our credence in the moral hypothesis. In stage 3, the debunker introduces a

[6] See Climenhaga (2018) who argues that philosophers do use intuitions as evidence.

debunking story about these intuitions. This changes the evidentiary weight accorded to the intuitions, mandating a reduction from the stage 2 credence in the moral hypothesis.

Stage 1

I will model debunking concerning a moral hypothesis h – keeping in mind that h could stand for any kind of moral claim; for instance, it might range from a specific claim like "It is morally impermissible to turn the trolley in the Loop case", to a more general claim like "There can be no moral difference between two actions without there also being some non-moral difference between them". The evidence for h would be your having the intuition that h.

Remember the priors represent an argumentative 'free move' – a probability distribution over the hypothesis and its alternatives, which both the debunkers and their opponents agree needs no further justification. What might this distribution look like? One plausible answer is that h is as likely to be true as it is to be false:[7]

$$Cr_1(h) = 0.5$$

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

Next, the likelihoods, which represent the evidentiary weight we believe should be accorded to your having the intuition that h. Let $Cr_1(int|h)$ be the likelihood that you have the intuition that h, given that h is true, and $Cr_1(int|\sim h)$ be the likelihood that you have the intuition that h, given h is false.[8] If we grant for now that intuitions are good indicators of the truth, we should accord them significant evidentiary weight. This is represented by a high $Cr_1(int|h)$ and a low $Cr_1(int|\sim h)$.[9]

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

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

These likelihoods are analogous to the doctor's beliefs about the true positive rate and the false positive rate of the test results. We can likewise see our moral intuitions as test results

[7] This follows the principle of indifference: "Given n mutually exclusive and jointly exhaustive possibilities, none of which is favored over the others by the available evidence, the probability of each is 1/n." (Weisberg, 2017, sec. 2.1)

[8] If we took beliefs as evidence, these likelihoods represent the probability that we would believe as we do, given that the moral hypothesis is true or not. Then we are rather working backward to see if there are good arguments supporting our beliefs. Thanks here to Katie Steele.

[9] I used exact values to illustrate the model, but we shouldn't take them too seriously. We could instead also perform a robustness analysis, over a range of different possible values, in more specific cases.

that indicate something about the moral hypothesis – these intuitions likewise yield true and false positives at some rate.[10]

Stage 2

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

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

Substituting our values for Cr_1(h), Cr_1(~h), Cr_1(int|h) and Cr_1(int|~h), we have

$$
\begin{aligned}
&Cr_2(h)\\
&= \frac{0.9 \times 0.5}{(0.9 \times 0.5) + (0.1 \times 0.5)}\\
&= 0.9
\end{aligned}
$$

Call Cr_2(h) the *evidential credence*.[11] It represents the confidence we should have in the moral hypothesis, after updating on having the relevant intuition. Just like how the doctor takes a positive test result as indication of the disease, the moral philosopher takes their having the intuition as evidence for the moral hypothesis. We should adopt an evidential credence of 0.9, an increase from the prior credence of 0.5.

Stage 3

The debunker then introduces a causal story about the intuitions that were treated as evidence – call this *story*.[12] For example, the debunker might allege that these intuitions were

[10] See O'Neill (2015) and Sauer (2018, pp. 38–41), who talk about some debunking arguments in the language of true and false positives.

[11] This terminology comes from Kotzen (forthcoming), who talks about credences in defeat.

[12] I'll assume we learn *story* with certainty, but this assumption could be relaxed, given that the current evidence might be unable to adjudicate between debunking and non-debunking evolutionary genealogies (Isserow, 2018, secs. 3–5). Jeffrey conditionalization should then be used – this tells us how to update our credence in the hypothesis when our credences in the evidential statements change, but are not raised to certainty. Even more

produced by evolutionary processes aimed at promoting reproductive success, rather than at tracking the moral truths. This is analogous to the doctor discovering that their tests are unreliable. I'll flesh out more details of the debunking story soon – for now, notice that in stage 3, we have two pieces of evidence, *int* and *story*. So the stage 3 credence in h should be

