Defusing the Regress Challenge to Debunking Arguments

Shang Long Yeo

Australian National University, National University of Singapore

(Forthcoming in the *Canadian Journal of Philosophy* – please cite the final version, thanks!)

Abstract: A debunking argument contends that some target moral judgments were produced by unreliable processes and concludes that such judgments are unjustified. Debunking arguments face a regress challenge: to show that a process is unreliable at tracking the moral truth, we need to rely on other moral judgments. But we must show that these relied-upon judgments are also reliable, which requires yet a further set of judgments, whose reliability needs to be confirmed too, and so on. Some argue that the debunker faces an insurmountable regress, which disables the debunking conclusion. In this paper, I explore and defuse this regress challenge.

Keywords: debunking arguments, regress challenge, moral epistemology, moral psychology

1. Introduction

A debunking argument contends that some target moral judgments are flawed, because such judgments have dubious origins in unreliable processes. Hence, the argument concludes, these judgments are unjustified or untrustworthy. A significant number of debunking arguments are selective – they undermine some, but not all, of our moral judgments. Regina Rini (2016) and others[1] have argued, however, that such debunking arguments cannot be sustained. To show that the target moral judgments are flawed, we need to rely on some other moral judgments. But how do we know that these relied-upon judgments are not also flawed? To check these relied-upon judgments, we might appeal to a further set of judgments. But then we will need to confirm the [illegible]ents! And so this confirmation continues. Rini argues that the defender of selective debunking arguments is committed to an insurmountable regress, and that this disables the debunking conclusion. If this is correct, then the regress challenge will have disabled an exciting new method for undermining

[1] See Sidgwick (1907, pp. 212–213) and Berker (2009, n. 76).

our moral judgments.[2] Fortunately, I believe this regress challenge can be overcome. In this paper, I examine the challenge as presented by Rini, and argue that it does not pose a problem for debunking. In section 2, I show how the regress challenge works against a specific debunking argument by Liao et al. In section 3, I explore some ways of stopping the regress, using Liao et al.'s debunking argument as a case study. Along the way, I answer some potential objections – some put forth by Rini herself. In section 4, I argue that even if there is a regress, the debunking argument still works. I conclude by drawing a more general conclusion about the success conditions for debunking.

2. Rini's regress challenge and Liao et al.'s debunking argument

Rini cites four targets of her regress challenge: Greene's (2008) debunking of 'characteristically deontological' intuitions, Horowitz's (1998) debunking of intuitions about Quinn's rescue dilemmas, Liao et al.'s (2012) debunking of Loop case intuitions, and de Lazari-Radek and Singer's (2012) debunking of the principle of egoism. In principle, however, this regress challenge could affect any selective debunking argument[3] – so long as the debunking argument relies on some moral judgments in order to debunk some other moral judgments, it is potentially vulnerable to regress, because it must justify the use of these relied-upon judgments.[4] Even more generally, the regress challenge raises the issue of how we could use a philosophical judgment to check some other philosophical judgments, and when (if ever) such checking is legitimate. I believe these issues have deep significance for moral and philosophical methodology. To make my investigation concrete and tractable, however, I'll examine the regress challenge in relation to Liao et al.'s debunking of Loop case moral judgments, which have been found to be subject to order effects. In doing so, I'm not hoping to defend the overall soundness of order effects debunking – I merely wish to use Liao et al.'s argument to illustrate how to respond to a regress

[2] Huemer (2008), Sinnott-Armstrong (2011), and McPherson (2014) all argue that something like debunking arguments can help improve our moral theorizing. The general idea of undermining our moral judgments on the basis of their origins is, however, an old one – stretching as far back as Nietzsche (2009).

[3] Thanks to an anonymous reviewer for urging me to clarify this. In my view, the regress challenge potentially affects *all* debunking arguments – whether selective or not. Roughly, this is because any debunking argument must make potentially contestable assumptions about what morality is like, and must justify these assumptions in some way. But see Rini (2016, pp. 690–694) for dissent.

[4] Consider, for example, the debunking of disgust-based judgments – such debunking must assume that disgust responses are somehow incompatible with the way the moral facts are. O'Neill (2015, pp. 1074–1076) assumes, for instance, that moral properties like badness are not easily transmissible between objects, like disgustingness is. This assumption could also be called into question, leading to a regress challenge.

challenge like Rini's. To that end, while I examine the challenge in relation to Liao et al., I will conclude with more general lessons for all moral debunking arguments.

Liao et al. investigated subjects' responses to a thought experiment known as the Loop case, where a trolley is headed toward five innocent people and will kill these five, but could be diverted to a side track to kill one innocent person. This side track loops back to the main track with five innocent people – so if no one was on the side track, the trolley would loop back and still kill the five (see Fig 1). Subjects were asked if it was morally permissible for a person to push a button that would redirect the trolley onto the side track.

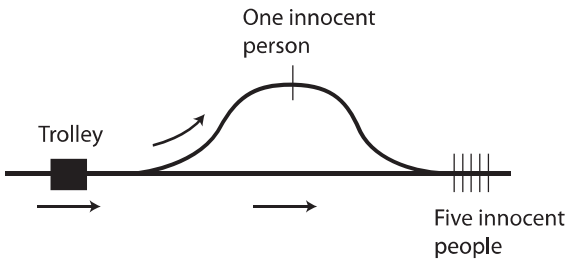

Fig 1. The Loop case

Thomson (1985) famously used this Loop case to argue against the Doctrine of Double Effect, but the details of her argument need not concern us. The interesting issue is this: Liao et al. found that subjects' intuitions about the Loop case varied depending on what case they saw before it (if any). In particular, they tested subjects' intuitions about the Loop case when it was the first case seen, versus when Loop was seen after the Standard case (where there is a side track with one person, but that track doesn't loop back), versus when Loop was seen after the Push case (where, instead of a side track, there is one person standing on a bridge over the track, and subjects are asked if it's permissible to push this one person over, killing them in order to stop the trolley and save the five). Subjects responded to Loop differently, depending on whether they saw Loop first, or whether they saw it after Standard or Push.[5] Liao et al. might then use these findings to argue that the subjects' Loop case intuitions are unreliable and should not be used in constructing moral theories. Their argument can be reconstructed as follows:

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

(*Causal Premise*) Loop case intuitions are caused by a psychological process which is subject to order effects, and such effects are not attributable to learning.

