PHILOSOPHICAL PERSPECTIVES

Philosophical Perspectives, 33, 2019
doi: 10.1111/phpe.12130

LOCAL EVOLUTIONARY DEBUNKING ARGUMENTS

Richard Rowland
School of Philosophy, Religion and History of Science, University of Leeds

0. Introduction

Evolutionary debunking arguments in ethics aim to use facts about the evolutionary causes of ethical beliefs to undermine their justification. *Global* Evolutionary Debunking Arguments (GDAs) aim to undermine the justification of all ethical beliefs–or all beliefs in non-analytic ethical truths realistically construed. *Local* Evolutionary Debunking Arguments (LDAs) aim to undermine the justification of only some of our ethical beliefs. GDAs are often used in arguments for skeptical or anti-realist metaethical views.[1] In contrast, LDAs are employed in arguments in normative ethics. Singer (2005) and Greene (2008: 43, 76) employ LDAs to argue that we should accept utilitarianism rather than a deontological view. And de Lazari-Radek and Singer (2012) use a LDA to argue that we should accept utilitarianism rather than egoism.

Kahane (2011) (2014), Rini (2016), Tersman (2008), and Vavova (2014) argue for skepticism about the possibility of LDAs. They argue that LDAs cannot be successful because they over-extend in a way that makes them self-undermining: if LDAs succeed, they undermine the justification of the ethical beliefs that their proponents wish to hold onto or lead to a form of moral skepticism. In this paper I argue that these arguments for skepticism about the possibility of LDAs are misplaced.

Assessing the plausibility of skepticism about LDAs is important for several reasons. First, the LDA for act-consequentialism made by Greene, de Lazari-Radek, and Singer is perhaps the most interesting new argument for act-consequentialism made in the last few decades. If good LDAs are impossible, as Kahane, Rini, Tersman, and Vavova allege, this would be a damning indictement of this argument, and would have a bearing on the more general contemporary case for act-consequentialism. More generally, the strategy of making a LDA in arguing for a particular ethical position is now relatively commonplace; for instance, LDAs are used against the belief that incest is always morally wrong and in arguments in defence of hedonism against the experience machine

objection.[2] If skeptics about LDAs are right, we should stop making LDAs in normative ethics altogether. Finally, it would be quite surprising if LDAs could not work at all. For generally, the causal history of our beliefs can undermine their justification: if we come to know that we only have a particular belief because we were hypnotised to have it a few weeks previously, this undermines the justification of our belief. And proponents of skepticism about LDAs share this view.[3] So it would be surprising if there were something about evolution or ethics that precluded the debunking of particular ethical beliefs via facts about their evolutionary causes.

My aim in this paper is not to argue for any particular ethical view. Rather my aim is to assess Kahane, Rini, Tersman, and Vavova's skepticism about LDAs and get clear on LDAs, when and where they can be made, and their prospects of success. But in order to show that Kahane et al.s skepticism about LDAs fails I need to show that LDAs can be made by proponents of particular ethical views without these LDAs undermining the ethical beliefs of their proponents. It seems to me that proponents of desire-constrained accounts of reasons—proponents of a first-order internalist account of practical reasons—are most clearly able to make LDAs without such self-defeat. So, I argue that such internalists can make LDAs. But because many ethicists are externalists rather than internalists, I also attempt to show that externalists about reasons can make LDAs. But readers shouldn't read too much into my arguing that LDAs can be made by internalists or particular kinds of externalists; I don't intend to endorse or argue for internalism or a form of externalism here. These arguments are only intended to show the possibility of making LDAs without self-defeat.

In §1 I sketch two prominent LDAs that Kahane, Tersman, and Vavova target. I then outline Kahane, Tersman, and Vavova's similar arguments that not only do these particular LDAs fail, but the way in which they fail shows that there is a more general problem with LDAs that should lead us to a general skepticism about the possibility of non-self-undermining LDAs. In §2 I argue that we should understand the epistemic principle and key claims of LDAs in a particular way and I argue that this way of understanding and constructing LDAs shows that Rini's argument for skepticism about LDAs fails. In §3 I argue that, given the way of constructing LDAs proposed in §2, proponents of internalist accounts of normative reasons can make LDAs that are not self-undermining in the way that Kahane, Tersman, and Vavova claim that all LDAs are. In §4 I argue that certain externalists about normative reasons can also make LDAs that are not self-undermining. In §5 I briefly sketch the slightly complicated implications of my argument for the work that LDAs can do in normative ethics.

1. LDAs and Skepticism about LDAs

LDAs normally aim to undermine the epistemic status of particular beliefs by undermining the status of the judgments they are based on which include moral intuitions (Tersman 2006: 391). We can understand moral intuitions broadly as moral judgements that are accepted by someone not merely on the basis that they follow from a moral theory or principle that they accept or more specifically as intellectual seemings on the basis of which we come to hold certain particular moral beliefs *(ibid*.; Stratton-Lake 2016: §1.1).

Kahane (2011) understands LDAs as having two components. First, a normative epistemic component which involves a claim about how irrelevant influences on our judgments—such as evolution—can undermine the justification of our beliefs based on these judgments. Second, a causal empirical component claiming that particular evolutionary factors caused particular ethical judgements of ours. For the time being we can understand the epistemic component to involve the principle that if our belief that *p* is the result of a process that is irrelevant to the truth of whether *p*, then our belief that *p* is not justified. To see how a principle along these lines can seem plausible, suppose that you see an object in front of you that seems blue to you. You then come to know that you've taken a pill that makes you see red things as blue things. It seems that your justification for believing that the object is blue, which you previously had based on your perceptual seeming, is undermined.

1A. LDA's Empirical Claims

The empirical component of a LDA has two parts: first, *a debunking claim*, that evolutionary factors caused particular ethical beliefs of ours; second, a *vindicatory claim*, that there are ethical beliefs of our which are not similarly explained by evolutionary irrelevant influences.

Singer (2005) articulates a LDA. Its debunking claim holds that our deontological intuitions, such as the widespread intuition that we should not push the heavy man to his death to save five people in the footbridge trolley case, are the result of a non-moral-truth-tracking evolutionary process. He argues that we have evolved to have immediate, strong, negative emotional responses to hitting, pushing, or strangling others but not to have such reactions to the infliction of violence by switch turning because historically we lived in small groups where violence could only be inflicted in such an up close and personal way (*ibid*. 347–348). In contrast, according to Singer's LDA's *vindicatory claim*, no such evolutionary story can be told for the intuition that it is bad if one person is killed and the (impartial consequentialist) intuition that it is less of a tragedy if 1 person is killed than if 5 people are killed. He claims that this intuition 'does not seem to be one that is the outcome of our evolutionary past' because it is not evolutionary fitness-enhancing for us to have the intuition that it is wrong or bad

that those who are strangers to us are killed; we have no reason to believe that 'love of mankind, merely as such' would have evolved through natural selection (*ibid*. 350).

Non-derivative reasons to φ are reasons to φ the normative force of which does not derive from the normative force of other reasons to φ. For instance, according to one view about promise-keeping, there are (non-derivative) reasons for us to keep our promises even if no good would come of our keeping them. According to another view, the only reasons for us to keep our promises are (derivative) reasons that derive from the good consequences of our keeping our promises or our being disposed to keep them. De Lazari-Radek and Singer (2012: 22) make a LDA. Its debunking claim targets (non-derivative) partial and self-interested reasons. They argue that the judgment or intuition that we have stronger (non-derivative) reason to help our own children than to help complete strangers just because they are our children is likely to lead to reproductive success. This is because making such a judgment disposes one to do more to help one's own children than strangers, and individuals who help their own children more than strangers are more likely to have children who survive and reproduce. And they similarly argue that having the judgment that we have non-derivative self-interested reasons, that is, stronger reasons to promote our own interests rather than others (just because they are our own), would lead to reproductive success—by leading us not to sacrifice ourselves for complete strangers for instance.