$$\begin{aligned} &Cr_3(h) \\ &= Cr_1(h|int\ \&\ story) \end{aligned}$$

Call $Cr_3(h)$ the *debunked credence* – it represents how confident we should be in h, after learning both the intuition and the debunking story. To relate this debunked credence to easily interpretable terms, I'll expand it in a slightly different way:

$$\begin{aligned} &Cr_3(h) \\ &= Cr_1(h|int\ \&\ story) \\ &= \frac{Cr_1(h\ \&\ int\ \&\ story)}{Cr_1(int\ \&\ story)} \\ &= \frac{Cr_1(h\ \&\ int\ \&\ story)}{Cr_1(int\ \&\ story\ \&\ h) +\ Cr_1(int\ \&\ story\ \&\ {\sim}h)} \\ &= \frac{Cr_1\left(int\,\middle|\,story\ \&\ h\right).Cr_1(h|story).Cr_1(story)}{Cr_1\left(int\,\middle|\,story\ \&\ h\right).Cr_1(h|story).Cr_1(story) +\ Cr_1\left(int\,\middle|\,story\ \&{\sim}h\right).Cr_1({\sim}h|story).Cr_1(story)} \\ &= \frac{Cr_1\left(int\,\middle|\,story\ \&\ h\right).Cr_1(h|story)}{Cr_1\left(int\,\middle|\,story\ \&\ h\right).Cr_1(h|story) +\ Cr_1\left(int\,\middle|\,story\ \&{\sim}h\right).Cr_1({\sim}h|story)} \end{aligned}$$

Line 3 follows from the Ratio Formula. Line 4 expands the denominator Cr_1(int&story) into the sum of credences in the two possibilities, Cr_1(int&story&h) and Cr_1(int&story&~h). Line 5 expands all terms in the numerator and denominator using the chain rule for conditional credences.[13] Line 6 factors out Cr_1(story) from the numerator and denominator – arriving at a formula for the debunked credence that uses the priors and some new conditional credences, Cr_1(int|story&h), Cr_1(int|story&~h), Cr_1(h|story), and Cr_1(~h|story).[14]

The debunker constrains these new conditional credences – this then affects the value of the debunked credence. Let's start with the latter two, Cr_1(h|story) and Cr_1(~h|story). These represent the credence we should adopt in h and ~h respectively, given that we learn only

interestingly, our credences could also be split between different debunking stories that impact the evidential weight of our intuitions differently.

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

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

about *story*, but not about having the relevant intuitions. It seems that learning *story* should only affect our credence in h *through* undermining the intuitions that were used as evidence. If we just learned *story* alone, however, this shouldn't affect our credence in the moral hypothesis at all. This is analogous to learning about the reliability of a medical test instrument, *before* we have learned about any results from that instrument – here, it seems like our opinions about the patient shouldn't change.[15] Similarly, in the debunking case, the following two conditions should apply

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

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

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

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

We can compare this with the stage 2 evidential credence,

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

Notice that the likelihoods in these two formulas have changed. The stage 2 evidential credence was computed with $Cr_1(int|h)$ and $Cr_1(int|{\sim}h)$; in the stage 3 debunked credence, these are replaced by $Cr_1(int|story\&h)$ and $Cr_1(int|story\&{\sim}h)$ respectively. The debunker can be interpreted as constraining $Cr_1(int|story\&h)$ and $Cr_1(int|story\&{\sim}h)$, in order to achieve a reduction in credence as we move from stage 2 to 3. To understand how this works, we have to look at the details of two prominent debunking stories.