(*Theoretical Premise*) If a process is subject to order effects, and such effects are not attributable to learning, then this process does not track the moral truth.

(*Epistemic Conclusion*) Therefore, Loop case intuitions are epistemically undermined. (from *Causal Premise, Theoretical Premise*)[6]

Causal Premise claims that the target judgments[7] about the Loop case are subject to order effects that are not attributable to learning.[8] *Theoretical Premise* then casts doubt on processes of this kind, claiming that such processes do not track the moral truth – that is, they produce outputs that are unreliable indicators of the moral truth.[9] These yield the *Epistemic Conclusion* that the target judgments are epistemically undermined.

Before turning to the regress challenge, note a terminological point: Rini does not think that debunking argument itself is committed to a regress. Rather, she thinks that in making debunking arguments, the supporter of such arguments – call them the debunker – will be

[6] I omitted a bridge premise, along the lines of "If *Causal Premise* and *Theoretical Premise*, then *Epistemic Conclusion*". I also used the more specific conclusion of the target judgments being "epistemically undermined", because I think this is the most plausible version. Finally, I only presented Liao et al.'s specific argument – see Rini (2016, p. 677) and Kahane (2011, p. 106) for a more general schema.

[7] Following Rini (2016, p. 679), I use 'judgment' to cover our moral beliefs, intuitions, and any other mental state that could be an indicator of the moral truth. These distinctions will not matter here.

[8] It's possible that the order effects observed are due to learning, which would be epistemically unproblematic (Horne & Livengood, 2017). For Liao et al.'s debunking argument to work, they need to rule out this possibility. They might do so by arguing that the prior case doesn't provide any relevant information to be learned, or they might conduct further empirical tests to rule out learning (Machery, 2017, pp. 74–75). While learning hasn't been conclusively ruled out by this debunking argument, I will simply assume that it has been – since my concern is with defending debunking arguments in general against the regress challenge, rather than with defending the soundness of this specific debunking argument. Thanks here to two anonymous reviewers.

[9] Order effects debunking faces a few other problems, which I hope to set aside: for example, even if moral judgments are generally found to be subject to order effects, it might still be highly unlikely that any random person makes inconsistent judgments across different orders of presentation – see Demaree-Cotton (2016), Machery (2017, pp. 107–108), and Sauer (2018, Chapter 3) for discussion of this. I read Rini's challenge as contending that *even if* all these problems can be overcome, the debunking argument still fails in principle, because of the regress. And I merely hope to defend the in-principle viability of debunking against this regress challenge. Thanks here to an anonymous reviewer.

committed to the regress, and hence cannot support their debunking conclusion. I return to this dialectical point later – for now, focus on the challenge itself.

Rini first argues that *Theoretical Premise* needs to be justified by further first-order moral judgments – that is, judgments about the moral properties of specific acts, agents, or consequences (ie. judgments like "Act X is impermissible"). In particular, she thinks that *Theoretical Premise* "is an abstract generalization that gets its support from our unwillingness to accept moral judgments like 'act X is wrong when read-about-first but permissible when read-about-second'." (Rini, 2016, p. 682) That is, Liao et al. can only support *Theoretical Premise* by thinking about specific actions – for instance, harming others for fun, making a false promise, etc. – and considering whether that action could be impermissible when read about first, but permissible when read about second. If we're reluctant to think about actions in this way – and it seems like we are – then we can infer that the moral truths don't change across order of presentation. But we might observe moral judgments that do change with order – and if we can rule out the possibility that these changes are due to learning, then we can conclude that such judgments fail to track the truth.

Rini (2016, p. 681) draws an analogy to the task of determining whether a cat's visual system can track spiders. We might check if the cat pounces on spiders at different locations and heights, or spiders with varying amounts of camouflage. We might also test whether the cat pounces on non-spider objects that look like spiders – dust balls, for instance. When testing the cat, it seems like we must rely on our own (human) judgments about which objects are spiders. In the same way, in demonstrating that a process doesn't track the moral truth, the debunker needs to rely on their own first-order moral judgments. Rini calls these relied-upon moral judgments the *basis set* – they form the basis for the debunking argument. She thinks that Liao et al.'s basis set must consist of first-order judgments that are reliable.[10]

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

[10] It is unclear what counts as a reliable basis set. Should that set be free from *all* epistemic flaws, or just the epistemic flaw alleged by the debunking argument? I tackle this issue in section 3.2.