They then argue that Sidgwickian self-evident intuitions in universal benevolence cannot be similarly evolutionarily debunked; this is their vindicatory claim. An intuition is self-evident if it is an intuition that is not derived from another intuition; self-evident intuitions are intuitions that are 'evident in and of themselves' rather than intuitions that one has on the basis of an argument.[4] The Sidgwickian self-evident intuitions that de Lazari-Radek and Singer (2012: 24) have in mind are intuitions that 'the good of any one individual is of no more importance, from the point of view . . . of the Universe, than the good of any other' and 'as a rational being I am bound to aim at good generally'.[5] They argue that when we have these intuitions self-evidently—after engaging in a process of rational reflection—these intuitions are not prone to evolutionary debunking. If these Sidgwickian intuitions are self-evident to us, we see (non-derivative) reasons to care about the good of all individuals, including those who are total strangers to us, equally and impartially. But de Lazari-Radek and Singer (2012: 19–20) argue, citing a barrage of evolutionary biologists, that

> *No Direct Stranger Concern*. Evolution would only select us to see direct (non-derivative) reasons to care about ourselves, our relatives, and those to whom we bear a special group or partner relationship for their own sake; evolution would not select us to see reasons to care about strangers for their own sake. Although evolution might select us to see reasons to help and not harm strangers because of the benefits of helping them (e.g. new trading partners) and the possible costs

> of harming them (e.g. their group harming us or our families), evolution would not select us to see reasons to non-instrumentally care about strangers.

So, according to de Lazari-Radek and Singer, if we come to believe in the truth of impartial consequentialism or utilitarianism on the basis of direct self-evident Sidgwickian intuitions, then these beliefs in impartial consequentialism are not debunkable.

1B. Skepticism about the possibility of LDAs

Kahane, Tersman, and Vavova all make a similar argument against LDAs. They *assume the truth of LDAs' debunking claims* but argue that their truth would prove too much. They make two claims. First,

> (1) The LDAs that have been made implausibly over-extend in a way that makes them self-undermining.

Kahane (2011:119-120) and Vavova (2014: 93–95) discuss Singer's (2005) LDA. They argue that if this argument's debunking claim is true, it's vindicatory claim is false because if kin altruism was evolutionarily selected for, then impartial altruism is a reasoned extension of kin altruism. The debunking premise claims that kin altruism is an irrelevant influence. But the reasoned extension of an irrelevant influence is still an irrelevant influence. Tersman (2006: 401) also discusses Singer's (2005) LDA. He argues that there are many possible debunking stories that debunk the impartialist intuitions that figure in Singer's vindicatory premise. For instance, impartial consequentialism is plausibly deeply influenced by Christian ethics, so impartialist intuitions may be the result of our Christian heritage.

Kahane (2014) assumes that de Lazari-Radek and Singer's claim, *No Direct Stranger Concern* is correct but argues that this is not enough for them to establish the kind of vindicatory claim that they want. Kahane (2014: 331–334) argues that if de Lazari-Radek and Singer's (2012) debunking premise is true, then our beliefs that pain is bad and that pleasure is good would be evolutionarily debunked too; for many, such as Street (2005: 150), have argued that being disposed to treat pain as bad and to pursue pleasure helped our ancestors to avoid injury and death. This counts as an over-extension because de Lazari-Radek and Singer's (2012) argument is an argument in favour of utilitarianism. But utilitarianism requires a theory of well-being. And Kahane argues that a form of utilitarianism that did not hold that pain is one of things that we should minimise would be an extremely odd form of utilitarianism that de Lazari-Radek and Singer would not accept. So, if their debunking premise is true, their vindicatory premise is true only in a way that they would not be interested in and we should not care about: the kind of beliefs in utilitarianism that are immune to debunking are beliefs in

a very unattractive form of utilitarianism that they do not hold and would think implausible.

Kahane, Tersman, and Vavova argue that their case for (1) generalizes to show that

(2) LDAs in general are implausible because they over-extend: they extend to undermine the justification of such a large set of ethical beliefs that they entail a form of moral skepticism.

Tersman (2006: 403) says that LDAs in general threaten to collapse into arguments for a more general moral skepticism. He holds this more general claim because there are so many different possible debunking explanations of different moral judgments that we hold. And if non-consequentillaist intuitions are debunked on the basis of the evolutoinary foces that lead to them, then similarly all other moral judgments willl be debunked on the basis of the possibility of debunking explanations of what caused them. Kahane (2011) (2014) argues that (2) is true because if LDA's debunking claims are true, then we are not justified in believing that pain is bad and that pleasure is good. Kahane (2011: 120–121) says that if anything would survive a purge of our ethical and normative beliefs that are infected by irrelevant evolutionary influences it is likely to be more in line with a form of Nietzschean egoism than utilitarianism before concluding that '[i]f [LDAs] work at all then, in one way or another, they are bound to lead to a truly radical upheaval in our evaluative beliefs'. Similarly, Kahane (2014: 334) says,

> Utilitarianism is often viewed as an extremely counterintuitive view; many find Singer's normative views troubling, even repugnant. But if we take the goal of purging all evolutionary influence from our normative views seriously enough, we will end up with a view that is so radically divorced from common sense, and so distant from any familiar ethical theory, that, by comparison, Singer's own utilitarianism will seem almost like old-fashioned common sense.

Vavova (2014: 94–95) similarly claims that LDAs 'just cannot provide an appropriately selective argument that targets all and only the intended beliefs.' She makes this conclusion just on the basis of her argument that the debunking premise of Singer's LDA generates the debunking of consequentialist and utilitarian intuitions too. Presumably, she thinks that showing that Singer's argument has this implication shows that LDAs more generally implausibly overgeneralize because our consequentialist intuitions are our intuitions about reasons to do good for others (*ibid*: 95). And a moral system without such altruistic reasons wouldn't be much of a moral system at all.

In the rest of this paper I will argue that there is not a more general problem here: LDAs can be made that do not generate a form of skepticism; even assuming (1) is true, (2) is false. In making this argument I will assume, as the skeptics about LDAs that I have been discussing do, that the debunking premise of the

LDAs discussed in §1A are true, and that the arguments that proponents of these LDAs make for these premises are sound.

2. How to make a LDA

How we understand the epistemic principle in LDAs has implications for whether we should be skeptics about the possibility of non-self-undermining LDAs or not.

2A. How to Understand LDAs Epistemic Principle

Kahane (2011: 111) and Clarke-Doane (2012: 319) understand the epistemic principle at work in LDAs as:

> *Insensitivity*. If we would have believed that *p* even if not-*p*, then this fact (or our reasonably believing this) undermines or defeats the justification of our belief that *p*.

On this view, consequentialist LDAs argue that we would have had deontological beliefs even if they were false, but it's not the case that we would have had impartial consequentialist beliefs even if they were false.

However, White (2010: 581) as well as Bogardus (2016: 644)—in a paper discussing the use of *Insensitivity* in evolutionary debunking arguments—have argued that *Insensitivity* is implausible. Suppose that we are jury members and we have overwhelming evidence that a defendant is guilty: we have eyewitness testimony, DNA evidence, finger prints, etc. It seems that we are justified in believing that the defendant is guilty. But we would have believed that the defendant is guilty even if they were not. For if a vast conspiracy was in operation, we would have the same seemingly overwhelming (but misleading) evidence. And so *Insensitivity* implausibly entails that we are not justified in believing that the defendant is guilty.

There is an epistemic principle close to *Insensitivity* that avoids this problem, namely

> *Good Reason to Believe Unreliable*. If (i) there is good reason for us to believe that *X* occurred and is causally responsible for it seeming to us that *p*, and (ii) we should believe that *X* would make it seem to us that *p* even if not-*p*, then our justification for believing that *p* on the basis that it seems to us that *p* is undermined or defeated.