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

Two Debunking Stories

The first debunking story concerns the contingency of our having certain intuitions. This story says we could easily have had different moral intuitions, even when the moral facts remained the same. For example, Joyce (2006, 2016) argues that our moral intuitions are evolutionarily contingent – we could easily have a different evolutionary history, leading to different moral intuitions, even though the moral facts remained the same. Challenges concerning demographic variation in our moral intuitions, or variation across different frames or conditions of a thought experiment (Liao et al., 2012; Machery, 2017, Chapters 2–3; Sinnott-Armstrong, 2011) work in the same way, but at a different temporal scale: our moral intuitions vary across different experimental conditions (or across demographic variables like culture), yet the moral facts remain the same. These challenges all highlight the low rate of true positives, given some debunking story – that is, alleging the low probability of your having the intuition that h, given that h and story are true. Call this kind of epistemic impact *true positive less likely*. This can be modelled in our example by lowering $Cr_1(int|story\&h)$ to 0.5 – as compared to $Cr_1(int|h)$, which was 0.9. Assuming the other values remain the same, constraining $Cr_1(int|story\&h)$ to 0.5 means we should adopt a debunked credence of 0.83 in h – a reduction from stage 2's evidential credence of 0.9 in h.[16] By showing that the true positive rate for our intuitions is lower than what we initially thought, the debunker achieves a reduction in credence in h from stage 2.[17] More generally, for *true positive less likely* to work, the debunkers need to ensure that

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

The second debunking story involves the inevitability of one's intuitions. Here, the idea is that we would still have the same intuitions, even when the moral facts are changed. For example, Morton (2016, p. 242) argues that evolutionary processes would have made us believe that categorical moral reasons exist, regardless of whether or not such reasons actually do exist. This challenge alleges the high rate of *false* positives, given some debunking story – in

[16] The 0.5 value for the constrained $Cr_1(int|story\&h)$ was chosen purely out of convenience. I believe its actual value will vary from case to case, depending on the strength of the debunking argument under consideration. I believe the relevant empirical evidence (for instance, about the counterfactual robustness of some evolutionarily-influenced moral intuition) will provide the basis for setting this value. To be clear, I am not suggesting that the principle of indifference helps us set this value – though this principle might be used to model some other challenges, such as the purely odds-based argument suggested by Street (2006, p. 122) and formalised by Shafer-Landau (2012, pp. 10–11). The above points will also apply to my modelling of constraints on $Cr_1(int|story\&{\sim}h)$ below. Thanks here to an anonymous reviewer.

[17] This can be used to model debunking involving contingency or lack of safety (Bogardus, 2016, pp. 645–647; Handfield, 2016, p. 68; Joyce, 2006, p. 181), and counterfactual disagreement (Bogardus, 2016, pp. 655–657).

particular, it highlights the high probability that you have the intuition that h, given that h is *false* and story is true. Call this kind of epistemic impact *false positive more likely*. This can be modelled in our example as constraining Cr_1(int|story&~h) to 0.5 – as compared to Cr_1(int|~h), which was 0.1. If all the other values are unchanged, constraining Cr_1(int|story&~h) to 0.5 would mean we should adopt a debunked credence of 0.64 in h – a significant reduction from the 0.9 evidential credence in stage 2. By showing that the false positive rate for our intuitions is higher than initially thought, the debunkers also achieve a reduction of credence from stage 2.[18] Generally, for *false positive more likely* to work, the debunkers need to show that

$$Cr_1\big(int \,\big|\, story \,\&\, {\sim}h\big) > \; Cr_1(int|{\sim}h)$$

Either of the above two constraints will suffice for reducing confidence in h as we move from stage 2 to 3. Notice that the model doesn't commit us to saying that the debunked credence must go below some specific threshold (like 0.5), as many debunking arguments seem to conclude.[19] Instead it just says we should reduce confidence by some amount, and leaves open how large this reduction should be.[20]

Moreover, the formula for the debunked credence allows us to quantitatively estimate the impact of debunking, which can come in useful. First, some theorists argue that we should resolve conflicts of intuition by comparing the best debunking argument for each opposing intuition (McPherson, 2014) – this formula gives us a precise way of doing so. Second, the formula can estimate the *joint* impact of *false positive more likely* and *true positive less likely* – this represents an advance, since they are usually only considered separately (for instance, sensitivity-based debunking arguments are run separately from safety-based ones).

5. Implications and Extensions of the Bayesian Model

Let's now consider some implications and extensions.