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

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

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

Rini (2016, p. 685) argues that this regress disables the debunking argument – if the regress continues, the debunkers "never acquire suitable grounds" for supporting the basis set or its further iterations, so they cannot support their argument. Thus, "[i]f the regress does not terminate somewhere, we never reach the debunking conclusion." (Rini, 2016, p. 685) To summarize the regress and help us keep track of it, here it is in a diagram:

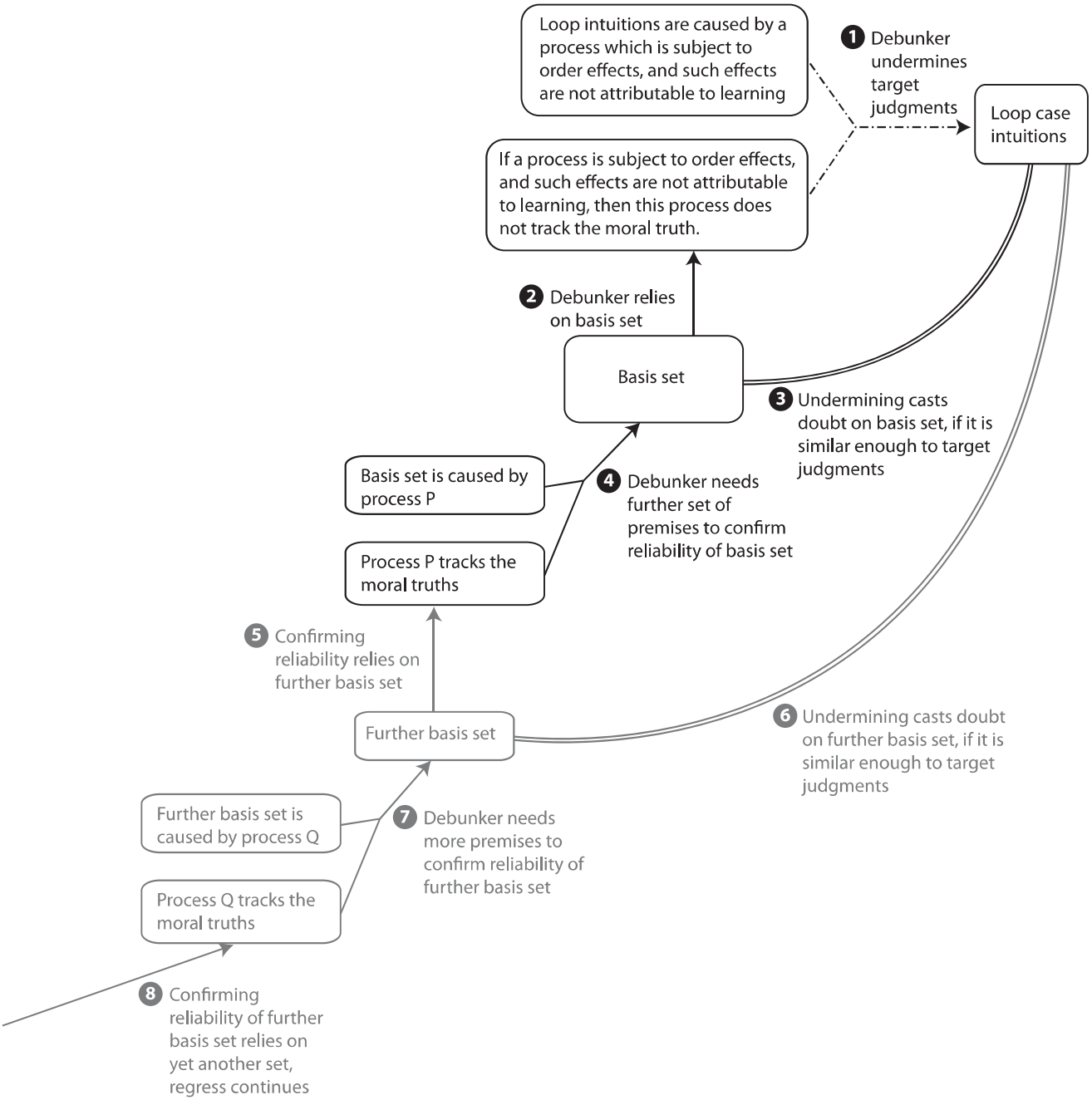

Fig 2. The regress challenge, as applied to Liao et al.'s debunking argument

The boxes in Fig 2 represent the judgments of the moral theorist or debunker. The lines represent relations of support or similarity – single solid lines with arrows denote relations of positive support; dot-dash lines with arrows denote relations of negative support or undermining; and the double solid lines denote a relation of similarity that is relevant for epistemic status. Start with the three boxes on the top right. These represent Liao et al.'s initial

debunking, where they use *Empirical Premise* and *Theoretical Premise* to undermine the target judgments – that is, the Loop case intuitions (step 1 in Fig 2).[11] Liao et al.'s argument relies on a basis set to support their *Theoretical Premise* (step 2). If this basis set is similar enough to the target judgments, then they have reason to doubt the basis set itself – this set could be flawed too (step 3). To be epistemically responsible, the debunker needs to rely on a further set of premises, which confirm that the basis set is caused by a process that *does* track moral truths (step 4). But these premises themselves rely on a further basis set (step 5), whose reliability could also be called into question if it is similar enough to the initial target judgments (step 6), and so the regress continues (steps 7, 8). Given this regress, the debunker allegedly cannot support their conclusion.

3. Resisting the Regress

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

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

Start with whether debunking arguments rely on a basis set of moral judgments. I agree that Liao et al.'s argument relies on a basis set, but this might not be true of all debunking arguments. Some debunkers instead say that to track the moral truths, we must also track the relevant *non-moral* truths – then they take issue with our ability to track these non-moral truths. For instance, O'Neill (2015, pp. 1076–1078) argues that we should not trust moral judgments that have been influenced by sympathy, because sympathy is demonstrably unreliable at tracking when various entities are in pain.[12] In effect, she argues that we fail to track the moral truths (about whether we should avoid stepping on insects, for instance) because we fail to track the relevant non-moral truths (about whether insects are in pain when we do so). Such debunking arguments do not seem to rely on a basis set of moral judgments, because they are criticising our ability to track non-moral truths. Another kind of debunking argument observes that a process – like emotion-driven processing – fails in some non-moral context, and extrapolates to conclude that it will

[11] The Y-shaped line represents the fact that *Theoretical Premise* and *Empirical Premise* together undermine the Loop case intuitions – rather than each premise doing so independently. Berker (2015, pp. 330–331) introduces the notion of Y-support – in this paper, I use its negative version, which we could call Y-undermining.

