Debunking debunking: A regress challenge for psychological threats to moral judgment

Regina A. Rini   NYU Bioethics

This paper has been accepted to appear in Philosophical Studies but has not yet been through final copy-editing. It is possible that the published version will differ in small ways.

Abstract   This paper presents a regress challenge to the selective psychological debunking of moral judgments. A selective psychological debunking argument conjoins an empirical claim about the psychological origins of certain moral judgments to a theoretical claim that these psychological origins cannot track moral truth, leading to the conclusion that the moral judgments are unreliable. I argue that psychological debunking arguments are vulnerable to a regress challenge, because the theoretical claim that 'such-and-such psychological process is not moral-truth-tracking' relies upon moral judgments. We must then ask about the psychological origins of *these* judgments, and then make a further evaluative judgment about these psychological origins… and so on. This chain of empirical and evaluative claims may continue indefinitely and, I will argue, proponents of the debunking argument are in a dialectical position where they may not simply call a halt to the process. Hence, their argument cannot terminate, and its debunking conclusion cannot be upheld.

1. Psychological debunking and regress challenges

In *The Methods of Ethics*, Henry Sidgwick swiftly rebuts the idea that our moral judgments are undermined simply by being the causal result of a psychological process. According to Sidgwick, evaluating moral judgments in that way leads to an infinite regress:

> I cannot see how the mere ascertainment that certain apparently self-evident judgments have been caused in known and determinate ways, can be in itself a valid ground for distrusting this class of apparent cognitions. I cannot even admit that those who affirm the truth of such judgments are bound to show in their causes a tendency to make them true: indeed the acceptance of any such *onus probandi* would seem to me to render the attainment of philosophical certitude impossible. For the premises of the required demonstration must consist of caused beliefs, which as having been caused will equally stand in need of being proved true, and so on *ad infinitum*... (Sidgwick 1907, 212-213)

According to Sidgwick, the fact that our moral judgments have psychological causes does not undermine them, nor even does it necessitate a positive argument that their causes are conducive to moral truth. Anyone who starts to play the game of psychologically debunking moral judgments has already made a mistake, as the game will never end. Any premise introduced to evaluate the cause of moral judgments is itself the causal result of some psychological process, and this further process will need to be understood and evaluated, and *that* evaluation will also be caused by some process, and so on…

Here Sidgwick is talking about a causal process involved in the production of *all* moral judgments – what we might now call the moral faculty or central moral cognition. To claim, as Sidgwick does, that we cannot debunk *all* moral judgments on their being rooted in this *common* cause is distinct from claiming that we cannot debunk *particular* moral judgments on their being rooted in *particular* psychological causes. In fact, Sidgwick explicitly allows for the latter possibility:

> It may, however, be possible to prove that some ethical beliefs have been caused in such a way as to make it probable that they are wholly or partially erroneous: and it will hereafter be important to consider how far any Ethical intuitions, which we find ourselves disposed to accept as valid, are open to attack on such psychological grounds. (*ibid*., 213)

In effect, Sidgwick draws a distinction between the *global* debunking of all moral judgments, which gets caught up in an infinite regress, and the *selective* debunking of some but not all moral judgments, which possibly avoids the regress challenge.[1]

In this paper I argue that a version of Sidgwick's regress challenge threatens even selective debunking. I will claim that nearly *any* attempt to debunk a particular moral judgment on grounds of its psychological cause risks triggering a regress, because a debunking argument must involve moral evaluation *of the psychological cause* – and this evaluation is itself then subject to psychological investigation and moral evaluation, and so on. I will argue that, given the dialectical context in which selective debunking arguments are typically invoked, there is no satisfying way to halt the regress. If I am right, psychological debunking is a far less attractive tool for moral theory than some have supposed.

In section 2 I provide an argument schema for selective psychological debunking. In section 3 I present the core of the regress challenge. Section 4 considers various ways that proponents of debunking might seek to halt the regress, and argues that none are successful. In section 5 I return to the question of *global* debunking, using Richard Joyce's recent debunking argument to illustrate what is distinctive about selective debunking. I conclude with a brief discussion of the limits that the regress challenge might impose on psychologically-informed moral theory.

2. Selective psychological debunking arguments

What is a psychological debunking argument? In this section I will provide an argument schema, introduce some examples, and clarify some of its central terms. I will turn to the regress challenge in the next section.

Here is a *Psychological Debunking Schema*:[2]

Empirical Premise (ψ): A set of moral judgments *M* is caused by psychological process *P*.

Theoretical Premise (ϕ): Psychological process *P* does not track the moral truth.

Conclusion: Therefore, moral judgments *M* are unreliable.

[1] The distinction between global and selective debunking appears in Kahane (2011), though he uses the term 'local' instead of 'selective'. Kahane presents several challenges for the feasibility of selective debunking (at least when motivated by evolutionary concerns). Proponents of debunking have also appealed to the distinction: Lazari-Radek and Singer (2012, 12) approvingly quote the Sidgwick passages above, and present their project as fitting Sidgwick's characterization of selective debunking.

[2] Similar schema appear in Mason (2011, 450) and Kahane (2011, 106) . It might be objected that some of the terms of my schema are vague. This is true – they are vague on purpose. The schema ranges over a number of instances of the argument type (some examples of which appear below). The vague terms of the schema allow agnosticism about differences among the various instances. The regress challenge is meant to apply to all instances of the schema, for which these differences are irrelevant.

It will be helpful to look at a few instances of psychological debunking. Here is one, from philosopher-neuroscientist Joshua Greene (2008, 69-70):

> There are good reasons to think that the distinctively deontological moral intuitions (here, the ones that conflict with consequentialism) reflect the influence of morally irrelevant factors and are therefore unlikely to track the moral truth. … [W]e respond more emotionally to moral violations that are "up close and personal" because those are the sorts of moral violations that existed in the environment in which we evolved.[3]

According to Greene, (ψ) the set {deontological intuitions} is caused by a certain emotion-driven psychological process, and (ϕ) this emotion-driven process does not track the moral truth.

Here is another, from Tamara Horowitz (1998, 385):

> It is therefore possible that prospect theory, augmented by the empirical hypotheses I described earlier, provides the correct account of the reasoning engaged in by people who come to have [certain moral intuitions]. If this is so, then Quinn's philosophical thought experiments do not provide us with an argument for his philosophical conclusions.[4]

According to Horowitz, (ψ) the set {certain moral intuitions discussed by Quinn} is caused by a psychological process described in prospect theory, and (ϕ) this process does not track the moral truth.

Another from Liao et al. (2011, 667), showing that the order in which cases are presented affects subjects' responses to them:

> [O]ur study suggests that intuitions about Loop are sensitive to the context in which the case is considered, in particular, when the intuitions are elicited in close temporal proximity to certain related hypothetical cases. Since evidence is only trustworthy to the extent to which its sensitivity is limited to things that are relevant to the truth or falsity of the claims for which it is supposed to provide evidence, the evidentiary status of Loop intuitions is significantly challenged by our findings.[5]

According to Liao and colleagues, (ψ) the set {Loop case intuitions} is caused by a psychological process which is sensitive to the order of case presentation and (ϕ) this process does not track the moral truth.

Finally, another from Lazari-Radek and Singer (2012, 28):

> [Prioritizing one's own good] may indeed be in accordance with common sense, but here common sense seems likely to have been formed by the evolutionary influences we have been discussing. Since the claim that egoism is rational clashes with the principle of universal benevolence… we have grounds

[3] Greene argues that neuroscientific evidence shows a correlation between deontological intuition and "up close and personal" forms of harm (e.g. physically pushing a person to his death rather than using a machine). He claims similar origins of other deontological intuitions, such as those about punishment. For critiques of the details of Greene's argument, see Berker (2009) and Kahane (2012).