[18] This can be used to model debunking arguments involving insensitivity (Bogardus, 2016, pp. 638–640; Morton, 2016, pp. 235, 242–243; Ruse & Wilson, 1986, pp. 186–187), and the screening-off of evidence (White, 2010, pp. 580–581). Also see Climenhaga (2018, n. 26) who uses this to model the practice of offering error theories in philosophy.

[19] Joyce argues that the evolutionary evidence renders our moral beliefs "unjustified" (2006, p. 180), that we should "cultivate agnosticism" (2006, p. 181) until we get further evidence. Street (2006, p. 125) argues that our moral beliefs are "likely to be false". Machery (2017, p. 95) argues that when we find demographic and presentation effects in philosophical cases, we "ought to suspend judgment".

[20] Joyce (2016, p. 125) later argues that confidence in our moral beliefs should be "dented".

Debunking effect only conditional on a specific updating history

The debunkers' constraints mean that we should reduce credence in the moral hypothesis as we move from stage 2 to 3. For this debunking effect to happen, at least one of the following conditions must hold:

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

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

Notice that these conditions depend crucially on the values of Cr$_1$(int|h) and Cr$_1$(int|~h) that were used for updating in stage 2. For instance, upon learning about the debunking story, you should reduce credence in the moral hypothesis only if you previously updated with a Cr$_1$(int|h) that was higher than the constrained Cr$_1$(int|story&h). That is, you should reduce credence in the moral hypothesis only if you updated on the intuitions thinking that the true positive rate was higher than it actually is. And vice versa with the false positive rate – you should only reduce confidence in h if you updated thinking that the false positive rate was lower than it actually is. This highlights the essentially historical nature of debunking arguments – such arguments don't dictate that *every* agent should reduce confidence in h. Rather, only those who updated on their intuitions using a specific range of evidentiary weights are affected (as set out in the above inequalities). If someone had previously updated using different weights – for instance, if they updated using a Cr$_1$(int|h) that was equal to Cr$_1$(int|story&h), and Cr$_1$(int|~h) equal to Cr$_1$(int|story&~h) – the debunker's constraints shouldn't affect their credence at all.[21] The Bayesian model makes clear that the debunking argument assumes a descriptive claim: it assumes that we have previously updated on our intuitions using a specific range of evidentiary weights.

Recognising this essentially historical nature also reveals how constraints on Cr$_1$(int|story&h) and Cr$_1$(int|story&~h) could have a perverse effect. Suppose someone initially updated their beliefs with a very low evidentiary weight on their intuitions – taking Cr$_1$(int|h) to be 0.001, for instance. The debunker then introduces *story* as evidence, constraining Cr$_1$(int|story&h) to be 0.5. This person should now give their intuitions *more* evidentiary weight than before, since true positives are more likely than they initially thought. Other things being equal, the debunker's constraint dictates that this person should *increase* credence in the moral

[21] Also see Climenhaga (2018, p. 86), who argues that if a philosopher didn't take intuitions as evidence, they shouldn't be affected by debunking explanations of those intuitions.

hypothesis.[22] This starkly illustrates how the debunking argument's impact depends on what evidentiary weight we initially assigned to our intuitions.

Two Special Cases

Two special cases of epistemic impact are also worth highlighting. First, consider when the likelihoods are constrained so that

$$Cr_1(int \,|\, story \mathbin{\&} h) = Cr_1(int|story\&{\sim}h)$$

That is, it's equally likely that you have the intuition that h, given that h and *story* are both true, as it is given that h is *false* and *story* is true. Here, the intuition is no guide to the moral hypothesis at all, and we should adopt a debunked credence that's *equal* to our prior credence.[23] Let $Cr_1(\text{int}|\text{story\&h}) = Cr_1(\text{int}|\text{story\&\textasciitilde h}) = x$. Then the debunked credence is

$$\begin{aligned} &Cr_3(h) \\ &= \frac{Cr_1(int \,|\, story \mathbin{\&} h).Cr_1(h)}{Cr_1(int \,|\, story \mathbin{\&} h).Cr_1(h) + Cr_1(int \,|\, story \mathbin{\&}{\sim}h).Cr_1({\sim}h)} \\ &= \frac{0.5x}{0.5x + 0.5x} \\ &= 0.5 = Cr_1(h) \end{aligned}$$