Good Reason to Believe Unreliable does not entail that we are not justified in believing that the defendant is guilty in the jury case. This is because in the jury case we have no good reason to believe that a vast conspiracy has taken place to frame the defendant. And *Good Reason to Believe Unreliable* would only entail

something about the status of our belief that the defendant is guilty if we had good reason to believe that a vast conspiracy involving the planting of misleading evidence had taken place.[6]

I have encountered the following objection: we have good reason to believe that (*) the court has behaved just as it would if a conspiracy had taken place. And the fact that (*) fact undermines the justification of our belief that the defendant is guilty.

However, (*) does not undermine our justification for believing them to be guilty; this is just what White and Bogardus take the jury case to show. And so it is a virtue of *Good Reason to Believe Unreliable* that it does not entail that (*) undermines our justification for believing the defendant to be guilty. Furthermore, concerns about evolutionary and other kinds of genealogical debunking are supposed to be distinct from the concerns motivating wholesale epistemological skepticism. But the kind of reasoning in this objection —and *Insensitivity*—motivates wholesale epistemological skepticism. For if we were brains in vats being manipulated by an evil demon to believe that we have bodies, we would believe that we have bodies even though we do not. But our being a brain in a vat is a merely logical possibility, rather than a live possibility that we have good reason to believe might be happening; evolutionary or genealogical debunking arguments are supposed to be stronger than arguments for wholesale epistemological skepticism precisely because we know that evolution happened and happened in various ways. So, although we have no positive good reason to believe that we are brains in vats, we *do have positive* good reason to believe that we have been manipulated by evolution in various ways. And this makes a difference to what we can justifiably believe (if we are not epistemological skeptics!)[7]

Our intuitions about cases provide us with reasons to accept *Good Reason to Believe Unreliable* too. Suppose that Blake signs up for an experiment. She's shown a short video and then given some data on virtual reality therapy for anxiety disorders, which is a topic that she knows nothing about. She finds the data compelling. But then she arrives home to find an email from the experimenters which tells her that the experiment was one attempting to ascertain the effectiveness of visual priming. Participants were divided into two groups. The video that one group were shown used subliminal cues to prime that group towards an interpretation of the data they were later shown; the other group's video had no subliminal cues. The experimenters tell her that they found a striking correlation: all the primed participants believed as they were primed and thought that the data on VR therapy showed that it was effective. However, the experimenters tell her that they cannot disclose to her whether she was in the primed group or not. It seems that in this case Blake's justification for believing that VR therapy is effective is undermined: after receiving the email she is no longer justified in having this belief. She might have been lucky enough to be in the unprimed group and to in fact respond to the data rationally; however, she has some good reason to believe that she was not in that group for half of the subjects were primed to

misread the data and so there is a good chance that she was one of the primed participants.[8] In this case (a) Blake should believe that if *something* happened to her (being primed by experimenters), then she would form a particular belief about some matter (that there is very good evidence of the effectiveness of VR therapy) and that belief would not be sensitive to the truth of that matter: she would believe that there is very good evidence that VR therapy is effective even if there were not. And (b) there is good reason for her to believe that this *something* (being primed by the experimenters) in fact happened to her. And it seems that the justification of the relevant belief of Blake's is undermined in virtue of (a) and (b) Vavova (2018: 143–144). So, cases like this support.

> *Good Reason to Believe Unreliable*. If (i) there is good reason for us to believe that X occurred and is causally responsible for it seeming to us that p, and (ii) we should believe that X would make it seem to us that p even if not-p, then our justification for believing that p on the basis that it seems to us that p is undermined or defeated.

One quick clarification: 'Good reason' in *Good Reason to Believe Unreliable* should be taken to mean some non-trivial or relatively strong reason. I want to leave it open exactly how strong the reason has to be for our justification for believing that p to be undermined, but the reason can't be too weak. For instance, suppose that the investigators' email explained that there was a 1/1000 chance that Blake had been primed. This might be some weak reason for her to believe that she's been primed. But in this scenario, the justification of Blake's belief that VR therapy is effective would not seem to be defeated or (significantly) undermined. (For those seeking a bit more precision, suppose, for instance, that R is a good reason to believe that p only if the probability of p given R is >0.5).[9]

So, it seems that we should understand LDAs as involving *Good Reason to Believe Unreliable* as their epistemic principle—or that they are better constructed using this principle than *Insensitivity*; indeed some skeptics of LDAs, such as Vavova (2014), already argue for a very similar conclusion. (I discuss how Good Reason to Believe Unrealisble can be tweaked to avoid certain further objections in a footnote).[9] But, as we'll see this has implications for how we should understand LDAs as well as for the plausibility of skepticism about LDAs.

2B. Implications of this way of constructing LDAs

For convenient short-hand, let's say that when *Good Reason to Believe Unreliable* holds of some X, p, and belief that p, we have good reason to believe that our belief that p is caused by a non-truth-tracking process with regards to whether p. If *Good Reason to Believe Unreliable* is the epistemic principle at work in LDAs, then we should understand the two empirical premises in LDAs as having the following form:

> *Debunking*. For some set of moral judgments *S1* we have good reason to believe that *S1* are caused by a non-truth tracking process such as evolution. (E.g. we have good reason to believe that our moral intuition that we ought not push the heavy man was caused by a non-truth-tracking process).
>
> *Vindicatory*. For some other set of moral judgments *S2*, we do not have good reason to believe that *S2* are caused by a non-truth tracking process such as evolution. (E.g. we do not have good reason to believe that Sidgwickian self-evident intuitions are caused by a non-truth-tracking process).

With this account of how LDAs can be plausibly constructed in hand we are in a position to see how LDAs can be made that avoid Regina Rini's (2016) argument for skepticism about LDAs. Rini argues that all LDAs fail because LDAs necessarily generate an infinite justificatory regress. She first argues that all LDAs rely on moral judgments. LDAs rely on claims of the form: if particular moral judgment set S were caused by a particular psychological or evolutionary process P, then S are insensitive to the moral truth because P is insensitive to the moral truth. But, according to Rini, if we make LDAs against some moral judgments or intuitions and thereby show that this set of moral judgments or intuitions are insensitive to the moral truth, then we have to consider whether we should generalize 'from this to worry about the reliability of other sets of moral judgments'. Rini (2016: 683, 681–682) says that

> [t]he point here is meant to be intuitive . . . [f]or 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.

This shows, according to Rini, that good LDAs must schematically involve the claims that

(a) moral judgments of set S were caused by process P, which is insensitive to the moral truth; and
(b) the belief that process P is insensitive to the moral truth was caused by process P*, which *is sensitive* to the moral truth.

Rini then argues that in this case LDAs must lead to a vicious infinite justificatory regress. This is because, if in order to be justified in making LDAs involving claims like (a) we need to be justified in making claims like (b), then in order to be justified in making claims like (b) we must be justified in making claims like: (c) the belief that process P* is sensitive to the moral truth was caused by process P**, which is sensitive to the moral truth. And in this case, we will have to be justified in claiming that: (d) the belief that process P** is sensitive to the moral truth was caused by process P***, which is sensitive to the moral truth; and so on *ad infinitum*. So, all LDAs must fail because they lead to a vicious infinite justificatory regress.

However, we should reject Rini's claim that good LDAs must involve (b) as well as (a). Good LDAs only need to make claims of the form of (a) and claims of the form:

(b*)*we have no good reason to believe that* the belief that process *P* is insensitive to the moral truth was caused by a process that is *in*sensitive to the moral truth.

And we are not under any pressure to justify claims of the form of (b*) via some further claim (some analogue of (c)) in the way that we are under epistemic pressure to justify claims of the form of (b) via claims of the form of (c). So, LDAs do not lead to an infinite justificatory regress. This just follows from the account of LDAs I've proposed. I've argued that we should understand LDAs as only involving:

Debunking. For some set of moral judgments *S1 we have good reason to believe that S1* are caused by a non-truth tracking process such as evolution; and

Vindicatory. For some other set of moral judgments *S2*, *we do not have good reason to believe that S2* are caused by a non-truth tracking process such as evolution.