[4] Prospect theory (Kahneman and Tversky 1979) is a psychological theory about the differential effects of gain- and loss-framing on decision making. Horowitz is discussing Quinn (1989). For critiques of the details of Horowitz's argument, see Kamm (1998) and van Roojen (1999). See also Sinnott-Armstrong (2008).

[5] The Loop case in question figures centrally in Thomson (1976) and Kamm (2007). Liao and colleagues are cautious about the strength of their conclusion; see also Liao (2008).

> for supporting the intuition for which there is no evolutionary explanation rather than the one for which there is an evolutionary explanation.[6]

According to Lazari-Radek and Singer, (ψ) the set {egoistic intuitions} is caused by a psychological process which is susceptible to evolutionary explanation, and (ϕ) this process does not track the moral truth.

The range of these claims should make clear the utility of the schema. Some of these arguments target only particular judgments (the Loop case or Quinn's intuitions), while others target much larger sets (all deontological intuitions, or all egoistic judgments). They differ somewhat in the degree of skepticism they urge toward the targeted judgments. But they share a common structure: they identify some psychological process (emotional triggers, order effects, etc.) as causally responsible for the targeted judgments and claim that this psychological process is unlikely to track moral truth.

In a moment I will introduce the regress challenge. But first, some ground-clearing is in order, to avoid any misunderstandings about the psychological debunking schema itself.

The Empirical Premise (ψ) discusses a *set* of moral judgments. This set may range in size from a singleton – one specific moral intuition – to the set of all moral judgments. The size of the set relates to the dialectical purpose of a given instance of the argument. For small sizes, debunking is usually a move *within* normative ethics, intended to advance some moral principle by disqualifying intuitions that run against it (e.g. Greene 2008). For very large sizes, debunking is usually understood as a metaethical claim, one aiming to challenge the epistemic credentials of intuition-based moral theory (e.g. Sinnott-Armstrong 2008). As already noted, very large sizes for this set are characteristic of *global* debunking.

ψ refers to moral 'judgments' and a 'psychological process'; both of these terms have imprecise meanings. In this schema they are intended to be as broad as possible. Moral 'judgments' may include intuitions, Rawlsian 'considered judgments', dispositions to believe, or occurrent reflective beliefs. Similarly, the term 'psychological process' is imprecise, ranging across *proximal* processes (e.g. cognitive mechanisms in the brain) and *distal* processes (e.g. evolutionary pressures). If we were interested in only a particular *instance* of this premise, we would need to be more precise – but my challenge is intended to apply to the schema itself.[7]

The Theoretical Premise (ϕ) expresses doubt that *P* 'tracks the moral truth'. To say that a psychological process *does* track the truth in a particular domain is to say that the psychological process subjunctively responds to the facts of that domain: the psychological process would indicate the presence of a domain-property if it were present, and would not do so if the property were not present.[8] For instance, a visual process tracks the truth about spiders if it would indicate the

[6] The arguments here rely in part on Greene's, above. See also Singer (2005) and, for criticism, Sandberg and Juth (2011) and Kahane (2014).

[7] The verb 'cause' may also deserve some caution, as any particular moral judgment will in fact be the causal result of many different psychological processes. It may be better to say that *the quality or manner* of judgments *M* is *influenced* by process *P* (see Lewis 2000). I will set aside such issues here.

[8] The terminology comes from Nozick (1981, 178), who defines a person possessing *knowledge* as one whose belief is subjunctively sensitive to the truth. Here I will not be concerned with whether the judgments in question meet the criteria of moral 'knowledge'. I also elide any difficulties that may arise from speaking about

presence of spiders when spiders are about, and would not do so when they are thankfully absent. A process that tracks the moral truth will exhibit this sort of relationship to properties like *wrongness*, *permissibility*, and the like. Truth-tracking comes in degrees (a mediocre spider-tracker can sometimes miss the vicious deceivers that appear to be ants until they crawl into one's hair). ϕ expresses doubt that process *P* tracks moral truth to any significant degree.

Finally, the conclusion of the schema claims that the judgments in *M* are 'unreliable'. This could have several meanings, including: that *M* cannot count as evidence; that beliefs based upon *M* lack justification, or have their justification undermined; that acting on the basis of *M* is likely to lead to moral error. Depending on the nature of epistemic reliability, some of these proposals may just be different ways of saying the same thing. If there are differences, they do not matter to the argument of this paper.

Qualifications deployed, I'll now turn to the regress challenge.

3. The regress challenge

This section presents the core of the regress challenge to selective psychological debunking. There are two steps to constructing the regress challenge. First, I will argue that the selective psychological debunking schema not only *targets* a set of moral judgments, but also *relies upon* another set of moral judgments. Second, I will show that this reliance on moral judgments makes the schema vulnerable to demands that the psychological origins of these relied-upon judgments be investigated and morally assessed – which triggers another round of psychological inquiry and moral assessment, and so on. This section presents the regress argument itself; in the next section I will discuss ways in which debunking proponents might try to stop it.

3.1 Why debunking moral judgment relies on moral judgment

Here again are the premises of the psychological debunking schema:

ψ: A set of moral judgments *M* is caused by psychological process *P*.

ϕ: Psychological process *P* does not track the moral truth.

ψ is an empirical claim, a claim made on the basis of scientific investigation. But what sort of claim is ϕ, and what sort of basis might it have? I will argue that ϕ *relies upon* moral judgments. Our confidence in ϕ itself must be generated in part by moral cognition. The regress challenge gets going by asking about the psychological status of *that* cognition.

First, though, why does ϕ rely upon moral judgments? One might suppose that it is a metaethical claim – a claim *about* the moral domain, not *within* the moral domain. After all, ϕ does not itself make any first-order moral claims about acts, agents, characters, or consequences. It instead makes a higher-order claim: that psychological process *P* does not have a certain relationship to moral

a *process* (rather than a full person) as tracking truth. See Street (2006) for an influential use of the term in a way that is identical to my own.

properties. Seemingly, ϕ needn't take *any* position on what the moral properties are or which acts, agents, characters, and consequences instantiate them.

I disagree. The boundary between metaethics and morality is debated, but I am broadly in agreement with Allan Gibbard's contention that many purportedly metaethical statements are just combinations of ordinary moral claims, "arrayed in sumptuous rhetoric" (Gibbard 2003, 186).[9] In the case of psychological debunking, we can see this if we imagine how we might go about assessing the truth of premise ϕ: how might we *find out* whether or not a particular psychological process *P* tracks the moral truth?

But morality is tricky, so let's start with a simpler example. How can we find out whether a particular psychological process does a good job of tracking the truth about *spiders*? Consider the visual system of my cat. My cat seems to have a great interest in spiders (primarily for purposes of dismemberment and consumption). I find her engaged in spider-related activities with some regularity. She seems to be good at noticing them. But this might just be a coincidence; maybe I live in a very spider-intensive neighborhood and the cat cannot avoid coming across spiders all the time. If I wish to rule this out and conclude that my cat is good at seeing spiders, I will have to do some investigation.

I could catch some spiders and place them in salient locations, at varying heights and amid varying sorts of camouflage. Will my cat pounce on the ones within range? Will she move closer to the distant ones? If so, this suggests that her visual system does reliably indicate the presence of spiders. But I will also have to test her ability to discriminate spiders from non-spiders. Will she pounce on dust balls blown by the wind, or little specks of dirt in the corner? If she reacts differently to these non-spiders than she does to actual spiders, this suggests her vision reliably does not indicate spiders when they are absent. And if both these conditions are met, I have strong evidence for the claim that my cat's visual system is spider-tracking.