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

likely fail in the moral domain too. Such arguments might avoid using a basis set too, although these arguments are contentious.[13]

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

I agree that Liao et al. *could* support their *Theoretical Premise* this way – but that is not the only way to do so. On an alternative picture, we also have conceptual intuitions about the nature of morality, alongside our first-order moral judgments. When entertaining a claim like "The moral truth about what to do in a case does not depend on that case's order of presentation", it might just appear to be true.[15] Alternatively, we might have the conceptual intuition that "There can be no change in the moral properties described in a case, without a change in the non-moral properties described in that case".[16] If such conceptual intuitions are possible, Liao et al. might appeal to them directly to support their *Theoretical Premise*.[17]

Rini (2016, pp. 693–694) anticipates this reply from conceptual intuitions,[18] and argues that "as a purely conceptual matter, it could turn out that the truth of moral judgments is sensitive to their order of presentation", even if we are deeply inclined against such a view. She points to other kinds of truths that might depend on order of presentation – for example, some pairs of counterfactual claims may both seem true when presented in one order, but not so when

[13] See Tersman (2008, pp. 392–393), Berker (2009, pp. 316–317) and Kahane (2016, pp. 292–293) for more discussion of this.

[14] Vavova (2014, pp. 92–93) makes this point in relation to evolutionary debunking.

[15] Huemer (2008, pp. 383–387) talks about *formal intuitions* – these only impose constraints on moral theories, but do not positively or negatively evaluate anything.

[16] This intuition is similar to a formal intuition cited by Huemer (2008, pp. 386–387) – he thinks that intuitions at this level of generality "should be given special weight in moral reasoning".

[17] This presupposes an intuitionist methodology – for a prominent advocate, see Huemer (2005).

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

presented in another order. The order-insensitivity of moral truths is not a conceptual claim, she concludes – we only arrive at this order-insensitivity by generalizing from first-order moral judgments.

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

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

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

Pleasure differs from the order-insensitivity claim in important respects. Our confidence in *Pleasure* can vary as we encounter confirming and disconfirming instances of it. For example, when we encounter putative counterexamples like Nozick's (1974, pp. 42–45) experience machine, we tend to reduce confidence in *Pleasure* – or to at least reconsider and take stock of the number of confirming versus disconfirming instances observed. In contrast, when we encounter disconfirming instances of the order-insensitivity claim – that is, when we encounter moral judgments that vary with a case's order of presentation – we just reject these disconfirming instances outright.[19] Importantly, we do not pause to weigh the number of confirming instances

[19] Similarly, Michael Huemer argues that the intuition that 'better than' is transitive is not the result of considering specific cases, but rather is produced by our insight into the nature of 'better than' – this explains why we do not

against disconfirming ones – this suggests that the order-insensitivity claim is not just a high-confidence claim that is supported by many confirming instances.

In all, the comparison between moral and counterfactual truths does not help Rini's case, and we have further reasons, from our treatment of other moral generalizations, to think that the moral truths do not change across orders of presentation. So when we find moral judgments that do change with order of presentation, and can ascertain that such changes are not due to learning, we can conclude that such judgments fail to track the moral truth.

What, then, should we make of Rini's analogy to testing whether a cat can track spiders? To test this, we seemingly must appeal to our own first-order judgments about which objects are spiders. In the same way, we seemingly need to use first-order moral judgments to tell whether a process tracks the moral truths. The analogy is useful, but we have drawn the wrong lessons from it. Because we *can* reach conclusions about the cat's spider-tracking abilities without relying on first-order spider judgments. To see this, suppose I am told that a spider and a dust ball have been placed in separate coloured boxes – a red box and a green box. Each box has a hole cut out at the side – so the cat sitting on the floor can see the item in each box. I cannot see the items, however, so I cannot tell which item is in which box. Suppose I observe, over multiple trials, that the cat pounces on the item in the red box roughly half the time, and on the item in the green box the other half of the time. If the cat always pounces whenever it sees a spider, I can conclude that the cat is not good at tracking spiders. Because regardless of which box the spider is in, I can conclude that the cat correctly identifies the spider only half the time. To reach this conclusion, I need only be sure that the items stay in their respective boxes and never switch places. I do not need information about which item is in which box – which, in effect, would be a first-order spider identification judgment. Analogously, the debunker doesn't need first-order moral judgments to conclude that processes subject to order effects do not track the moral truth. They can get by with something weaker – with just the assumption that the moral truth about a case doesn't change with its order of presentation.

Still, an opponent might take issue with the assumption that the items stay in their respective boxes and never switch places. This assumption, they might argue, can only be supported by further first-order spider identification judgments.[20] However, even here, I think our beliefs might not bottom out in first-order judgments either. Consider our understanding of object

immediately accept counterexamples as disproving transitivity, but instead declare such situations "paradoxical" (Huemer, 2008, pp. 386–387).

[20] Thanks here to an anonymous reviewer.

permanence and spatiotemporal continuity – as adults, we understand that objects persist in time and space, and that if an object appeared at one point and then at another, it must have taken a path between these points. Empirical findings suggest that we might not have learned these things from experience – because even very young infants display expectations of object permanence and spatiotemporal continuity too (Samet & Zaitchik, 2017, sec. 2.2.1). Thus it is possible, even in this perceptual analogy, that we are not just generalizing from first-order judgments. (Of course, it is a further claim whether morality works like this too.)