The second case involves constraints such that

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

That is, it's more likely that you have the intuition that h, given h is *false* and *story* is true, than it is that you have the intuition, given that h and *story* are both true. This constraint turns the intuition into an anti-reliable indicator – it is evidence for ~h! Consequently, the debunked credence could go *below* the prior credence. For instance, let $Cr_1(\text{int}|\text{story\&h}) =$

[22] This might happen when we discover surprising agreement or robustness in our intuitions, as presented by Knobe (2019).

[23] The calculation in this section pertains to the case where $Cr_1(h) = Cr_1(\sim h) = 0.5$, but the result – of the debunked credence being equal to prior credence – holds more generally. It obtains as long as $Cr_1(h) = 1 - Cr_1(\sim h)$. Here's the proof: Let $Cr_1(h) = p$. So $Cr_1(\sim h) = 1-p$. We can then compute the debunked credence, $Cr_3(h) = xp / [xp + x(1-p)] = xp / [xp + x - xp] = p = Cr_1(h)$. Thanks here to an anonymous reviewer.

0.3, $Cr_1(int|story\&{\sim}h)$ =0.8, then the debunked credence in h will be 0.27 (a great reduction from the prior credence of 0.5).

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

Brosnan's (2011, p. 54) framework implies that this as a tracking success – since our moral intuitions are indicators, albeit anti-reliable ones, of the moral facts. I beg to differ – since it would be cold comfort to learn that we should conclude exactly the opposite of what our moral intuitions tell us.[24]

Marginal Effects of Constraining False Positive vs. True Positive Rate

The model also has implications for the relative size of the two kinds of epistemic impact. It seems that point for point, *false positive more likely* has greater marginal impact on the debunked credence in h than *true positive less likely*. As an illustration, recall the initial examples from section 4. The example of *false positive more likely* – where the false positive rate of our intuitions was shown to be higher than we thought – constrained $Cr_1(int|story\&{\sim}h)$ to a value that's 0.4 greater than $Cr_1(int|{\sim}h)$, a 0.4 increase in the false positive rate. This lowered the debunked credence to 0.64. On the other hand, consider the example of *true positive less likely*, where the true positive rate of our intuitions was shown to be lower than initially thought. This constrained $Cr_1(int|story\&h)$ to a value that's 0.4 lower than $Cr_1(int|h)$ – a 0.4 decrease in the true positive rate. This only lowered the debunked credence to 0.825.

This difference in marginal impact – where an instance of *false positive more likely* has greater impact on our credence in the moral hypothesis than an equal-sized instance of *true positive less likely* – is robust across all values that are relevant for the debunking debate. As long as $Cr_1(int|h) > Cr_1(int|{\sim}h)$ – that is, as long as the intuition is taken as some positive indication of the hypothesis – *false positive more likely* will have a greater marginal impact than *true*

[24] I'm unaware of any real-world intuitions that would correspond to this case, but a hypothetical example would be learning that our moral intuitions were created by a powerful evil demon who is a moral expert, but who also wishes to instill non-veridical intuitions in us. Thanks here to an anonymous reviewer.

positive less likely (see Appendix for proof). Put in more traditional epistemological terms, the Bayesian framework shows how sensitivity-based debunking arguments (which correspond to the impact of *false positive more likely*) have a greater effect on our moral views than safety-based ones (which correspond to *true positive less likely*).