Notice that this procedure relies on *my* judgments about the nature of spiders. I am assuming that such-and-such small creeping thing in front of me *is* a spider, and not an ant or centipede or some still more horrible arthropod. Obviously, my own spider classification practices will have great significance for the outcome of my investigation into my cat's visual spider-tracking ability. Given this fact, it would be unhelpful for someone to insist that the claim 'the cat's vision is spider-tracking' is *not* a spider claim, but is instead a *meta-arachnid* claim. True, the statement itself does not explicitly make any first-order claims about particular spiders. But it *does* rely upon a range of such claims, in that my grounds for it require an investigation where making first-order spider claims is unavoidable.

The same is true for ϕ. It is beside the point to call ϕ a metaethical claim in this context. Our grounds for accepting an instance of ϕ – that is, our grounds for accepting that some particular psychological process does not track the moral truth – require an investigation where making first-order moral claims is unavoidable. An instance of ϕ relies upon some set of moral judgments.

We can see this if we examine again some actual instances of ϕ. Recall that Joshua Greene claims to have psychologically debunked deontological moral intuitions by showing that "these intuitions

[9] See also Blackburn (1985) and Dworkin (1996) for doubts about the distance between metaethics and ethics.

appear to have been shaped by morally irrelevant factors having to do with the constraints and circumstances of our evolutionary history" (Greene 2008, 75). According to Greene, deontological moral intuitions are distinctively responsive to factors such as whether a particular means of harming someone involves "up close and personal" physical violence – the sort of factors most salient to our distant ancestors and evolutionarily adaptive for them to attend to. Greene claims that whether a harm is done through up close and personal violence or through some other means is "morally irrelevant" in assessing the status of the action. This claim relies upon a set of moral judgments: that certain sorts of harm are morally permissible, whether or not up close and personal violence is involved.[10]

Even the most seemingly obvious instance of ɸ still relies upon certain moral judgments. You'll recall that Liao and colleagues (2011) found that Loop case intuitions are caused by a psychological process sensitive to the order in which cases are presented, and they claim that the order in which cases are presented has nothing to do with the moral truth. The latter is surely an uncontroversial claim; I doubt anyone would wish to defend the view that the moral status of an act could somehow depend upon whether any given moral judge had read about a different act before or after reading about it. But a claim can be uncontroversial and still *rely upon* a set of (uncontroversial) moral judgments. In this case, Liao et al.'s ɸ is an abstract generalization, which 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'. As Frances Kamm puts it (in a slightly different context):

> [a] framing effect is arbitrary only if people respond differently to puzzle cases that are framed by different descriptions when inspection reveals that the descriptions are equivalent in all morally relevant respects…. Whether two characterizations are morally equivalent - that is, whether they include the same morally relevant factors - is a judgment that depends on moral theory, which picks out some factors as relevant and other factors as irrelevant. (Kamm 1998, 481)

To say that a particular psychological process does not track moral truth is to say that the process generates judgments which are not subjunctively sensitive to certain moral properties. We cannot say this without making some moral judgments ourselves.[11]

3.2 Why relying on moral judgment triggers a regress

So far I have offered a schema for the debunking argument and claimed that any instance of its premise ɸ relies upon some set of moral judgments. I will now argue that this fact about ɸ leads the debunking argument into a regress challenge.

[10] Greene thinks these sorts of harm are permissible. One might also conclude that these sorts of harm are *impermissible*, whether or not up close and personal violence is involved. Either way, we get the judgment that up close and personal violence is morally irrelevant. The locution 'X is a morally irrelevant factor' operates something like this: 'the moral status of an action does not change when factor X is altered'. It is hard to see where we might get a conclusion of that sort, other than by comparing moral judgments about cases with X present to moral judgments about cases without X present.

[11] An error theorist might disagree. If you think there just *are no moral properties*, then obviously you can conclude that a psychological process does not track the moral truth. There is no moral truth to track. And (allegedly) error theorists reach this conclusion without relying on any of their own moral judgments. But notice that error theory does not lead only to *selective* debunking; it leads to global debunking. I will return to this point in section 5.

Suppose we have come to accept some instance of the debunking argument. That is (to rehearse once again), we accept (ψ) that some set *M* of moral judgment is caused by a psychological process *P* and we accept (ϕ) that *P* cannot track the moral truth, so we have concluded that set *M* is unreliable.

Now suppose someone presents us with the following worry: per the argument of section 3.1, our premise ϕ relies upon a set of moral judgments. Call it the ϕ-basis set. The ϕ-basis set is not the same as set *M*. But if we have discovered reason to doubt the reliability of *M* on psychological grounds, shouldn't we at least wonder what psychological investigation of the ϕ-basis set might reveal?

Call this the *generalization worry*. The worry is that, after concluding that psychological investigation makes *M* unreliable, we have to consider whether we should generalize from this to worry about the reliability of other sets of moral judgments, such as the ϕ-basis set. The point here is meant to be intuitive, as shown from comparison to other domains. For instance, if you've just learned that some of your important perceptual experiences are unexpectedly unreliable, then you have at least some reason to wonder about the reliability of other perceptual experiences.

The strength of the generalization worry will depend on how large set *M* is, and on how its members are related to the members of the ϕ-basis set. (I will return to this point in section 4.2.) Suppose, for the moment, that we are only interested in instances of the debunking argument where set *M* is large enough to make the generalization worry reasonable. Hence we *should* now wonder what we might discover about the ϕ-basis set of moral judgments if they were subjected to psychological investigation. What sort of psychological processes cause the ϕ-basis set?

Suppose we (or some psychologist friends) have the means to carry out this investigation. Would it be epistemically responsible to simply *refuse* to do so? I think the answer is no, for two reasons. The first reason is general to justification: if we take ourselves to have default justification in a particular domain, then discover unexpectedly that some important segment of our judgments in that domain are not reliable, we ought to at least check on the reliability of other judgments within that domain. This is precisely what Walter Sinnott-Armstrong says about those who resist psychological debunking on the grounds that only *some* moral judgments have been shown to be unreliable:

> At least when we should know that moral beliefs in general are often subject to distortion, we cannot be justified in trusting any moral belief until we confirm that it is an exception to the rule that most moral beliefs are problematic. (Sinnott-Armstrong 2006, 363)

The second, related, reason is a dialectical one. Proponents of debunking often present fairly limited psychological data; they can cast *direct* empirical doubt over only a limited subset of moral judgments (the ones actually tested in a lab) but they aim to draw a debunking inference about a larger set of intuitions. Roughly, their claim becomes that, now that we (apparently) know that *some* tested moral judgments are unreliable, the burden of proof falls on anyone who claims that certain other untested moral judgments are exempt.

Here is an example of this sort of dialectical move. Tobia et al. (2013) have presented some limited psychological data against the idea that philosophical training confers 'expertise', which here means lowered susceptibility to judgment-distorting factors. In discussing the impact of their results, they write:

> If an advocate of the expertise defense contends that, despite our results, philosophers’ intuitions are much less likely to be subject to framing effects and other problematic influences, he must show that there are a number of other cases in which the intuitions of ordinary folk are subject to framing effects and other problematic influences, and philosophers’ intuitions are not. Stomping one’s foot and insisting that philosophers are experts simply will not do. The issue is an empirical one. And at this point, we submit, the advocates of the expertise defense need to offer some empirical evidence that supports their view. (Tobia et al. 2013, 634)

Proponents of psychological debunking are typically keen to claim that their argument pushes the burden of positive evidence onto their opponents. And if this is true for their opponents, it is equally true for them. If any instance of the debunking argument *itself* relies on certain moral judgments, then proponents of psychological debunking must show that these judgments are reliable.