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

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

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

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

do not; promise-keeping need not involve stakes of bodily harm, whereas the Loop case does; and so on. Thus, contrary to Rini, the debunker might find a basis set that is different enough from the Loop case intuitions, such that the initial undermining does not cast doubt on the basis set. The possible alternatives are illustrated in Fig 3:

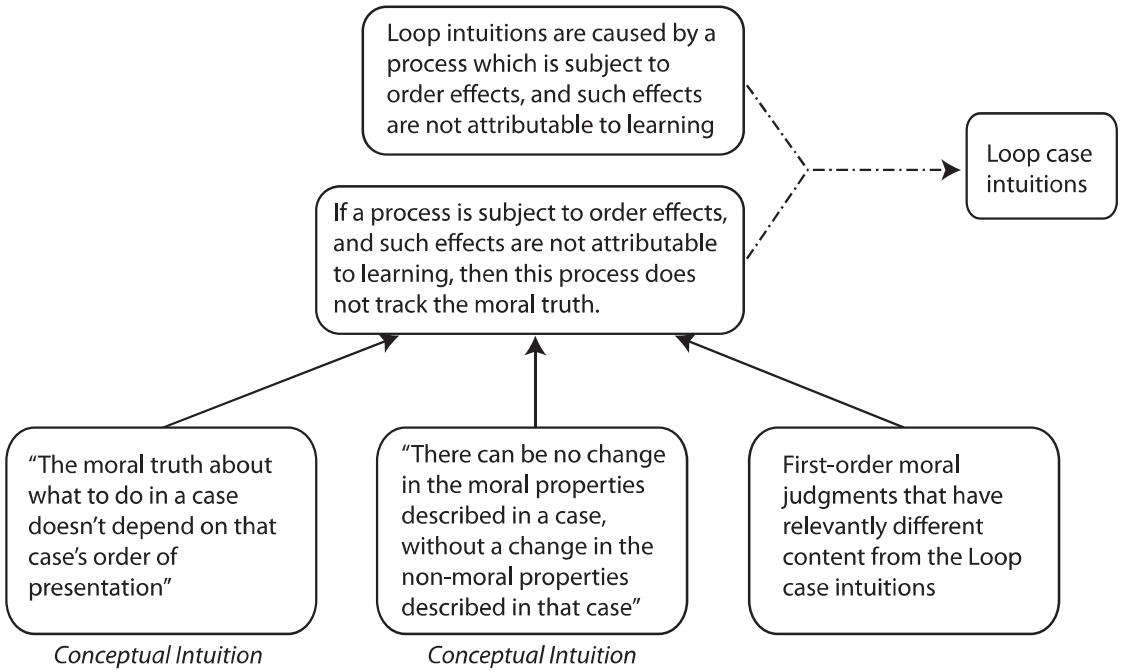

Fig 3. Alternative ways of supporting *Theoretical Premise*

Rini anticipates the reply that the basis set will be sufficiently different from the target Loop case intuitions. She argues, however, that Liao et al. cannot invoke this reply, because their debunking argument targets a large set of moral judgments (Rini, 2016, pp. 687–688). Here Rini (2016, n. 15) takes the number of target judgments as a proxy for similarity to the basis set: the more target judgments there are, the more likely there will be a relevant similarity between the target judgments and the basis set – such that undermining the target judgments will likely also undermine the basis set.

In response, I think we should just reject this strategy of using the number of target judgments as a proxy for similarity. First, because the number of target judgments is a highly imperfect proxy – if the many target judgments are of the same kind, for instance, increasing their number will not increase similarity to the basis set. Secondly, we can directly assess the thing being proxied – similarity to the basis set – so why bother with proxies at all? For instance, I argued above that Liao et al.'s basis set differs from the target Loop case intuitions, because this basis set could have

a different level of generality (if they used conceptual intuitions), or because the basis set could have different content (if they used first-order judgments about promise-keeping). In doing so, I focused directly on the similarities involved, rather than on imperfect proxies. This makes my assessments of similarity – and ultimately, of the epistemic status of the basis set – more fine-grained and likely more accurate than Rini's.

Of course, it is unclear which dimensions of similarity are important for inferring epistemic status. For any particular moral judgment M, this judgment is similar to other moral judgments in various respects. We want to infer M's epistemic status from the epistemic status of these other moral judgments, by virtue of their similarity. To do this, we need to determine which of these other judgments count as 'most similar' for our purposes, which would mean identifying the dimensions of similarity that provide good indications of a judgment's epistemic status. This is difficult, but we might still offer arguments here. Huemer (2008, pp. 383–384), for instance, argues that the level of generality of a moral judgment matters – because concrete judgments about specific cases are more likely to be susceptible to emotional bias, and will more likely to be the results of biological and cultural programming. Abstract intuitions, on the other hand, are more likely to be the products of rational reflection.

Remember that the whole point of assessing similarity between the target judgments and the basis set is to see whether we have reasons to doubt the basis set. But note a crucial ambiguity here: 'reasons to doubt to the basis set' could either mean reasons to think that the basis set is *flawed in the same way* as the target Loop case intuitions (that is, the basis set might also be subject to order effects), or it could mean reasons to think that the basis set is *flawed in some unspecified way* that might not relate to order effects at all.[21]

If we take the first reading of 'reasons to doubt', there is a clear way out of the regress for Liao et al. – since their basis set can just comprise of moral judgments *that are not subject to order effects, or are subject to order effects that are due to learning*. This results in the structure of support shown in Fig 4 – the basis set plays a role in supporting itself, when combined with empirical results showing that it was produced by a process that isn't subject to order effects.