Rini's argument relies on the assumption that in order to make a good debunking argument, debunkers need to show that the moral judgments they wish to preserve were not the product of an irrelevant influence; but this is not the case, they just need to show that we lack good reason to believe this.

The view that LDAs involve (b*)/*Vindicatory* rather than (b) fits with our intuitions about the perception case that Rini uses as an analogy. If we find out that some of our perceptions are caused by a process that we have good reason to believe is not truth-tracking, we should thereby question whether the same is true of all our perceptions. But if we then find that our other perceptions are not (obviously) caused by the same off-track process and we have no good reason to believe that the process that caused them is a non-truth-tracking process, then we no longer have reason to worry about our other perceptions; the fact that we found that some of our perceptions are off-track no longer gives us any reason to worry about whether our other perceptions are off-track or not.[10]

Understanding LDAs in this way also seems to undermine Tersman's argument for

(2) LDAs in general are implausible because they over-extend: they extend to undermine the justification of such a large set of ethical beliefs that they entail a form of moral skepticism.

Tersman (2006: 403) claims that in order to show that one's LDA does not collapse into a GDA 'one must show that there are intuitions for which no

debunking explanation can be given or where the debunking explanations are inferior to non-debunking ones'. But this is not right: in order to show that one's LDA does not collapse one only needs to show that there are intuitions that we lack good reason to accept a debunking explanation of. So, *contra* Tersman, the fact that for any ethical judgment, there are many possible debunking explanations of this judgment does not on its own create problems for LDAs.[11]

A further clarification of the structure of LDAs will yield further implications later in this paper. We should not pick out the intuitions in *Debunking* and *Vindicatory* by their content, but rather by the process and/or reason for having these beliefs. This fits with our intuitions about the VR therapy case discussed in 2A. Suppose that Alice is not in the study—and does not know about the study or about Blake—but she has read all the evidence about VR therapy and has come to believe that it is effective. The fact that the epistemic status of Blake's belief that VR therapy is effective is undermined—due to being in the study—does not establish that the epistemic status of Alice's belief that VR therapy is effective is undermined. This is because the process leading to Alice's judgment is distinct from the process leading to Blake's judgment. Furthermore, the way we pick out the judgments or intuitions in *Debunking* and *Vindicatory* should be sensitive to the reasons for which these judgments are held (if there are such reasons: some such judgments may be self-evident intuitions). Suppose that Alex believes that there are true moral claims because she believes that it is wrong to push the heavy man and if that is right, there are true moral claims. But Singer's debunking story is correct, so that her belief about the trolley case is undermined. In this case, her belief that there are true moral claims is undermined too. But suppose that Beth believes that there are true moral claims because she believes that it is analytic that murder is wrong and so she judges that the conditional claim, if ϕ-ing is a murder, then ϕ-ing is wrong, is a true moral claim. None of the premises on the basis of which Beth believes that there are true moral claims is debunkable. (The view that we should not pick out the sets of intuitions in the debunking and vindicatory claims via their content also fits with de Lazari-Radek and Singer (2012) and Kahane's (2014) discussions of LDAs; see §1).

All this establishes that we should understand LDAs as taking the following schematic form:

> *Epistemic Premise*. Suppose that: (i) there is good reason for us to believe that *X* occurred and caused us to have judgment *J* that *p;* and (ii) we should believe that if *X* occurred and caused us to have judgment *J* that *p*, then we would have *J* even if not-*p*. If (i) and (ii) hold, then our justification for believing that *p* on the basis that we have *J* is undermined or defeated.
>
> *Debunking Premise*. For some set of moral judgments *S1* that *p* (e.g. pain is bad), we have good reason to believe that (a) (evolutionary process) *X* caused us to have *S1*, and that (b) if *X* caused us to have *S1*, we would have *S1* even if not-*p*.
>
> *Vindicatory Premise*. For some set of moral judgments *S2* that *q* (e.g. it is better that more lives are saved), there is no *X* such that *we have good reason to believe*

that (a) (evolutionary process) *X* caused us to have *S2*, and (b) if *X* caused us to have *S2*, we would have *S2* even if not-*q*.

Conclusion. Moral beliefs that *p* based on a a judgment in set *S1* are not justified, but moral beliefs that *q* based on a judgment in *S2* are, or may still be, justified; irrelevant evolutionary influences undermine the epistemic status of moral beliefs that *p* based on a judgment in set *S1* but do not undermine the epistemic status of moral beliefs that *q* based on a judgment in set *S2*.[12]

In the next two sections I'll show, *contra* LDA skeptics, that we can hold that the debunking premise is true for some sets of moral judgments and beliefs whilst simultaneously holding that the vindicatory premise is true for another set of moral judgments and beliefs.

3. Internalist LDAs

Many accept an internalist constraint on reasons according to which, a consideration is a reason for an agent to perform an action only if that action promotes some desire or aim that they have.[13] For instance, if classical music does nothing for you, and going to a particular classical concert wouldn't do anything else for you (your friends aren't going, you wouldn't fulfil a promise that you want to keep, etc.), there's no reason for you to go to the concert. There is a first-order normative view held by the majority of those who accept an internalist constraint. According to this view, (i) this internalist constraint holds, and (ii) we have reasons to promote at least most of the desires and aims that we have; hereafter I'll refer to this first-order view as internalism.[14]

My contention in this section is that at least many beliefs in internalism are not vulnerable to evolutionary debunking because these beliefs are based on judgments that we *lack good reason to believe to be caused by non-moral-truth-tracking evolutionary processes*—importantly, this is not to say that we have good reason to believe that these judgments *are the result of a moral-truth-tracking process* (see 2B above). To make good on this contention I'll first explain the judgments and intuitions on the basis of which internalists normally claim to accept internalism. (In explaining this, however, I'm not claiming that their reasons for accepting internalism are good reasons—LDAs try to undermine the justification of one set of moral beliefs without undermining the justification of another set, they do not simultaneously try to show that this other set is positively justified).

3A. The Internalist Judgments that LDAs do not undermine

Internalists commonly cite one of the following four judgments (which may be constituted by intuitions or other reasons) for their acceptance of internalism.

First, some internalists, such as Williams (1995b: 189, 194), hold that claims about an agent's reasons for action must say something distinctively about that agent and how that action and its normative status link up to that agent in particular. Since, otherwise reasons claims wouldn't be distinctive claims: to claim that an agent has a *reason* to ϕ would be to say nothing more than that it would be good if they ϕd. If we accept this claim, then, these internalists claim, we must accept internalism.

Second, some, most clearly Manne (2014: 91), argue that it is only when we take-up an interpersonal or second-personal stance towards another that we can be said to reason *with* them 'as opposed to ordering them about, coercing them, or trying to "manage" their behaviour'. And a consideration is a reason for an action only if it is apt to offer it (or ideally would be apt to offer it) to another person when we are 'reasoning with her, or (similarly) offering her collaborative advice or friendly suggestions about what she ought to do'. But Manne (*ibid.*: 103) says that intuitively to her, if an agent A genuinely has no motivation that would be served by their φ-ing, and we come to know this, we cannot aptly claim that there is a reason for A to φ in a rational interpersonal conversation with them about (or when advising them about) whether they ought to φ. And if all this is right, then internalism holds.[15]

Third, many internalists claim that we can preserve the relationship between normative and motivating reasons only if we accept internalism. According to these internalists, if R is a normative reason for A to φ, R must be a counter-factual version of A's motivating reason to φ; R must be the reason for which a counter-factual version of A φs. And only internalism can secure this relationship between normative and motivating reasons.[16]

Fourth, Markovits (2014: 58–65) and Goldman (2009: 68–73) argue that we should accept internalism because of the relationship between practical and theoretical reasons. Reasons for us to believe propositions depend on what we already believe and the standards of procedural rationality; we have reason to believe that p only if some consideration constitutes good evidence for p given what we already believe. So, reasons for belief are constrained by our internal belief-related states. But in this case, we should analogously hold that our reasons for action depend on our aims or desires. That is, that reasons for action are analogously constrained by our internal action-related states. For practical and epistemic reasons stand in the same warranting (reason) relation but just warrant different things (beliefs and actions respectively).[17]