So, if we can carry out psychological investigation of the ϕ-basis set of moral judgments, we ought to do so. Our confidence in the truth of premise ϕ depends upon the outcome of this psychological investigation. We want to know what psychological process causes the ϕ-basis set of moral judgments, and whether *this* psychological process itself tracks the moral truth. We need to establish this because, if it turns out that the ϕ-basis set is caused by a psychological process that is not moral-truth-tracking, then the ϕ-basis judgments are themselves unreliable, and we cannot trust premise ϕ. An epistemically-responsible instance of the debunking argument needs to be able to confidently assert the following two claims:

ψ': The ϕ-basis set of moral judgments is caused by psychological process *P’*.

ϕ’: Psychological process *P’* *does* track the moral truth (or at least there are no reasons to doubt this).

If we can affirm ψ’ and ϕ’, then we can conclude our debunking argument in an epistemically-responsible way. Together ψ’ and ϕ’ establish that ϕ is supported by truth-tracking judgments, and ϕ licenses our conclusion that set *M* is unreliable.

But here is the problem. We might still raise the following challenge. What is the status of premise ϕ’? Just as with ϕ, any instance of ϕ’ will rely on some set of moral judgments. This is because, just as in ϕ, finding out that a particular psychological process does or does not track moral truth will rely upon some set of moral judgments. There must be a ϕ’-basis set of moral judgments. And now what about the generalization worry? If learning that set *M* are unreliable gave us grounds to psychologically investigate the ϕ-basis set, don’t we also have grounds to psychologically investigate the ϕ’-basis set? Absent further argument, it would seem to be epistemically-irresponsible to accept premise ϕ’ and refuse to carry out such psychological investigation. It seems we will also need to be able to affirm these two claims:

ψ'': The ϕ’-basis set of moral judgments is caused by psychological process *P’’*.

ϕ’’: Psychological process *P’’* *does* track the moral truth (or at least there are no reasons to doubt this).

You can see where this is going. The challenge can be repeated for ϕ’’: what is the ϕ’’-basis set, and is the psychological process causing it a moral-truth-tracking process? And so on. We are headed into a regress: each iteration of ϕ requires a further iteration in order to be made epistemically-responsible.

The regress *challenge* is that this iterative process will never end.[12] There will always be a further psychological investigation to carry out, and a further theoretical premise (relying on further moral judgments) to establish. And if the process never ends, we never acquire suitable grounds for affirming ϕ', and therefore we cannot rely on the ϕ-basis set, and therefore we cannot establish ϕ to complete the argument. If the regress does not terminate somewhere, we never reach the debunking conclusion.

4. Can the regress be halted?

4.1 Criteria for a good halting strategy

So can the regress terminate? In this section I will discuss several strategies for showing that the regress can be brought to an end. I will argue that none are fully satisfactory, at least for the dialectical circumstances in which psychological debunking arguments are typically invoked.

What we are looking for here is a *halting strategy*: some account of why and how the regress stops, and of how we get to a point where we can stably affirm the ϕ-basis set. A halting strategy allows us to affirm the debunking argument itself – that is, to affirm ψ and ϕ – without triggering an endless chain of claims like ψ' and ϕ'. There are therefore two criteria for a halting strategy: it explains why we get to stop, *and* it does so in a way that doesn't undermine our grounds for employing psychologically debunking at all. The second criterion will be clearer in a moment.

First, and obviously, a halting strategy must tell us why we get to stop – why we don't have to keep following the regress down its endless spiral. It is worth taking a moment here to consider what might seem like an equally obvious answer. Here is the answer: we get to stop because the whole endeavor is pointless if it goes on forever. We want to reach a conclusion about a set of moral judgments, and we can only reach that conclusion if we avoid the regress. It is a general fact about belief-confirmation that one can always go on seeking more and more certainty. If I want to know whether there is a spider in my bed, I can check under the sheets once, twice, and so on. What if a really big spider crawled in just after I checked the *third* time? But if I keep checking for spiders, it will eventually be morning and I won't have slept at all. I do care whether I am right about the presence of spiders in my bed, but my caring about this is in service of another goal (sleeping soundly) that is incompatible with an endless hunt for certainty about the matter.

This *seems* like a pretty good response to the regress challenge to psychological debunking. We do want our moral judgments to be reliable, but if we allow ourselves to fall into an endless spiral of psychological reappraisal, we'll never actually reach a conclusion about their reliability and never actually get on with making important moral choices. So we might follow the regress down a few iterations – check up on the psychological causes of the ϕ'''-basis set – but at a certain point, if

[12] Or at least it will never end *for practical purposes*. Supposing that there are a finite number of psychological processes in the human mind, the regress is not literally infinite. But the number is surely so incredibly vast that no actual human being could ever hope to exhaust it.

everything has been satisfactory so far, we just get to stop. When is that certain point? That's not clear – but it is again in the nature of belief-confirmation that we cannot always specify in advance how many iterations confirmation requires. I cannot tell the severe arachnophobe precisely how many times to check the sheets, but I can definitively say that it needs to stop somewhere, and ideally not too long into the night.

But I do not think this is an adequate response to our regress challenge. There is an important difference between hunting for spiders and debunking moral intuitions. The difference is that psychological debunking takes place in a dialectical context where some participants argue that it should not even *start*. A number of moral philosophers believe that psychological debunking fundamentally misconstrues the role of moral judgments in our practical reasoning. Moral judgments, according to these philosophers, have a special authority, such that we could reasonably hang on to them *no matter what* we might discover about their psychological origins. Virginia Held, for instance, writes:

> [T]hose of us interested in ethics should insist, I think, on pursuing our agenda: finding a conception of mind compatible with what we understand and have good reason to believe about moral experience. If this conception of mind is incompatible with cognitive science, so much the worse for cognitive science. It has only been so much the worse for ethics because we have let the agenda be set by those who have little, or only marginal, interest in ethics. (Held 1996, 73)[13]

I do not have space here to describe this view in full. I do not accept it myself (Rini 2013). But its proponents are accredited parties to the debate over the role of psychology in moral methodology. A good strategy for halting the regress will not provide them with further ammunition against the entire project of psychological debunking. Yet so far that seems to be what we are doing. If we admit that we aren't going to ruthlessly follow the psychology wherever it may lead, we concede to our opponents that ultimately it won't be psychology that decides the matter – it will be some sense of ours that it is okay to stop at some point down the regress. Why not, then, stop before we get started? Why engage in psychologizing moral judgment at all?[14] It is as if the arachnophobe's bedmate were saying, 'look, you're just going to check once or twice and then get in bed, even though you know there might still be a spider you've missed. So let's just skip the whole spider-checking business and go to sleep!'

This shows the second criterion for a halting strategy. Not only must a halting strategy explain why we get to *stop* the regress, but it must also show why we should *start*. The answer cannot be simply that psychologizing moral judgment is good in moderation, because there are parties to the debate who deny that psychologizing moral intuition is good at all – and they are happy to use its

[13] Similarly, Ronald Dworkin: "Morality is a distinctive, independent dimension of our experience, and it exercises its own sovereignty. We cannot argue ourselves free of it except by its own leave, except, as it were, by making our peace with it" (Dworkin 1996, 128). See also Nagel (1997).

[14] This seems to be Selim Berker's point when he deploys a regress challenge against Greene: "[S]hould we perform additional experiments to see what parts of the brain light up when certain people make a judgment *that such-and-such-factors-picked-out-by-deontological-judgments are not morally relevant*, and what parts of the brain light up when other people make a judgment *that such-and-such-factors-ignored-by-consequentialist-judgments are in fact morally relevant*? What would that possibly tell us? (Shall we then perform yet more experiments to see what parts of the brain light up when people make a judgment about the relevance of the factors to which those second-order judgments are responding, and so on, ad infinitum?)" (Berker 2009, 327).

susceptibility to the regress as evidence of this fact. A good halting strategy is one that concedes nothing to the opponent of psychological debunking.