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

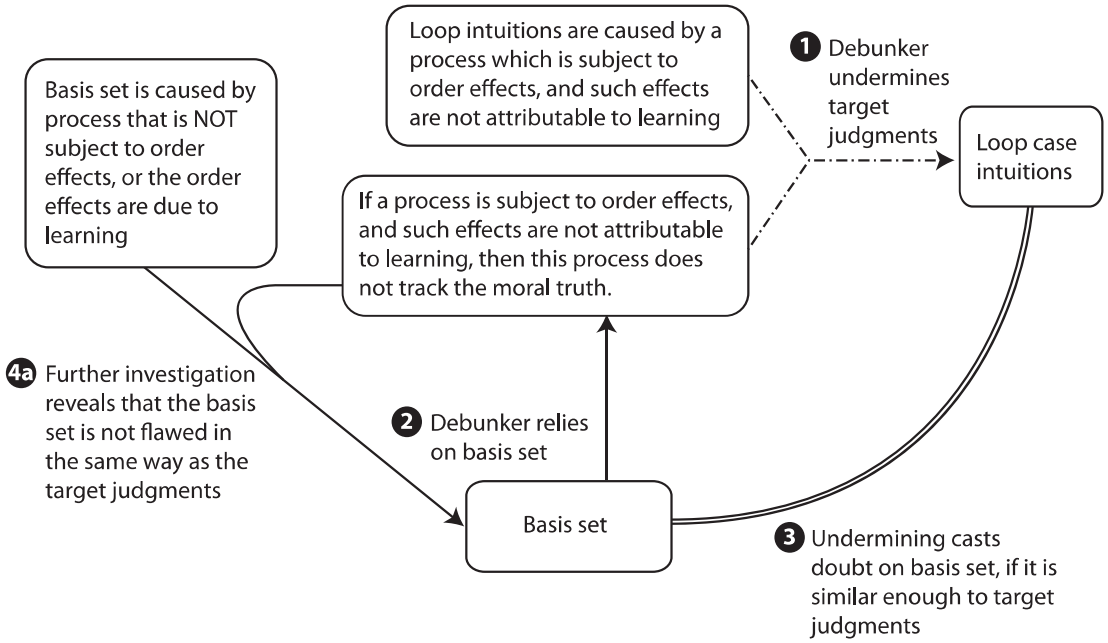

Fig 4. Supporting the basis set without relying on a further basis set

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

If we take the second reading of 'reasons to doubt', then the standard to meet is much higher. Not only do the debunkers have to show that the basis set is not flawed in the same way as the target judgments, they have show that the basis set *is not flawed at all*. This reading of 'reasons to doubt' is problematic in several ways. Firstly, it's not supported by the empirical, targeted nature of the regress challenge. We don't need empirical results or similarity between different judgments to tell us that our moral judgments are fallible and could be flawed in various ways. If

that is all the regress challenge amounted to, then this seems no different to pressing a general sceptical concern against the debunker.[22] Compare again with the perceptual case: if I realise that some of my perceptions are unreliable in some specific way, this only gives me reason to worry that similar perceptions will be unreliable in similar ways. So instead, I believe the regress challenge is better read as a more targeted one: if the target judgments are flawed in some specific way, and such judgments are relevantly similar to the basis set, then we should worry that the basis set is also flawed in the same way. Secondly, this reading of 'reasons to doubt' is impossibly demanding. We don't demand of others that they eliminate all suspicion of error before considering their arguments. In the same way, we might be holding the debunkers to an overly demanding standard – beyond the scope of proper epistemic responsibility – when asking that they give conclusive proof that their basis set is free from all epistemic flaws.

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

Finally, quite separate from all the previous issues, notice that the target judgments do all the work in generating the reasons to doubt each iteration of the basis set (see steps 3 and 6 in Fig 2). The debunker might argue that at some point, one of the further basis sets is just going to be too different from the target judgments. It is difficult to assess this strategy, because it is unclear what kinds of judgments are in the further basis set. But as long as the notion of similarity used does not deem the target judgments similar enough to *all further basis sets*, the regress will stop at some point.

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

4. Even with a regress, the debunking argument still works

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

4.1 Flows of Justification

First, if we pay attention to the flows of justification, it is likely that the regress challenge only dampens the impact of the debunking argument, rather than neutralizing it entirely. To see this, first notice that we can cast the regress challenge as a challenge about neutralizing justification: an increase in justification for *Theoretical Premise* might initially undermine the target judgments. But because the target judgments are similar to the basis set, the initial undermining also calls the basis set into doubt – this then reduces support for *Theoretical Premise*. So the initial increase in justification for *Theoretical Premise* might ultimately neutralize itself. These flows of justification are labelled in Fig 5 below.