These four judgments (intuitions and/or acceptances of considerations or arguments) lead internalists to accept internalism. We lack good reason to believe that the acceptance of these judgments by internalists is the result of an off-track process such as a non-moral-truth-tracking evolutionary process. Regarding the first consideration, we do not have good reason to believe that non-moral-truth tracking evolutionary forces favoured our judging that claims about an agent's reasons for action must say something distinctive about that agent. For instance, our having these views would not make us and our family and friends more

likely to survive. And similarly, we do not have good reason to believe that a non-moral-truth-tracking evolutionary process would favour our believing that reasons must be the kinds of things that it is apt to cite to another in reasoning with and advising others, or that normative and motivating reasons and practical and epistemic reasons must be the same kinds of things and related to our internal mental states in the same or analogous ways to one another.[18] Call this set of four judgments on the basis of which internalists accept internalism, *internalist judgments*. As I've explained, there seems to be a good case that

> *Internalist Vindicatory Premise*. There is no *X* such that we have good reason to believe that (a) evolutionary process *X* caused internalists to have internalist judgments, and (b) if *X* caused internalists to have internalist judgments, internalist would have them even if internalism were false.

I have encountered the following objection to my argument for this claim: (A) a belief in internalism makes one more effective at persuading others to adopt one's normative preferences; (B) being more effective at persuading others to adopt one's normative preferences is fitness-conducive. (B) holds because effective persuaders have a better chance of encouraging social norms that favour the persuader or their offspring. (A) holds because according to internalism, reasons for agents to perform actions are always tied to their motives. And appealing to a person's existing motives is a much more effective means of persuading someone than demanding that they do things that are disconnected from their motives. So, *Internalist Vindicatory Premise* is false.

There are two problems with this argument. First, *we lack good reason to believe (A)*. Perhaps (A) does hold but in order for us to have good reason to believe it we would need correlational data that shows that those who are disposed to accept internalism are more effective persuaders than those who do not accept it. But internalism isn't overwhelmingly dominant in moral philosophy; many philosophers reject it including most if not all impartial consequentialists—for whom reasons are just tied to impartial value—and others such as Scanlon (1998: appendix) and Brunero (2017). We might think that if internalists were better at persuasion than non-internalists, we should expect internalism to be dominant in the field. Furthermore, non-philosophers tend to be disposed to reject internalism because of its implications such as that serial killers have no reason to refrain from killing someone if so refraining would not serve one of their desires or aims.[19] Finally, it's not obvious why a belief in internalism would make one better (or *pro tanto* good) at persuasion. Since non-internalists agree with internalists that when you want to persuade someone to do something it is better to cite a consideration that is connected to their motives because our actions are the products of our motives.

The second problem with this argument is that, as I explained in §2B, LDAs try to undermine/vindicate the justification of beliefs based on particular judgments. To undermine the epistemic status of a particular belief with content *C* based on judgment *J* it is not enough to undermine the epistemic status of

some beliefs with that same content; one needs to undermine the epistemic status of a belief with content *C* based on *J*. But in this case to debunk internalist beliefs based on the four judgments that I've been discussing this objection would need to debunk these judgments. But it cannot do this: even if believing in internalism is fitness enhancing because such a belief makes one more adept at persuading others, believing that claims about reasons and claims about value should be understood to be distinct or that practical and theoretical reasons should be understood to have a similar structure does not make one more adept at persuading others.

It might seem that my argument does not show that internalists' positive first-order normative judgment that *there are* normative reasons for everyone to promote their desires cannot be debunked. However, we lack good reason to believe that evolution would have selected for beings who hold that everyone has normative reasons to promote all of their desires. There are two reasons for this. First, *desires are essentially dispositional states* such that to the extent that we desire to have or do *X*, we are to the same extent disposed to have or do *X* (at least if we can).[20] In this case, judging that we have reasons to satisfy our desires could not further dispose us to do that which promotes our desires. Second, *desires are too contingently linked up to our own and our family's survival*. We have desires for all kinds of things. Many people's desires for pleasure, power, or fame exceed their desires for what's good for their friends and family and what would enable their friends and family to survive. In this case, evolution's selection of beings for reproductive success would not involve the selection of beings who see reasons to promote these desires. Many people desire above everything else that they complete particular projects (works of art for instance), the completion of which do not benefit—and in fact harm—their friends and family (as is made clear by Williams's (1981: 22–23) Gauguin who neglects his family because he finds it more important to become a painter). And some people desire their own pain or pain for their own family and friends. In this case, belief that there are reasons to promote all our desires is too contingently tied up with that which promotes reproductive success to be a judgment that evolution would dispose us to have. Having this belief would too often lead us to do things that are harmful to ourselves and our kin and would frustrate reproductive fitness. So, we at least do not seem to have good reason to believe that the judgment that we have reasons to promote our desires would be fitness-enhancing (and may have some positive good reason to believe that having this judgment would not be fitness-enhancing).

Perhaps evolutionary biologists or philosophers of biology will show that the judgment that we have reasons to promote all our own desires whatever they are is fitness-enhancing. I have not been able to find any considerations that favour this view in the relevant literature. Nor do the arguments that LDA skeptics make show that the judgment that we have reasons to promote all of our desires would be selected for. For instance, Kahane (2011) (2014) argues that beings who didn't think that pain is bad and enjoyment and the absence of pain is good

wouldn't last very long because they would be going around injuring themselves. But beings that didn't think they have reasons to promote all of their desires wouldn't be doing this; for so long as they desire certain things they would be just as motivated to do them anyway.

3B. Internalists can make LDAs without over-extension

With the internalist vindicatory claim in hand, internalists can make a LDA. They can hold the debunking premise of Singer (2005) or de Lazari-Radek and Singer's (2012) arguments, according to which we have good reason to believe that our deontological judgments or judgments about partial reasons are the result of an off-track evolutionary process. Either of these debunking claims combined with the internalist vindicatory premise and the epistemic principle discussed in §2 will yield the conclusion that: we can be justified in believing internalism but we cannot be justified in believing that there are deontological or partial reasons that outstrip such internal reasons. Evolutionary irrelevant influences do not undermine the justification of beliefs in internalism but do undermine the justification of beliefs in partial and deontological reasons that outstrip internal reasons. In this case, Tersman, Kahane, and Vavova are mistaken that LDAs must over-extend and undermine themselves.

It might be objected that if (a) we lack good reason to believe that internalist judgments are the product of an off-track evolutionary process, then (b) we lack good reason to believe that our judgments about deontological and partial reasons are the result of an off-track evolutionary process. Assessing whether that is the case is beyond the scope of this paper. I am merely arguing that if we accept skeptics about LDAs' assumptions, which include the negation of (b), skepticism about LDAs does not follow.

I have encountered two responses to my argument that internalists can make LDAs without self-defeat.[21] First, the LDAs that I've been discussing do no interesting work for internalists because internalists' already believe that there are no non-derivative deontological or partial reasons for agents to perform actions that outstrip those reasons they have to promote their own desires and goals.

However, in order for these LDAs to do interesting work for internalists it only needs to be the case that internalists can wield LDAs to show that others, their opponents—as well as those who have no belief (yet) about whether internalism holds or who are on the fence about this—are not justified in believing that there are non-derivative external deontological and/or partial reasons. And, as I've shown, internalists can wield these arguments for this purpose.

Second, internalists who make evolutionary debunking arguments are just making GDAs because internalism is a metaethical view, which involves a reductive analysis of reasons, rather than a first-order view in normative ethics. And

in this case I have not shown that LDAs can be made without over-extension or self-defeat.