The Role of Priors

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

This is correct, but I wonder if it's a strong reply. Firstly, recall that we're interpreting the priors as an argumentative 'free move' that is allowed by both sides of the debate. The debunkers wouldn't allow the 'free move' of assigning a high prior credence to the moral hypothesis. Even if they did, all the work of defending against debunking is done by this assignment, rather than by any substantive argument. Leaning on the priors thus merely stipulates the problem of debunking away, rather than giving a positive answer to it. Secondly, even if the debunkers don't show that our moral beliefs are likely false, they might still secure some weaker epistemic impact – like concluding that we should reduce confidence in the hypothesis. Finally, just constraining the likelihoods still counts as a significant achievement for the debunkers. This undermines at least some of the evidence cited by their opponents, and constrains the range of possible prior credences that could still vindicate our moral beliefs (that is, by still leading to a high debunked credence).

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

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

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

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

These, recall, concern how we shouldn't change our credence in the moral hypothesis h if we just learned about the debunking *story* alone, without learning about the intuition. Opponents of debunking will contend, however, that these conditions don't hold. Brosnan

(2011, pp. 43, 59–61) argues that if we make some background normative assumptions, then some non-moral facts obtaining might raise the probability of some moral facts obtaining. So it could be that $Cr_1(h|story) \neq Cr_1(h)$; likewise for $Cr_1(\sim h|story)$ and $Cr_1(\sim h)$. For example, if we assume that wellbeing is generally morally good, then it might be that a belief's being reproductively advantageous would raise the probability of its being true. This is because what promotes reproductive fitness will also generally promote wellbeing, which is itself a morally good thing. Responses along these lines are called *third-factor accounts* – they posit a third factor that explains how our evolutionarily-produced moral beliefs are correlated with the moral facts. To list two other examples: Enoch (2010, pp. 430–432) argues that evolutionary causes tend to produce beliefs that promote survival, and survival tends to be a morally good thing. Wielenberg (2010, pp. 449–450) contends that for an organism to form moral beliefs about rights, it must have sufficiently sophisticated cognitive faculties to do so. But if an organism had such faculties, it would possess rights of its own.

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

Proximate and Ultimate Causes

Debunking stories could also focus on different kinds of causes that contributed to producing our moral intuitions. The proximate causes operate within our lifetimes (like the immediate psychological mechanisms that process the framing of a case) while the ultimate causes operate outside our lifetimes (such as natural selection operating over many generations). How might information about these causes contribute to the final moral credence? O'Neill (2015) argues that when it comes to the reliability of our moral beliefs, information from the proximate causes has priority over information from ultimate causes.[25] This is because if we can tell whether our moral beliefs are reliable just from information about the proximate causes, then information about the ultimate causes tells us nothing more. I find her arguments convincing – and if we accept her conclusion, we get further direction on how to use information about these causes in our computation of the debunked credence.

In the Bayesian framework, we can interpret O'Neill as concluding that information about proximate causes *screens off* information about ultimate causes, when it comes to the reliability of our moral beliefs. This places further constraints on the relevant likelihoods. Suppose now we have two different debunking stories about a moral intuition – a *proximate story* that pertains to proximate causes of that intuition, and an *ultimate story*, which pertains to its ultimate causes. O'Neill's arguments can be read as constraining our likelihoods such that:

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

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

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

6. Conclusion

In this paper, I presented a Bayesian model of the debunking debate. The model first involves updating on the intuition that h, yielding an increased evidential credence. The debunkers then introduce a debunking story. Updating on both the intuition and the story yields the

[25] O'Neill (2015, n. 1) defines proximate and ultimate causes in a different way – but this won't matter here.

debunked credence – which is computed using some new likelihoods, $Cr_1(int|story\&h)$ and $Cr_1(int|story\&{\sim}h)$. The debunker hopes to constrain these likelihoods, such that the debunked credence is lower than evidential credence. This happens when at least one of the following two conditions hold:

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

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

The first condition leads to the impact of *true positive less likely* – the true positive rate of our intuitions is revealed to be lower than we initially thought. The second condition leads to the impact of *false positive more likely* – the false positive rate is revealed to be higher than we thought. This model has important implications: firstly, that the debunking effect is only conditional on the agent's having a specific update history – the debunker shouldn't argue that every agent should reduce confidence. Secondly, the debunker is only licensed to conclude that we should reduce confidence by some amount, and not that we should reduce beyond some threshold. Third, we can quantitatively integrate evidence about the two kinds of epistemic impact, about legitimate background assumptions, about the different possible origins of our moral beliefs, and about the different kinds of causes of our moral beliefs, in order to arrive at a final rational moral credence. I also argued that point-for-point, the epistemic impact of *false positive more likely* has greater marginal effect on our credence in h than *true positive less likely*.