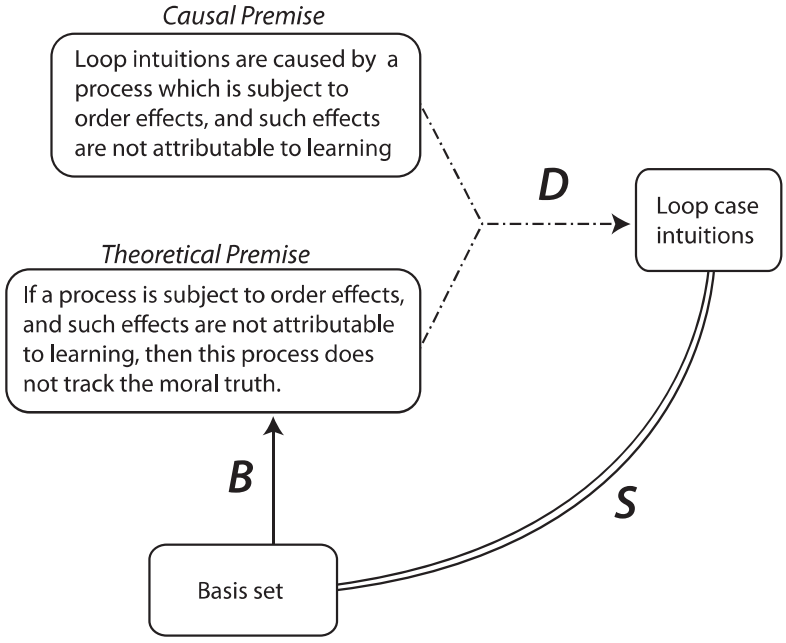

Fig 5. Abridged diagram with flows of justification

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

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

We started out with a 0.1 increase in justification of *Theoretical Premise*, but this also generated a $[0.1 * D * S * B]$ decrease in justification of that same premise. So the overall increase in justification of *Theoretical Premise* amounts to $0.1 - [0.1 * D * S * B]$. For the initial 0.1 increase to fully neutralize itself, we need D, S, and B to all be equal to 1. Remember that D, S, and B represent the marginal debunking potential of the debunking argument, the undermining potential of the similarity between the target judgments and the basis set, and the supporting potential of the basis set respectively. This reveals simple ways for the debunker to avoid the regress challenge, when construed as challenge about neutralizing justification. The debunker might aim for a weaker debunking argument with a lower debunking potential – one which says, for instance, that increasing the justification for *Theoretical Premise* by x units should only lead

to a $[0.5 * x]$ decrease in justification for the target judgments (debunking potential of 0.5). The debunker might also argue that while there is some similarity between the target judgments and the basis set, this similarity is not complete, so the undermining potential of this similarity is less than 1: each x unit decrease in justification for the target judgments should only lead to, say, $[0.7 * x]$ units decrease in justification of the basis set (undermining potential of 0.7).[23]

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

4.2 Debunking arguments need not be dialectical challenges; they merely highlight an internal inconsistency

Next recall that the debunking argument itself is not committed to a regress – rather the supporter of such arguments is committed to the regress. Rini views debunking arguments as being put forth by some opponents, the debunkers – these opponents have judgments of their own, and they could be defeated by a regress challenge. But this is not the best way of seeing the situation. While we often speak of the debunkers as real opponents to be defeated – I myself have done so in this paper – there need not be such opponents at all. Instead, debunking arguments merely highlight an internal tension between *one's own* moral judgments.[24] Liao et al., for instance, highlight a tension between believing, on the one hand, that moral truths don't depend on a case's order of presentation, and trusting the Loop case intuitions on the other.[25] They can be read as saying that insofar as we are committed to *Theoretical Premise*, we should not trust Loop case intuitions. In other words, Fig 2 represents our own judgments – not those of the

[23] They could also that the basis set does not strongly support *Theoretical Premise* – I set that aside here.
[24] Srinivasan (2015, p. 346) makes the same point against self-undermining objections to debunking arguments.
[25] There might be two other ways to resolve the tension: deny the empirical evidence in *Causal Premise*, or reject the epistemic principles that the argument implicitly relies on. I suppressed these options because they seem much less attractive.

debunker. Regardless of whether there is a regress, the debunking argument – as illustrated by the three boxes on the top right – still exposes a real tension that needs to be resolved, given the judgments we hold. This aligns with Kumar and Campbell's (2012, pp. 315–319) thought that debunking arguments reveal inconsistencies between our moral judgments.[26]

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

Rini is still right to point out, though, that we will never be certain whether the moral truths are sensitive to order of presentation or not. When doing debunking, we are just checking one moral judgment with another, without confirming whether any are ultimately accurate. This is a genuine problem, but I think abandoning debunking is the wrong response. Because this is a general problem that affects many other domains – and in these other domains, we still engage in something like debunking. Think, for instance, of the analogous case in sense perception. To check someone's perceptual judgments about whether there are trees or not, I need to use my own judgments about trees. But how do I know if my own perceptual judgments are reliable? I need to check them with some further judgments, which themselves need to be checked, and so on – a similar regress ensues. It does not follow from this, however, that optometrists should not check anyone's eyesight.[27] Even if we cannot confirm that our perceptual judgments are ultimately reliable, there is still epistemic value in trying our best to make these judgments consistent with each other, and with our current empirical findings. The same goes, I think, in the moral case.[28] Saying we should not do debunking because of the regress challenge is analogous to saying that we should not check anyone's eyesight because of the same epistemological regress for sense perception.

To be fair, it might also be that once the relied-upon basis set is made explicit, it is controversial whether this basis set is reliable, so the debunking argument fails. (Kumar and Campbell (2012,

[26] See also Rini (2017).

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

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

pp. 317–318) allege this of Greene's debunking argument.) This would be like discovering that my optometrist was drunk when they diagnosed me as colourblind – given their unreliability, it might be rational to resist the debunking and hang on to the allegedly undermined target judgments. It is difficult to give exact conditions for when a debunking argument succeeds, but here's a tentative proposal: when we are more justified in trusting the basis set than we are in trusting the target judgments, and when our justification for trusting the basis set exceeds a certain threshold, then the debunking works. While I believe Rini's regress challenge doesn't disable the debunking argument, our investigation into it has drawn attention to this important and neglected issue: how much justification we have for trusting the basis set and the target judgments, and how this affects the success of debunking.