However, the internalist view that I discussed involved no such reductive analysis, and many, such as Korsgaard (1986) and Scanlon (2014: 5), understand internalist views like the one that I've discussed to involve a substantive first-order view of the practical reasons there are rather than a metaethical view, namely, that the only normative reasons there are are reasons that link up to our desires. Furthermore, given that, as I've argued, internalists can wield LDAs against specific types of (non-internalistically constrained) reasons without self-defeat, this fact on its own shows that internalists can make LDAs regardless of whether internalism involves or entails a metaethical view. The important feature here is whether the arguments that are being made are local or not, that is, whether the arguments that are being made aim to undermine the justification of all ethical judgments (or all non-analytic ethical judgments construed objectively or realistically) or just some sub-set of ethical judgments; and, as I've been arguing, internalists can make such a local debunking argument.

As I discussed in §1B, Kahane (2011: 103, 120–121) argues that successful LDAs will entail that the only ethical views that we can be justified in believing are views that constitute 'a truly radical revision of our evaluative outlook'. So, do (all) internalist views constitute radical revisions of our evaluative outlook? If they do, internalist LDAs will over-extend in the way that Kahane claims that all LDAs over-extend.

Most of us have aims that would be served by acting morally: most of us care about the well-being of others, as well as caring about avoiding blame and punishment, and not looking bad in front of others. Internalism might entail that some extremely unusual people have no reasons to refrain from doing things that we believe to be wrong but it's not obvious that this would constitute a truly radical revision of our evaluative outlook rather than just a sad truth about how some human beings are constituted (Manne 2014). Furthermore, many including Korsgaard (1996), Markovits (2014: ch. 4–6), Schroeder (2007: ch. 7), and Finlay (2008) have argued that not only is internalism not radically revisionary but that the counter-examples to it fail; so the view that internalism is radically revisionary is at least *controversial*. And it seems that Kahane believes that the only successful LDAs will be *uncontroversially* radically revisionary. For Kahane (2011: 120) seems to believe that successful LDAs would entail that we are only justified in accepting a radically revisionary evaluative outlook because they would entail that we are only justified in accepting something like a Nietzschean perfectionist anti-moralistic outlook which is both intended to be radically revisionary and is uncontroversially radically revisionary.[22]

4. Externalist LDAs

Externalists about reasons do not accept internalists' subjective constraints on reasons. There is at least one way in which externalists can also make LDAs that avoid LDA skeptics' charge of over-extension and self-undermining. This way of making a LDA involves a controversial empirical claim. Namely, the controversial empirical claim of de Lazari-Radek and Singer's that I discussed in §1A:

> *No Direct Stranger Concern*. Evolution would only select us to see direct (non-derivative) reasons to care about ourselves, our relatives, and those to whom we bear a special group or partner relationship for their own sake; evolution would not select us to see reasons to care about strangers for their own sake. Although evolution might select us to see reasons to help and not harm strangers because of the benefits of helping them (new trading partners) and the possible costs of harming them (their group harming us or our families), evolution would not select us to see reasons to non-instrumentally care about strangers.

De Lazari-Radek and Singer (2012: 19–20) say the following in defence of this claim:

> Richard Dawkins has argued—as the title of his early work, *The Selfish Gene*, suggests—that actions that involve sacrificing an organism's prospects of surviving and reproducing have evolved because they benefit the organism's genes, largely through favoring kin. He does not hesitate to draw the conclusion that "much as we might wish to believe otherwise, universal love and the welfare of the species as a whole are concepts that simply do not make evolutionary sense." Pierre van den Berghe has said flatly, and no doubt too bluntly, that "we are programmed to care only about ourselves and our relatives." Richard Alexander, in *The Biology of Moral Systems*, writes: "I suspect that nearly all humans believe it is a normal part of the functioning of every human individual now and then to assist someone else in the realization of that person's own interests to the actual net expense of those of the altruist. What this greatest intellectual revolution of the century [i.e., the individualistic perspective in evolutionary biology] tells us is that, despite our intuitions, there is not a shred of evidence to support this view of beneficence, and a great deal of convincing theory suggests that any such view will eventually be judged false."
>
> In *Unto Others*, Elliot Sober and David Sloan Wilson have forcefully challenged this individualistic perspective in evolutionary theory. They argue that evolution could have selected for actions that benefit groups to which individuals belong, rather than for actions that benefit the individuals themselves. For the argument we are about to make, therefore, it is vital to understand that, while Sober and Wilson are challenging the views of Dawkins, van den Berghe, and Alexander, they do not argue that evolution could have selected for the kind of universal benevolence required by Sidgwick's axiom. As they put it, "our goal in this book is not to paint a rosy picture of universal benevolence. Group selection does provide a setting in which helping behavior directed at members of one's

> own group can evolve; however it equally provides a context in which hurting individuals in other groups can be selectively advantageous. *Group selection favors within-group niceness and between-group nastiness*."[23]

I don't know whether *No Direct Stranger Concern* is true. But the skeptics of LDAs that I've been discussing generally claim to not be interested in contesting this or similar claims. For instance, Vavova (2014: 79), noting that not even proponents of debunking arguments think that they have conclusive grounds for their empirical claims, asks '[s]o why take [these arguments] seriously? Because the philosophically interesting question is not whether some empirical claim is true, but what follows about the rationality of our beliefs if something like it were true'. Tersman (2006: 400–401) appears to grant Singer a claim along the lines of *No Direct Stranger Concern*.[24] And in his response to de Lazari-Radek and Singer's LDA, Kahane (2014: 329) says that 'although [he has] some serious reservations about the way that de Lazari-Radek and Singer defend [their] empirical claims' his argument that their LDA and all LDAs are self-undermining 'will not directly challenge them or rely on any similarly controversial empirical speculation . . . Instead, I will argue that even if we grant these claims, the authors fail to address the worry that utilitarian appeals to evolutionary debunking are ultimately self-undermining'. (It's also worth noting that Kahane notes what is presumably his most serious reservation about the way de Lazari-Radek and Singer defend their empirical claims in a footnote and it has nothing to do with *No Direct Stranger Concern*).[25] So, for the purpose of assessing Kahane, Tersman, and Vavova's arguments for skepticism about LDAs we can grant *No Direct Stranger Concern*.

Consider the following propositions,

> *Non-Violence to Strangers*. There is a strong non-derivative reason for us to refrain from killing or assaulting innocent strangers;

> *Save the Lives of Strangers*. There is a strong non-derivative reason for us to save the lives of innocent strangers if we can easily do so.

Non-Violence to Strangers and *Save the Lives of Strangers* are intuitively self-evident to many. But, assuming *No Direct Stranger Concern*, beliefs based on these intuitions, if they are direct and self-evident, are not vulnerable to debunking. This is because these intuitions are about how we ought to act regarding strangers rather than those with whom we have a kin, community, or other group relationship. And it follows from *No Direct Stranger Concern* that evolution would not select us to have direct self-evident intuitions as to the truth of these propositions. (Note that it is consistent with the view that evolution would not select us to see *Non-Violence to Strangers* and *Save the Lives of Strangers* as directly intuitively self-evident that evolution would selects us to see *derivative instrumental*, but not *non-derivative non-instrumental*, reasons to help strangers).

An obvious objection here is that our intuitions that *Non-Violence to Strangers* and *Save the Lives of Strangers* hold are not self-evident; rather these

intuitions are derived from direct intuitions of the truth of the more general claims that

> *Reasons of Non-Violence*. There is a strong non-derivative reason for us to refrain from killing or assaulting others; and
>
> *Reasons to Save Lives*. There is a strong non-derivative reason for us to save the lives of others if we can do so easily.