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

Acknowledgements Many thanks to Alan Hájek, Katie Steele, Christian Barry, Kim Sterelny, Toby Handfield, Ramon Das, Jeremy Strasser, James Willoughby, Edouard Machery, Joshua Knobe, Jessica Isserow, Neil Sinhababu, Weng Hong Tang, and Joshua Thong for the awesome discussions and feedback. I'm also very grateful to two anonymous reviewers for *Philosophical Studies*, and an anonymous reviewer for another journal, for their helpful and insightful comments. This research was supported by an Australian Government Research Training Program (RTP) Scholarship.

Appendix

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

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

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

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

The marginal impact of the two kinds of debunking can be found by partially differentiating $Cr_2(h)$ with respect to T and F respectively. Partial differentiation is useful here because it helps us compute the rate of change of one variable with respect to another, while holding all other variables constant.

To see how this is useful, first consider the marginal impact of *true positive less likely*. This impact can be found by first partially differentiating $Cr_2(h)$ with respect to T (the true positive rate). The resulting derivative represents how much $Cr_2(h)$ would change if we increased the true positive rate by a small amount, while holding all other variables (the priors and the false positive rate) constant. We then negate this result, because *true positive less likely* involves a *decrease* in the true positive rate:

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

Now consider *false positive more likely*. We simply partially $Cr_2(h)$ with respect to F (the false positive rate) – the result tells us how much $Cr_2(h)$ would change if we increased the false positive rate by a small amount, while holding all other variables constant. There is no need to negate the result like before, since *false positive more likely* involves an increase in the false positive rate:

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

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

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

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

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

$$TH(1-H) > FH(1-H) \qquad \textbf{Multiply by denominator. (Assume T, H > 0)}$$

$$T > F \qquad \textbf{Divide by } \boldsymbol{H(1-H)}\textbf{. (Assume 0 < H < 1)}$$

Thus when the initial true positive rate is greater than the initial false positive rate, *false positive more likely* has a greater marginal impact.

References

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

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

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

Climenhaga, N. (2018). Intuitions are Used as Evidence in Philosophy. *Mind*, *127*(505), 69–104. https://doi.org/10.1093/mind/fzw032 de Lazari-Radek, K., & Singer, P. (2012). The Objectivity of Ethics and the Unity of Practical Reason. *Ethics*, *123*(1), 9–31. https://doi.org/10.1086/667837

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.

Goldman, A. (2016). Reply to Schaffer. In B. P. McLaughlin & H. Kornblith (Eds.), *Goldman and His Critics* (1 edition, pp. 365–368). Wiley-Blackwell.

Goldman, A. I. (2015). Naturalizing metaphysics with the help of cognitive science. In K. Bennett & D. W. Zimmerman (Eds.), *Naturalizing metaphysics with the help of cognitive science* (Vol. 9). Oxford University Press. https://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198729242.001.0001/acprof-9780198729242-chapter-8

Handfield, T. (2016). Genealogical Explanations of Chance and Morals. In U. D. Leibowitz & N. Sinclair (Eds.), *Explanation in Ethics and Mathematics: Debunking and Dispensability*. Oxford University Press UK.

Horowitz, T. (1998). Philosophical Intuitions and Psychological Theory. *Ethics*, *108*(2), 367–385. https://doi.org/10.1086/233809

Huemer, M. (2005). *Ethical Intuitionism*. Palgrave Macmillan.

Isserow, J. (2018). Evolutionary Hypotheses and Moral Skepticism. *Erkenntnis*, 1–21. https://doi.org/10.1007/s10670-018-9993-8

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

Joyce, R. (2016). Reply: Confessions of a Modest Debunker. In U. D. Leibowitz & N. Sinclair (Eds.), *Explanation in Ethics and Mathematics: Debunking and Dispensability*. Oxford University Press UK.