4.2 The generalization worry

I briefly alluded to one halting strategy in my introduction of the regress challenge itself. A proponent of debunking might seek to block the *generalization worry*. Recall how the regress got going: if (per debunking) we have reason to doubt the reliability of certain moral judgments (set *M*) we should now do some checking on other judgments in the same domain, such as the ϕ-basis set. The generalization worry is the contamination of the ϕ-basis set by our (apparent) discovery of the unreliability of its co-domain set *M*.

But perhaps the generalization worry is a mistake, at least if set *M* is small or narrow enough. Think about perception again: learning that my perceptual experiences are unreliable in extremely specific and unusual environments (e.g. only while ten meters underwater and wearing mauve-colored goggles) should not lead to much worry about my ordinary perceptual experiences. Similarly, if for some instance of the debunking argument set *M* is very small or narrow, then we might not find the generalization worry compelling, and so have no reason to start down the psychological regress.[15]

Obviously the plausibility of this strategy will depend on the details of set *M*. If one aims to debunk only a very small and narrow set of moral judgments, such as moral judgments made after taking a particular narcotic, then the generalization worry can probably be dismissed and the regress avoided. But notice that the examples discussed above are not like this. Quinn's intuitions about the Doctrine of Double Effect (targeted by Horowitz) and general deontic intuitions about moral dilemmas (targeted by Greene) are related to large numbers of moral judgments – this is precisely why debunking them is taken to be important to moral theory. Even worse, 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.[16]

So although it is possible in principle to halt the regress by rejecting generalization from *M,* it is doubtful that this would work for typical instances of the debunking argument, which are aimed at sets of moral judgments large enough to be interesting to general moral theory. At minimum, anyone who wishes to advance a particular instance of the debunking argument owes an explanation of how their targeted set *M* is small and narrow enough to avoid generalization.

[15] To be precise, the issue here isn't directly about the size of set *M*. Rather, the real issue is the relationship between the judgments that are elements of *M* and the judgments that are elements of the ϕ-basis set. Are these similar *types* of moral judgments? Similar to what degree? But it's not an easy matter to provide a taxonomy of moral judgment types or a metric of similarity space. Since these are messy questions, I use set size as a proxy for similarity, on the purely mathematical assumption that a large set *M* is more likely to have elements similar to those of ϕ-basis set than a small set *M*.

[16] Though see Wiegmann et al. (2012) for suggestive evidence that order effects may affect moral judgments only under certain circumstances.

4.3 The limits of psychological science

A different halting strategy simply points to the limits of science. The regress challenge is conditional on what present psychological science makes possible: it says that *if* we can study the psychological causes of the ϕ-basis set, *then* we ought to do so. But what if the present state of psychological science is limited, such that while we can study the psychological causes of set *M*, we cannot study the psychological causes of the ϕ-basis set? If that is the case, then the proponent of debunking can call a halt to the regress. After all, if it is not possible for us to carry out a particular sort of psychological inquiry, then we can't be faulted for not doing so!

This strategy, if it turns out to be correct about the limits of psychological science, would seem to fulfill our criteria for a good halting strategy. It explains why we stop – because the science just doesn't allow us to go on – in a way that poses no problem for having gotten started. But, of course, at the moment we have no reason to assume this is how things will turn out. And if they do turn out that way, it's likely that this will be only a temporary reprieve. The techniques of psychological science are constantly improved, so that what is now an empirically intractable question may be resolved in a few years.

In fact, it's not clear that this would be a very satisfying way of halting the regress even if the scientific problems were irresolvable. Suppose that there is some in-principle limit on the ability of psychological science to investigate cognitively complex moral judgments, such that at a certain point down the regress we just *cannot* ever hope to understand the psychological process responsible for a particular set of judgments. In that case, we are in this position: we want to disqualify certain judgments (set *M*), on the grounds of our other judgments (the ϕ-basis set) about their psychological causes. We admit that it *might* be true that the ϕ-basis set is unreliable – in which case we are mistaken to discard set *M* – but we have no way of knowing whether this is true due to the limits of science.

If that turns out to be our situation, is it really obvious that we ought to discard set *M* after all? Isn't it a reasonable option to regard the status of set *M* as unresolved, since the evidence of its unreliability depends on the empirically irresolvable status of the ϕ-basis set? If that is a reasonable option, then the argumentative force of psychological debunking disappears. The logic of the regress argument intercedes even if the science is limited.

4.4 A self-approving process

I'll discuss one final halting strategy. In fact, this isn't quite a *halting* strategy, as it does not try to stop the regress. The claim here is that the regress becomes *unproblematic* under certain circumstances. Suppose that the regressing chain of premises turns out to include a loop. That is, somewhere down the line appears a sequence like this:

ψR: The ϕ-basis set of moral judgments is caused by psychological process *PR*.

ϕR: Psychological process *PR* <u>does</u> track the moral truth (or at least there are no reasons to doubt this).

ψRR: The ϕR-basis set of moral judgments is caused **by psychological process *PR*** [that is, the *same PR* in premise ψR].

ϕRR: Psychological process *PR* <u>does</u> track the moral truth (or at least there are no reasons to doubt this).

In this sequence, psychological process *PR* is shown to generate moral judgments *about itself*. (Or, more precisely, it generates moral judgments that support a claim affirming that the process itself tracks the moral truth.) If such a self-approving psychological process does turn up, then this sequence will fill the rest of the regress. It will continue to go on endlessly – but it does so in a harmless way. Remember that the regress is only a problem for debunking because of the danger that somewhere down the chain, there might emerge a reason to doubt the reliability of the ϕ-basis set. But this sequence does not provide such a reason, and if we see that this is what will repeat endlessly, we can comfortably conclude that the ϕ-basis set is secure. So, if a self-approving process turns up, then the regress is painless and can be ignored.[17]

I am prepared to concede that the regress can be made painless if psychological investigation turns up a self-approving process. We *can* get away with psychological debunking of particular moral judgments, if we can show that the theoretical premise in the debunking argument is ultimately grounded in a self-approving moral psychological process. But at the moment we have absolutely no evidence for such a thing. Until we find it, the regress challenge presses on any selective psychological debunking argument.

5. Global debunking and selective debunking: why selectivity leads to regress

In the previous section, I argued that some possible halting strategies are inadequate (or at least, in the last case, depend on facts not in evidence). In this section I will explain *why* it is so difficult to come up with a good halting strategy. I will show that the regress problem is fundamentally rooted in the nature of *selective* debunking itself. To see this, we need to return to that other sort of debunking, the sort Sidgwick himself charged with regress worries. Surprisingly, it turns out that a form of *global* debunking can evade my regress – precisely because it is *not* selective debunking.

Return again to the psychological debunking schema:

(ψ): A set of moral judgments *M* is caused by psychological process *P*.

(ϕ): Psychological process *P* does not track the moral truth.

Conclusion: Therefore, moral judgments *M* are unreliable.

[17] In fact, if there is such a self-approving psychological process, then it might be understood as the naturalistic foundation of morality itself. Christine Korsgaard (describing a view she attributes to Hume) argues that morality must meet the *reflectivity test*. "Call a purportedly normative judgment a 'verdict' and the mental operation that gives rise to it a 'faculty'.... [A] faculty's verdicts are normative if the faculty meets the following test: *when the faculty takes itself and its own operations for its object, it gives a positive verdict*" (Korsgaard 1996, 62; italics in original).

A global debunking argument is one in which set *M* is the set of all moral judgments. Notice, then, that a global debunking argument *must* avoid relying on particular moral judgments to motivate its premise ϕ. Since its conclusion directly entails the unreliability of *any* moral judgment, it would be immediately self-defeating if its premises required their truth. So if there *is* a successful global debunking argument, it must be one that avoids triggering the regress challenge. Is there a successful global debunking argument?