However, first, it is not obvious that we do derive *Non-Violence to Strangers* and *Save the Lives of Strangers* from *Reasons of Non-Violence* and *Reasons to Save Lives*. Suppose that we see someone that we don't know. It seems directly intuitive to us that we have a reason not to assault them and not just because of the bad additional consequences for us or others that this might lead to. The intuition that we have that there is a reason for us not to assault this person that we don't know is not derived from a more general intuition that there is a reason for us not to assault others.[26] So, at least some of the intuitions that we have that support *Non-Violence to Strangers* are not derived from the intuition that *Reasons of Non-Violence* holds; more generally, many have claimed that our particular moral judgments are not all derived from judgments about more general principles.[27] For instance, in the combination of the switch and footbridge trolley cases we judge that one action is wrong and the other is not but given that there is a dearth of plausible non-Kammian principles distinguishing the two cases, it does not seem that our judgment is the result of applying a principle to the cases.

Second, assuming *No Direct Stranger Concern*, even if our intuitions about *Non-Violence to Strangers* and *Save the Lives of Strangers* are derived from our intuitions about *Reasons of Non-Violence* and *Reasons to Save Lives*, there is no good evolutionary debunking explanation of our having direct self-evident intuitions that *Reasons of Non-Violence* and *Reasons to Save Lives* are true. For *Reasons of Non-Violence* and *Reasons to Save Lives* make no reference to ourselves, our relatives, and those to whom we bear a special group or partner relationship. And so *No Direct Stranger Concern* implies that evolution would not select us to have direct self-evident intuitions of their correctness. (The obvious objection here is that evolution just isn't that well targeted: it might select us to judge that there are non-derivative reasons not to assault all as a way of getting us to not assault our family. This may well be true, but if *No Direct Stranger Concern* is true, then this could not be the case. So, the objection here is to *No Direct Stranger Concern*, which, as I've discussed, LDA skeptics do not challenge).

Now, there are various plausible weak impartialist claims our belief in which do not seem vulnerable to debunking such as the claim that if we have non-derivative reasons to aid and refrain from killing some agents, then we have non-derivative reasons to aid and refrain from killing all agents. Or the claim that if we have non-derivative reasons to treat some agent in some way, then we have

reasons to treat all agents in that way.[28] These claims seem self-evident to many. And, assuming *No Direct Stranger Concern*, beliefs in these impartialist claims based on direct self-evident intuitions about these claims are not vulnerable to debunking. And the combination of such impartialist claims and *Non-Violence to Strangers* and *Save the Lives of Strangers* entails that *Reasons of Non-Violence* and *Reasons to Save Lives* hold.

Assuming *No Direct Stranger Concern*, there is another set of (external) reasons that we can believe in without our beliefs in these reasons being susceptible to debunking. Manne (2017: 9, 5–8) draws attention to what she thinks of as a subset of desires, bodily imperatives. Bodily imperatives are the mental states that we are in when we are in states of agony or hunger, are thirsty, or are freezing; they 'are the sorts of states which torturers are able to use against their victims. For when the body is protesting, people can be broken'. It seems that we have non-derivative reasons to stop people from being in these states. But if these states are just fundamentally desires—as Manne argues—then we have non-derivative reasons to promote the desires of others. Furthermore, it seems to me *directly intuitive* that we have non-derivative reasons to stop strangers from being in these states of agony, hunger, or great discomfort; when I come to judge that there are such reasons, I'm not first judging that there are reasons to stop myself and my friends and family from being in these states and then generalizing to the view that there must be reasons for me to stop anyone from being in these states. But, in this case, if we assume *No Direct Stranger Concern*, our intuition that there are such non-derivative reasons to promote others' desires are not susceptible to evolutionary debunking. In this case, if we generalize from such a judgment to the view that we have non-derivative reasons to promote others' desires—via a non-debunkable impartialist claim—this generalized belief will not be susceptible to evolutionary debunking.[29]

So we have a set of judgments: that we have (external) reasons of non-violence, to save others' lives, and to promote others' desires. This set of judgments are judgments that are based on impartialist generalizations from intuitions that we have reasons to promote strangers' desires, save their lives, and not be violent towards them. Call this set of judgments *externalist impartial judgments*. And call the view that there are such generalized externalist reasons of non-violence, to promote others' desires, and save their lives, *external impartial reasons*. If we assume *No Direct Stranger Concern*, it seems to follow that we can argue:

> *Externalist Vindicatory Premise*. There is no *X* such that we have good reason to believe that (a) evolutionary process *X* caused us to have externalist impartial judgments, and (b) if *X* caused us to have externalist impartial judgments, we would have them even if there were not external impartial reasons.

I do not have space to fully address objections to my argument that if *No Direct Stranger Concern* holds, then we should accept *Externalist Vindicatory Premise*. But analogues of the moves that I made to defend the internalist vindicatory

premise in §3 can be made to defend this claim. And with *Externalist Vindicatory Premise* in hand externalists about reasons can adopt the debunking premise of, for instance, de Lazari-Radek and Singer's (2012) argument, according to which we have good reason to believe that our judgments about partial reasons are the result of an off-track evolutionary process. Some externalists can then justifiably hold that they and others can be justified in their beliefs in externalist impartial reasons but that we cannot be justified in our beliefs in partial reasons. So, assuming *No Direct Stranger Concern,* externalists about practical reasons can also make LDAs without these LDAs being self-undermining.

5. The Role of LDAs in Normative Ethics

In this paper, I've shown that Kahane, Rini, Tersman, and Vavova's skepticism about the possibility of LDAs is misplaced. LDAs do not necessarily over-extend or undermine themselves. Non-self-undermining LDAs can be made by internalists about reasons. And given an empirical assumption that skeptics about LDAs themselves grant, externalists can too. (At least assuming with LDA skeptics that the debunking premises of Singer (2005) and de Lazari-Radek and Singer's (2012) LDA are correct). But what do these conclusions mean for the role of LDAs in normative ethics?

I can only very briefly sketch the implications of my arguments here—and my discussion of these implications will assume that the debunking premises of Singer (2005) and de Lazari-Radek and Singer's (2012) arguments hold. It seems that one implication is that, at least as things stand, *contra* de Lazari-Radek and Singer (2012), a LDA cannot be made that undermines the justification of all beliefs in all forms of egoism. Since, as I argued in §3, internalist views are not vulnerable to evolutionary debunking; and (certain) internalist views may seem to be akin to a form of egoism.[30] And *contra* Singer (2005), it seems that a LDA cannot be made that undermines the justification of all deontological beliefs. Since, in §4 I argued that, given the empirical claims made by Singer (2005) and de Lazari-Radek and Singer (2012), we can hold beliefs about reasons to refrain from harming or killing others, which do not seem to be consequentialist, that are not vulnerable to evolutionary debunking. So, LDAs are possible but the most well-known LDAs do seem to fail to establish what they try to establish.

However, this does not show that LDAs cannot do significant piecemeal work in normative ethics. For LDAs can undermine the justification of our beliefs in particular sets of non-derivative reasons. Although LDAs may not be able to undermine the justification of *all beliefs in egoism or deontology*, they can undermine the justification of all beliefs in external non-derivative partial and (external) self-interested or prudential reasons (§1, §2). And many consequentialists and deontologists, such as Crisp (2006), Portmore (2011: ch. 1) and Keller (2013), have held that there are non-derivative self-interested and/or partial reasons. Furthermore, if we should reject de Lazari-Radek and Singer's

(2012) controversial empirical claim, *No Direct Stranger Concern*, then it might be that LDAs can be made that undermine all beliefs in external reasons.