At this point, a critic might wonder: if it all comes down to whether we were initially justified in trusting the basis set, why bother with empirical investigation at all?[29] To answer this, notice that for the debunking argument to work, we need *both* the empirical findings (as exemplified in *Causal Premise*) and some theoretical claims about what morality is like (as stated in *Theoretical Premise*). Even if we agree that some class of judgments is epistemically flawed, we still need positive empirical support for thinking that a specific judgment (e.g. an intuition about the Loop case) falls into that class. So the empirical evidence is still necessary for us to identify epistemic flaws. The overall picture that emerges is this: we agree on some clear epistemic flaws of our moral judgments, and our empirical investigations help identify and weed out these flawed judgments from our moral theorizing.[30]

5. Conclusion

In conclusion, the regress challenge does not disable debunking arguments. I've explored different ways of resisting the regress: craft the debunking argument so that it does not rely on a basis set; use a basis set that is different enough from the target judgments (be it in its level of generality, its content, or some other relevant dimension); argue that the similarity between the target judgments and further iterations of the basis set will give out at some point; and finally, discover empirical results showing that the basis set is not flawed in the same way as the target judgments. These strategies are perfectly general – even if you thought that my specific illustration of them (involving Liao et al.) was implausible, they might still be helpful in

[29] Thanks to an anonymous reviewer for raising this.

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

defending a debunking argument of your choice against similar issues with regress. Moreover, while each strategy might not be equally effective across different debunking arguments, having several strategies on hand creates a formidable defence against the regress.

I have also argued that even with a regress, the debunking argument still works. Firstly, if we pay attention to the flows of justification, we can see that the regress challenge might only weaken the debunking argument, rather than neutralize it entirely. Secondly, debunking arguments should not be cast as a dialectical attack from opponents with judgments of their own. Instead, such arguments reveal an internal tension within our own moral judgments: either let go of the basis set, or discard the target judgments. This tension obtains, regardless of whether there is also a regress. And we should, epistemically speaking, resolve this tension. When we are justified in trusting the basis set over the target judgments – as I believe is the case with Liao et al.'s argument – we should accept the debunking and throw out the target judgments.

Acknowledgements Thanks to Christian Barry, Alan Hájek, Seth Lazar, Devon Cass, James Willoughby, Daniel Stoljar, and Abhishek Mishra for comments and conversations about this paper, and thanks to an audience at the 2016 Australasian Postgraduate Philosophy Conference for their helpful questions. I'm also very grateful to two anonymous reviewers for the *Canadian Journal of Philosophy*, and an anonymous reviewer for another journal, for their insightful comments, which have greatly improved this paper. This research is supported by an Australian Government Research Training Program (RTP) Scholarship.

References

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

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

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

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

Demaree-Cotton, J. (2016). Do framing effects make moral intuitions unreliable? *Philosophical Psychology*, *29*(1), 1–22. https://doi.org/10.1080/09515089.2014.989967

Greene, J. (2008). The Secret Joke of Kant's Soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol. 3*. MIT Press.

Horne, Z., & Livengood, J. (2017). Ordering effects, updating effects, and the specter of global skepticism. *Synthese*, *194*(4), 1189–1218. https://doi.org/10.1007/s11229-015-0985-9

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

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

Huemer, M. (2008). Revisionary Intuitionism. *Social Philosophy and Policy*, *25*(01), 368–392. https://doi.org/10.1017/S026505250808014X

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

Kahane, G. (2016). Is, Ought, and the Brain. In S. M. Liao (Ed.), *Moral Brains: The Neuroscience of Morality* (pp. 281–311). Oxford University Press.

Kumar, V., & Campbell, R. (2012). On the normative significance of experimental moral psychology. *Philosophical Psychology*, *25*(3), 311–330. https://doi.org/10.1080/09515089.2012.660140

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

Machery, E. (2017). *Philosophy Within Its Proper Bounds*. 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

Nietzsche, F. (2009). *On the Genealogy of Morals* (D. Smith, Trans.; 1st edition). Oxford University Press.

Nozick, R. (1974). *Anarchy, State, and Utopia*. Basic Books.

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

Rini, R. A. (2016). Debunking Debunking: A Regress Challenge for Psychological Threats to Moral Judgment. *Philosophical Studies*, *173*(3), 675–697.

Rini, R. A. (2017). Why moral psychology is disturbing. *Philosophical Studies*, *174*(6), 1439–1458. https://doi.org/10.1007/s11098-016-0766-4

Samet, J., & Zaitchik, D. (2017). Innateness and Contemporary Theories of Cognition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2017). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2017/entries/innateness-cognition/

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

Sidgwick, H. (1907). *The Methods of Ethics* (7th ed.). Macmillan. https://www.gutenberg.org/files/46743/46743-h/46743-h.htm

Sinnott-Armstrong, Walter. (2006). *Moral Skepticisms*. Oxford University Press.

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

Srinivasan, A. (2015). The Archimedean Urge. *Philosophical Perspectives*, *29*(1), 325–362. https://doi.org/10.1111/phpe.12068

Tersman, F. (2008). The reliability of moral intuitions: A challenge from neuroscience. *Australasian Journal of Philosophy*, *86*(3), 389–405. https://doi.org/10.1080/00048400802002010

Thomson, J. J. (1985). The Trolley Problem. *The Yale Law Journal*, *94*(6), 1395–1415.

Vavova, K. (2014). Debunking Evolutionary Debunking. *Oxford Studies in Metaethics*, *9*, 76–101.

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

Wiegmann, A., Okan, Y., & Nagel, J. (2012). Order effects in moral judgment. *Philosophical Psychology*, *25*(6), 813–836. https://doi.org/10.1080/09515089.2011.631995