I will not try to answer that question completely. But I will show that at least one prominent global debunking argument appears to avoid being immediately self-defeating or triggering a regress. Whether this argument fails for other reasons is a matter I leave for another time. The argument comes from Richard Joyce (2006a).[18] Joyce defends the following premises:

(1) We have an empirically confirmed theory that explains the origins of our moral judgments;
(2) This theory does not state, imply, or presuppose that these judgments are true.

According to Joyce, if both of these premises are right, then we ought to conclude that our moral judgments are undermined. We cannot rely on them. As he puts it: "we have no grounds one way or the other for maintaining these beliefs. They *could* be true, but we have no reason for thinking so." (Joyce 2006, 211).[19]

For our purposes, premise (2) is the interesting premise. It seems to be a generalized form of ϕ; it can be read as saying that, taken together, the psychological processes that produce all our moral judgments do not track the moral truth (or at least that we have no grounds for supposing that they do).

Suppose that we tried to catch Joyce in a regress of the sort featured in section 3. We say to him: what are your grounds for making claim (2)? If it is some set of moral judgments, then the machinery of the regress challenge starts turning.

But if we look at how Joyce argues for claim (2), it's not obvious that he *does* rely upon any set of moral judgments. In fact, he relies only on a *conceptual* claim about the nature of morality. The claim is this: it is an essential feature of the concept of moral reasons that they involve inescapable authority, or as Joyce calls it, 'practical clout' (see Joyce 2006a, 191-193). It is distinctive of moral concepts that our experiencing them as compelling does not depend on our desires or interests. When I think that I am morally obligated to do X, I think that my being obligated to do X does not depend on my having a certain preference or belonging to a certain group. In effect, I am judging that I have reason to do X, *whatever* I might want to do.

[18] See also Harman (1977), on which Joyce relies, and also Street (2006) for a somewhat similar argument (though Street does not intend the conclusion to be debunking in quite the same way as Joyce). Though Joyce's book deals centrally with evolution, the logic of his debunking argument does not require any particularly Darwinian premise. Indeed, in another publication (Joyce 2006b) he employs the same argument starting from points that are wholly general to the empirical sciences.

[19] There is a sense in which Joyce's debunking *is* selective: he wishes only to debunk moral judgments, not judgments in other normative domains. This is important, because if his argument were truly global – that is, applied to all normative domains – then it would risk self-defeat. The epistemic norms underwriting his empirical claims, and even the inferential threads of the argument itself, might be evolutionarily debunked. (See Joyce 2006a, 183-184, for a brief discussion of the point.) I leave this issue to the side for now, since my interest is only in moral debunking; for my purposes, 'global' and 'selective' are relative to the moral domain.

How does this claim about practical clout lead to (2)? According to Joyce, practical clout is in tension with scientific naturalism. He argues (pp. 193-198) that there is no "naturalistically respectable" account of practical clout. Nothing in our best naturalistic explanations of terms like 'reason' or 'desire' supports a claim that there are considerations with inescapable authority. There may be *non*-naturalistic accounts of practical clout, but then these will not be captured in empirical investigation.

So, if moral judgments essentially involve practical clout, and if practical clout is no part of naturalistically respectable science, then we get (2): an empirically confirmed theory that explains the origins of our moral judgments will not state, imply, or presuppose that these judgments are true.

We might challenge either of Joyce's assertions that practical clout is essential to morality or incompatible with naturalism.[20] I will not discuss these points here, as my aim is not to defend Joyce's view. Our immediate interest is in the logic, and in particular how this argument avoids the regress challenge. Joyce does not need to rely upon any particular moral judgments in order to motivate his premise (2). It follows from a purely conceptual claim about the nature of morality, a claim which is supposed to be clear *a priori* and not as a result of affirming any specific moral judgment.

If there is any way in which Joyce's (2) relates to *particular* moral judgments it is not the way in which ϕ relies upon the ϕ-basis set. To justify Joyce's claim about practical clout, we needn't *affirm* any particular moral judgments. All we need to do is reflect on our own personal moral experience – how it *feels* to view an action as morally wrong, or another as morally mandatory. 'Practical clout' is just a description of that motivational force, which Joyce takes to be essential to the moral domain. Crucially, we can make this point without assuming that any of our moral judgments are actually *true*. We could agree with Joyce that the experience of making moral judgments essentially includes this practical clout feature without having to affirm the judgments themselves (just as we might agree about how we are moved by our aesthetic experience without necessarily assuming that our aesthetic experiences are truth-related). Since Joyce needn't assume the truth of any moral judgments, he needn't worry about the regress. And, combined with the claim about practical clout being absent from naturalistically respectable science, this gets Joyce to his debunking conclusion.

It is as if Joyce were saying: look, if you thought more carefully about the *concept* 'spider', you would realize that your concept essentially involves certain non-natural properties. It turns out that the concept 'spider' involves (for instance) lack of spatial extension.[21] Properly understood, if spiders exist at all (maybe they don't!) then it is only as abstract objects lacking any physical instantiation. Regrettably, this is not a very plausible analysis of the concept 'spider'. But if it *were*, then Joyce could cast doubt on the claim that my cat's vision tracks the truth about spiders without getting into an argument over whether-or-not any particular creature is a spider. Our best account of my cat's

[20] For instance, see Finlay (2008), Stich (2008), or Tresan (2010).

[21] If this just sounds insane, rather than merely ridiculous, then keep in mind that in *Yates vs. United States*, the American Supreme Court decided that some fish are not tangible objects. See 'In Overturning Conviction, Supreme Court Says Fish Are Not Always Tangible' (*New York Times*, February 25, 2015; http://www.nytimes.com/2015/02/26/us/justices-overturn-a-fishermans-conviction-for-tossing-undersize-catch.html). Justice Kagan disagreed with the majority: "A fish is, of course, a discrete thing that possesses physical form," she said, and cited Dr. Seuss.

visual system will not state, imply, or presuppose the truth of any claims about uninstantiated abstract objects. Hence, if that really were an essential property of the concept 'spider', then we could conclude that my cat's vision is not a reliable guide to the truth about spiders without getting into the frightful business of spider-detection testing.

So, let's now grant that Joyce's global debunking argument avoids the regress challenge. Could we use his strategy to motivate any sort of *selective* debunking? I do not think so, and it is useful to consider why not. Joyce's emphasis on an *essential* feature of the moral domain (practical clout here) is what helps him avoid the regress, but this move is unavailable to a selective debunker. I'll now try to illustrate this point by showing how two examples of selective debunking fail to fit Joyce's mold.

We've already looked at one of these examples. Recall Lazari-Radek and Singer's (2012) selective debunking of egoistic and partial moral judgments. Lazari-Radek and Singer argue that it is an essential feature of moral judgments that *if* they are true, they are true in such a way that it would be an implausible coincidence for evolution to happen upon the right answer (the core of this argument comes from Street (2006)). Then, Lazari-Radek and Singer claim, egoistic and partial judgments are most plausibly understood as the result of an evolutionary process, whereas impartial judgments are not. Hence we have reason to doubt the reliability of egoistic and partial judgments, but not impartial ones. This argument, if successful, would appear to be a form of selective debunking that evades the regress: like Joyce's argument, it grounds its claims about truth-tracking on purportedly essential features of the moral domain (that its truths are non-natural if construed realistically), rather than claims about particular psychological processes.