Kahane, G. (2011). Evolutionary Debunking Arguments. *Noûs*, *45*(1), 103–125. https://doi.org/10.1111/j.1468-0068.2010.00770.x

Knobe, J. (2019). Philosophical Intuitions Are Surprisingly Robust Across Demographic Differences. *Epistemology and Philosophy of Science*, *56*, 29–36.

Kotzen, M. (forthcoming). A Formal Account of Epistemic Defeat. *Synthese Library*. http://matthewkotzen.net/matthewkotzen.net/Research_files/Klein.pdf

Lange, M. (1999). Calibration and the Epistemological Role of Bayesian Conditionalization. *The Journal of Philosophy*, *96*(6), 294–324. JSTOR. https://doi.org/10.2307/2564680

Liao, S. M., Wiegmann, A., Alexander, J., & Vong, G. (2012). Putting the trolley in order: Experimental philosophy and the loop case. *Philosophical Psychology*, *25*(5), 661–671. https://doi.org/10.1080/09515089.2011.627536

Lutz, M. (2018). What Makes Evolution a Defeater? *Erkenntnis*, *83*(6), 1105–1126. https://doi.org/10.1007/s10670-017-9931-1

Machery, E. (2017). *Philosophy Within Its Proper Bounds* (1 edition). Oxford University Press.

McGrath, S. (2014). Relax? Don't Do It! Why Moral Realism Won't Come Cheap. In R. Shafer-Landau (Ed.), *Oxford Studies in Metaethics: Volume 9* (1 edition, pp. 186–213). Oxford University Press.

McPherson, T. (2014). A Case for Ethical Veganism. *Journal of Moral Philosophy*, *11*(6), 677–703. https://doi.org/10.1163/17455243-4681041

Morton, J. (2016). A New Evolutionary Debunking Argument Against Moral Realism. *Journal of the American Philosophical Association*, *2*(2), 233–253. https://doi.org/10.1017/apa.2016.14

O'Neill, E. (2015). Which Causes of Moral Beliefs Matter? *Philosophy of Science*, *82*(5), 1070–1080. https://doi.org/10.1086/683441

Parfit, D. (1986). *Reasons and persons*. OUP Oxford.

Roush, S. (2007). *Tracking Truth: Knowledge, Evidence, and Science* (1 edition). Oxford University Press.

Ruse, M., & Wilson, E. O. (1985). The evolution of ethics. *New Scientist*, *108*(1478), 50–52.

Ruse, M., & Wilson, E. O. (1986). Moral Philosophy as Applied Science. *Philosophy*, *61*(236), 173–192.

Sauer, H. (2018). *Debunking Arguments in Ethics*. Cambridge University Press.

Shafer-Landau, R. (2012). Evolutionary Debunking, Moral Realism, and Moral Knowledge. *Journal of Ethics and Social Philosophy*, *7*(1).

Sinnott-Armstrong, W. (2011). Emotion and Reliability in Moral Psychology. *Emotion Review*, *3*(3), 288–289. https://doi.org/10.1177/1754073911402382

Sober, E. (2008). *Evidence and Evolution: The Logic Behind the Science*. Cambridge University Press.

Street, S. (2006). A Darwinian Dilemma for Realist Theories of Value. *Philosophical Studies*, *127*(1), 109–166. https://doi.org/10.1007/s11098-005-1726-6

Talbott, W. (2016). Bayesian Epistemology Supplement—Probability Laws. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/entries/epistemology-bayesian/supplement1.html

Vavova, K. (2015). Evolutionary Debunking of Moral Realism. *Philosophy Compass*, *10*(2), 104–116. https://doi.org/10.1111/phc3.12194

Weisberg, J. (2017). Formal Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2017). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2017/entries/formal-epistemology/

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

Wielenberg, E. J. (2010). On the Evolutionary Debunking of Morality. *Ethics*, *120*(3), 441–464.