The problem is that this argument isn't as selective as Lazari-Radek and Singer would like it to be. As Kahane (2014) points out, consequentialist intuitions (which Lazari-Radek and Singer want to exempt from debunking) lack content without some supplementary account of well-being. Yet, Kahane continues, our judgments about what would constitute well-being have been shaped by evolutionary forces just as surely as have the moral judgments Lazari-Radek and Singer have targeted for debunking. The point here generalizes: given the centrality of evolution in shaping human evaluative processes, it seems unlikely that we have *any* moral judgments completely exempt from evolutionary explanation. Hence, if Lazari-Radek and Singer's argument is debunking, it is not *selectively* debunking.[22]

Now consider another earlier example: Liao and colleagues (2011)'s evidence for order effects in trolley loop case judgments. We could imagine constructing the following selective debunking argument from this evidence. (What follows is *not* what Liao et al. say – it is an illustrative demonstration of how their findings *could* be employed.) It is, let us claim, essential to the moral domain that the truth of a set of moral judgments cannot depend on the order in which they are presented. This is a conceptual claim about morality, not apparently dependent upon any particular moral judgments. And if it turns out that only certain moral judgments (such as those about the loop

[22] Perhaps this should not be surprising, since the Street (2006) evolutionary argument on which they rely is explicitly targeted at *all* evaluative attitudes.

case) display order effects, then it seems that we have a *selective* debunking argument that evades the regress in just the way Joyce's does.[23]

But there is a problem here as well. This time the problem is with the idea that the order-insensitivity of moral judgments is an essential truth we can know without relying on any particular moral judgments. Consider: there *are* some cases in which the truth (or at least the proposition-assertibility) of a judgment can depend on order of presentation. For instance, some pairs of counterfactual judgments seem both true when presented in one order, but not both true when the order is reversed.[24] Now we are not tempted to see moral claims in this way, but my point is only this: as a purely conceptual matter, it *could* turn out that the truth of moral judgments is sensitive to their order of presentation. The fact that we easily insist upon the order-insensitivity of moral judgments is not a conceptual matter. It comes from immediate employment of moral judgment itself – we consider the variant loop cases in one order and then another, and we judge that each case has the same moral status in either position in the order. But then this assessment *does* depend upon particular moral judgments – and so does not evade the regress challenge.

Obviously there are very deep issues, about justification and the nature of conceptual domains, hanging not far in the background here. I have not been able to explore these issues at great depth, and so I hesitate from making overly sweeping claims. Still, the two examples just discussed seem to show two ways in which putative selective debunking arguments are likely to go awry. They can turn out to be global rather than selective, as in the Lazari-Radek and Singer example. Or they can turn out to rely upon particular moral judgments for their seemingly conceptual grounding, as in the example based upon Liao et al., and so tip back into the regress challenge.

I want to suggest that this bifurcation is built into the nature of psychological debunking. Debunking arguments that avoid the regress are those that (like Joyce's argument from practical clout) do not rely upon the truth of any particular moral judgments. But the only apparent way to do this is to appeal to conceptual claims about essential features of the moral domain – that is, features essential to *all* moral judgments, features that will be apparent to us *no matter which* particular moral judgments we think are true (or indeed even if we conclude that none are true). Since these are features of all moral judgments, they can ground only global debunking, not selective debunking.

I leave open whether global debunking arguments like Joyce's are vulnerable to other sorts of regress challenge, such as the one from Sidgwick discussed at the start. The lesson here is only this: the regress challenge explored in this paper follows *distinctively* from features of selective psychological debunking. So if psychological debunking of moral judgments works, it works globally.

[23] I owe this example to an anonymous referee.

[24] These are reverse Sobel sequences. Consider the following two counterfactuals, first in this order, and then in reverse order: (C1) If Sophie had gone to the parade, she would have seen Pedro. (C2) But if Sophie had gone to the parade and been stuck behind a tall person, she would not have seen Pedro. There is an active dispute over whether the problem exposed by this reversal reflects something deep in the semantics of counterfactuals or is a pragmatic effect. See von Fintel (2001) and Moss (2012).

6. Empirically-informed moral theory?

If psychological debunking did work, it would be quite useful. Moral philosophy often aspires to address moral *disagreement*. We want to be able to settle a dispute between two inconsistent moral judgments (whether between two moral judges, or within a single judge). Yet inevitably moral philosophers end up in intractable disputes about principles and intuitions. Psychological debunking seems to offer another way through the mess, by cutting down certain moral judgments as unreliable, leaving those that remain to guide us. Peter Singer expresses the hope like this: “Precisely because science is outside ethics, the scientific study of our ethical judgment is a fulcrum on which we can rest our critical lever” (Singer 1981, 73).

But if the argument of this paper is right, the prospects for such an Archimedean role for psychology are not bright. Selective debunking, I’ve argued, faces a regress challenge. And global debunking is not suited to arbitrate moral disagreement, as it undermines one moral judgment only if it undermines them all.

So it might seem that the lesson of this paper is dismal for one of the central ambitions of the movement toward psychologically-informed moral theory.[25] However, this is only true if we understand the relationship of empirical psychology to moral theory construction in purely epistemic terms. The debunking argument, for all its psychological baubles, is fundamentally a move in moral epistemology; it aims to show that our moral judgments are unreliable indicators of the moral *truth*. I’ve argued here, also on epistemic grounds, that such an argument faces a troubling regress.

Yet that conclusion is compatible with seeing a *non-epistemic* role for empirical psychology in moral theory. I don’t have the space here to fully explain such a role, but I offer a sketch simply to make the idea clear. In my view, the epistemic approach loses track of something important to the nature of moral judgments. Moral judgments are in part *willful*. They are in part about *making* ourselves and *committing* ourselves to certain courses of action – and to declaring certain values central, to giving expression to our identities. This *self-constituting* conception of morality’s aims has a long tradition, which has been increasingly influential in recent moral theory.[26] On this conception, the importance of moral psychology is that it allows us to become aware of how we are *actually* psychologically constituted, and to make informed, reflective choices about how to achieve a better reconstitution.

Recognizing the self-constituting aim of moral judgment makes clear that the lesson of this paper is about *how* to empirically inform moral theory, not about *whether* to do so. In my view, much of the existing literature on the normative significance of cognitive science has incorrectly assumed that the *only* way cognitive science could impact moral theory is via moral epistemology. If this were true, then the regress challenge would indeed be a serious barrier to empirically-informed moral theory. But it is not at all obvious that the regress challenge arises for the self-constitution approach. The regress challenge relies on certain concepts (‘truth-tracking’, ‘reliability’) that are distinctively epistemic, and may have no relevant counterparts in an inquiry focused on self-constitution.

[25] Representatives of this movement appear in Doris and Stich (2007), Appiah (2008), Levy (2009), Christen et al. (2014).

[26] See, for instance: Korsgaard (2009), Velleman (2009), Oshana (2010).

Hence there are two lessons to this paper. First, if you are inclined to engage in selective psychological debunking of moral judgments, you will need to deal with a regress challenge. Second, if you are sympathetic to the idea that empirical psychology can play *some* role in guiding our moral reasoning, then it may be time to look beyond epistemology, and think more about how psychological self-understanding can aid the project of making ourselves into moral agents.[27]

References

Appiah, Kwame Anthony. 2008. *Experiments in Ethics*. Harvard University Press.

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

Blackburn, Simon. 1985. "Errors and the Phenomenology of Value." In *Morality and Objectivity*, edited by Ted Honderich. Routledge & Kegan Paul.

Christen, M., C. van Schaik, J. Fischer, M. Huppenbauer, and C. Tanner, eds. 2014. *Empirically Informed Ethics: Morality between Facts and Norms*. Library of Ethics and Applied Philosophy 32. Springer. http://www.springer.com/social+sciences/applied+ethics/book/978-3-319-01368-8.

Doris, John M., and Stephen Stich. 2007. "As a Matter of Fact: Empirical Perspectives on Ethics." In *The Oxford Handbook of Contemporary Philosophy*, edited by Frank Jackson and Michael Smith, 1st ed., 1:114–53. Oxford: Oxford University Press. http://www.oxfordhandbooks.com/oso/public/content/oho_philosophy/9780199234769/oxfordhb-9780199234769-chapter-5.html.

Dworkin, Ronald. 1996. "Objectivity and Truth: You'd Better Believe It." *Philosophy & Public Affairs* 25 (2): 87–139.

Finlay, Stephen. 2008. "The Error in the Error Theory." *Australasian Journal of Philosophy* 86 (3): 347–69.

Gibbard, Allan. 2003. *Thinking How to Live*. Cambridge, MA: Harvard University Press.

Greene, Joshua D. 2008. "The Secret Joke of Kant's Soul." In *Moral Psychology, Vol. 3. The Neuroscience of Morality: Emotion, Brain Disorders, and Development*, edited by Walter Sinnott-Armstrong, 35–80. Cambridge, MA: MIT Press.

Harman, Gilbert. 1977. *The Nature of Morality: An Introduction to Ethics*. New York: Oxford University Press.

Held, Virginia. 1996. "Whose Agenda? Ethics versus Cognitive Science." In *Minds and Morals: Essays on Cognitive Science and Ethics*, edited by Larry May, Marilyn Friedman, and Andy Clark, 69–88. Cambridge, MA: MIT Press.

Horowitz, Tamara. 1998. "Philosophical Intuitions and Psychological Theory." *Ethics* 108 (2): 367–85.

Joyce, Richard. 2006a. "Metaethics and the Empirical Sciences." *Philosophical Explorations* 9 (1): 133–48.

———. 2006b. *The Evolution of Morality*. 1st ed. MIT Press.

Kahane, Guy. 2011. "Evolutionary Debunking Arguments." *Noûs* 45 (1): 103–25. doi:10.1111/j.1468-0068.2010.00770.x.

[27] The development of this paper has benefited immensely from detailed comments by Guy Kahane, Joanna Demaree-Cotton, and an anonymous referee for *Philosophical Studies*. I am also indebted to discussion with Hilary Greaves, Hossein Dabbagh, Nora Heinzelmann, Ned Block, Michael Strevens, Joshua Knobe, Thomas Nagel, Daniela Dover, Knut Olav Skarsaune, and Sharon Street. The ideas have been sharpened by engagement with audiences at NYU, Oxford, VU University Amsterdam, the University of Sydney, the University of Tokyo, Kansas State University, and UC Davis. This research was supported by the VolkswagenStiftung's European Platform for Life Sciences, Mind Sciences, and the Humanities (grant II/86 063).

———. 2012. “On the Wrong Track: Process and Content in Moral Psychology.” *Mind & Language* 27 (5): 519–45. doi:10.1111/mila.12001.

———. 2014. “Evolution and Impartiality.” *Ethics* 124 (2): 327–41. doi:10.1086/673433.

Kahneman, Daniel, and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision under Risk.” *Econometrica* 47 (2): 263–91. doi:10.2307/1914185.

Kamm, F. M. 1998. “Moral Intuitions, Cognitive Psychology, and the Harming-Versus-Not-Aiding Distinction.” *Ethics* 108 (3): 463–88.

———. 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York ;Oxford University Press.

Korsgaard, Christine M. 1996. *The Sources of Normativity*. Cambridge University Press.

Korsgaard, Christine M. 2009. *Self-Constitution: Agency, Identity, and Integrity*. New York: Oxford University Press.

Lazari-Radek, Katarzyna de, and Peter Singer. 2012. “The Objectivity of Ethics and the Unity of Practical Reason.” *Ethics* 123 (1): 9–31. doi:10.1086/667837.

Levy, Neil. 2009. “Empirically Informed Moral Theory: A Sketch of the Landscape.” *Ethical Theory and Moral Practice* 12 (1): 3–8. doi:10.1007/s10677-008-9146-2.

Lewis, David. 2000. “Causation as Influence.” *Journal of Philosophy* 97 (4): 182–97.

Liao, S. 2008. “A Defense of Intuitions.” *Philosophical Studies* 140 (2): 247–62. doi:10.1007/s11098-007-9140-x.

Liao, S. Matthew, Alex Weigmann, Joshua Alexander, and Gerard Vong. 2011. “Putting the Trolley in Order: Experimental Philosophy and the Loop Case.” *Philosophical Psychology* 25 (5): 661–71.

Mason, Kelby. 2011. “Moral Psychology And Moral Intuition: A Pox On All Your Houses.” *Australasian Journal of Philosophy* 89 (3): 441–58. doi:10.1080/00048402.2010.506515.

Moss, Sarah. 2012. “On the Pragmatics of Counterfactuals.” *Noûs* 46 (3): 561–86.

Nagel, Thomas. 1997. *The Last Word*. Oxford: Oxford University Press.

Nozick, Robert. 1981. *Philosophical Explanations*. Harvard University Press.

Oshana, Marina. 2010. *The Importance of How We See Ourselves: Self-Identity and Responsible Agency*. Lanham, Md.: Lexington Books.

Quinn, Warren S. 1989. “Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing.” *The Philosophical Review* 98 (3): 287–312. doi:10.2307/2185021.

Rini, Regina A. 2013. “Making Psychology Normatively Significant.” *The Journal of Ethics* 17 (3): 257–74. doi:10.1007/s10892-013-9145-y.

Sandberg, Joakim, and Niklas Juth. 2011. “Ethics and Intuitions: A Reply to Singer.” *The Journal of Ethics* 15 (3): 209–26. doi:10.1007/s10892-010-9088-5.

Sidgwick, Henry. 1907. *The Methods Of Ethics*. Hackett.

Singer, Peter. 1981. *The Expanding Circle: Ethics and Sociobiology*. Oxford: Oxford University Press.

———. 2005. “Ethics and Intuitions.” *Journal of Ethics* 9 (3-4): 331–52.

Sinnott-Armstrong, Walter. 2006. “Moral Intuitionism Meets Empirical Psychology.” In *Metaethics After Moore*, edited by Terry Horgan and Mark Timmons, 339–66. Oxford: Oxford University Press.

———. 2008. “Framing Moral Intuition.” In *Moral Psychology, Vol 2. The Cognitive Science of Morality: Intuition and Diversity*, 47–76. Cambridge, MA: MIT Press.

Stich, Stephen. 2008. “Some Questions About ‘The Evolution of Morality.’” *Philosophy and Phenomenological Research* 77 (1): 228–36.

Street, Sharon. 2006. “A Darwinian Dilemma for Realist Theories of Value.” *Philosophical Studies* 127 (1): 109–66.

Thomson, Judith Jarvis. 1976. “Killing, Letting Die, and the Trolley Problem.” *The Monist* 59 (2): 204–17.

Tobia, Kevin, Wesley Buckwalter, and Stephen Stich. 2013. “Moral Intuitions: Are Philosophers Experts?” *Philosophical Psychology* 26 (5): 629–38. doi:10.1080/09515089.2012.696327.

Tresan, Jon. 2010. “Question Authority: In Defense of Moral Naturalism Without Clout.” *Philosophical Studies* 150 (2): 221–38.

Van Roojen, Mark. 1999. “Reflective Moral Equilibrium and Psychological Theory.” *Ethics* 109 (4): 846–57.

Velleman, J. David. 2009. *How We Get along*. New York: Cambridge University Press.

Von Fintel, Kai. 2001. “Counterfactuals in a Dynamic Context.” In *Ken Hale: A Life in Language*, edited by Michael Kenstowicz, 123–52. Cambridge, MA: MIT Press.

Wiegmann, Alex, Yasmina Okan, and Jonas Nagel. 2012. “Order Effects in Moral Judgment.” *Philosophical Psychology* 25 (6): 813–36. doi:10.1080/09515089.2011.631995.