I believe that LDAs can be plausibly developed that undermine the justification of beliefs in further different kinds of non-derivative normative reasons. I'll sketch one example of such a possibility. Consider what we can call non-derivative backward-looking reasons, that is, non-derivative reasons for us to keep promises, punish others, and have hostile reactions towards wrongdoers; these reasons are backward-looking because they are reasons to do things just solely in virtue of actions in the past. The case that our beliefs in such non-derivative backward-looking reasons can be evolutionarily undermined has been made by philosophers including Joyce (2006: 24–26) and Olson (2014: 142). According to this case, evolution would favour our having the intuition that there are non-derivative (backward-looking) reasons for us to keep promises regardless of whether there were such reasons. For it would be evolutionarily beneficial for us to be disposed to judge that it is right in itself to keep promises and wrong in itself to break them and that there are reasons for us to keep these promises for their own sake. For individuals and families that are in communities that can make agreements with other individuals and communities to divide tasks, labour, and shares of goods are more likely to survive than individuals that are in communities that cannot make such agreements. Just as evolution favours kin altruism, evolution favours the acceptance of norms the acceptance of which by individuals disposes those individuals to keep their agreements. And accepting the view that it is wrong in itself to break a promise makes individuals more likely to keep their promises.[31]

Similarly, evolution would seem to favour the acceptance of norms of punishing and having hostile reactions (for its own sake) to those who breach agreements and who breach other norms regardless of whether such norms tracked the objective normative truth because communities in which people will punish (for its own sake) those who break agreements are communities in which people are less likely to break agreements.[32] So, intuitions that we ought to punish and/or blame those who breach agreements for its own sake will be favoured by evolution regardless of whether such intuitions track the normative truth.[33]

So, we have good reason to believe that evolution would favour our having the intuition(s) that we have non-derivative backward-looking reasons even if there were no such reasons. And in this case, given the epistemic principle that I developed in §2, *Good Reason to Believe Unreliable*, it seems to follow that our justification for believing that there are non-derivative backward-looking reasons (such as reasons to keep promises and punish for its own sake) is undermined. This case seems as strong as Singer's (2005) and de Lazari-Radek and Singer's (2012) case—which LDA skeptics grant—for the view that evolutionary influences undermine the justification of our beliefs in deontological, and self-interested and partial reasons.

The conclusion that our beliefs in non-derivative backward looking reasons are not justified would have significant implications for normative ethics. For

many significant normative ethical theories involve the claim that there are non-derivative backward-looking reasons to punish wrongdoers (for its own sake) and/or to keep promises (for its own sake). For instance, W.D. Ross's (1930) ethical theory involves such non-derivative reasons or duties.[34] And Darwall's (2006) influential second-personal moral view involves the claim that there are such reasons to keep promises for its own sake.

So, LDAs are possible and can do significant work in normative ethics, just not exactly the work that some have thought that they can do: instead of showing that we should accept one big theory (utilitarianism) over another (egoism, deontology) they can show that we cannot be justified in holding that there are certain types of (non-derivative) reasons though we *can be* justified in holding that there are other types of (non-derivative) reasons.

One final issue. It might seem that the conclusion that LDAs cannot show that we should accept one big theory over another is not right. For, as I discussed in §1B, Kahane (2014) argues that we cannot be justified in holding any plausible account of well-being on the basis of intuitions about what things are good or bad because intuitions such as that pain is bad for us and that pleasure is good for us are debunkable. And without a plausible account of well-being, utilitarianism is implausible. So, LDAs show that we cannot be justified in accepting a big theory, namely utilitarianism.

However, I do not believe that things are *quite* so simple, for an unfortunately complicated reason. According to the buck-passing account of value (BPA) for something to be good is for there to be reasons for us to have pro-attitudes towards it.[35] In Rowland (2019: 154–158) I argued that the BPA is compatible with internalism and that the combination of BPA and internalism does not yield an implausible account of value. So, even if the only beliefs in reasons that are not debunkable are beliefs in internalist reasons, we might be able to construct a plausible account of value out of these internal reasons. On this account of value it will be good for everyone to get what they want. We could then combine such an account of value with the content of the Sidgwickian intuitions that 'the good of any one individual is of no more importance, from the point of view . . . of the Universe, than the good of any other' and 'as a rational being I am bound to aim at good generally'; these are the intuitions that de Lazari-Radek and Singer (2012: 24) claim not to be debunkable.[36] So, if we have internalist intuitions as well as Sidgwickian intuitions, then, somewhat paradoxically we may be able to have non-debunkable beliefs in a form of utilitarianism.

There is still a lot of work to be done here. The implications of my argument are most generally that the role of LDAs in normative ethics is complicated and that LDAs can be made but with some difficulty. But *contra* LDA skeptics, it is possible for LDAs to play some significant role in normative ethics.[37]

Notes

1. See Joyce (2006) and Street (2006). The 'global' and 'local' debunking terminology is due to Kahane (2011).
2. See Crisp (2006: 121-122).
3. See Vavova (2018).
4. Stratton-Lake (2016: §1.1).
5. Sidgwick (1907: 381–82).
6. Vavova (2018:142-147).
7. Vavova (2014: 80-87).
8. This is Vavova's (2018: 143-144) case.
9. Good *Reason to Believe Unreliable* may need a little tweaking. As it stands it entails that if we acquire good scientific evidence that *p*, which might turn out to be misleading evidence that *p*, we are not justified in believing that *p* on the basis of this evidence; since in such a case we may have good reason to believe that our acquiring good evidence that *p* would make us believe that *p* even if not-*p*. This objection may show that it needs to be specified in *Good Reason to Believe Unreliable* that *X* does not bear on whether *p* and is irrelevant to whether that *p*. It is not circular or ad-hoc to add this stipulation. For it is uncontroversial in the VR therapy case that being primed by an experimenter does not bear on whether VR therapy is effective.

   However, it seems that we must specify that we have reason to believe that *X* does not bear on whether *p* and is irrelevant to whether *p* that is independent as to whether we accept *p* or not; see Vavova (2014: 81-82). For instance, suppose that you are an anti-vaxxer who believes that medical schools tout myths about the pedigree of vaccines and you know that your friend Alex only believes that vaccines are effective because they went to medical school. You cannot plausibly give Alex reason to believe that she is not justified in believing that vaccines are effective just by showing her that she only believes in the effectiveness of these vaccines because she went to medical school; this is because your reason for believing that medical school is an irrelevant influence on beliefs about the effectiveness of vaccines is not independent of your belief that vaccines are ineffective which she does not share. In contrast, in the VR therapy case we have independent reason to believe that Blake's judgment that VR therapy is effective is caused by a process that is irrelevant to whether VR therapy is effective; for our belief that being primed is irrelevant to whether VR therapy is effective does not depend on any beliefs about whether VR therapy is effective. (The same seems true for the view that evolutionary influence is an irrelevant influence. For many's belief that evolutionary influence is irrelevant influence is a belief that it independent of whether they believe that utilitarianism or deontology is true, for instance; see *infra* note 10).

   So, it might be that we should revise *Good Reason to Believe Unreliable* in the following way:

   *Good Reason to Believe Unreliable**. If (i) there is good reason for us to believe that *X* occurred and is causally responsible for it seeming to us that *p*, (ii) we should believe that *X* would make it seem to us that *p* even if not-*p*, *and (iii) we have good reason to believe that whether X occurred is irrelevant to whether*

p that is independent of our beliefs about whether p, then our justification for believing that *p* on the basis that it seems to us that *p* is undermined or defeated.

The additional third clause does not matter too much for the core purpose of this paper, for it only comes into play when (i) and (ii) are satisfied; and, as I'll argue in §3-4, they are not for certain internalist and externalist judgments. However, there are certain points at which holding *Good Reason to Believe Unreliable** rather than the unrevised version of this principle does matter, see *infra* notes 10 and 22.

10. If we understand the epistemic principle in LDAs as based on *Good Reason to Believe Unreliable**—*supra* note 9—then LDAs would seem to need positive reason to hold that evolutionary forces regarding our judgments about the moral status of ϕ-ing are irrelevant to the truth about the moral status of ϕ-ing that is independent of our beliefs about the moral status of ϕ-ing. And we may not be able to find such independent reason. For some philosophers hold that the moral status of ϕ-ing is determined by what evolution would select us to believe about the moral status of ϕ-ing. However, this would only show that LDAs have a slightly more limited target, namely the beliefs of those who assume that the moral status of ϕ-ing is not determined by what evolution would select us to believe. (Or that these LDAs work conditional on the assumption that evolution does not determine objective morality). Many egoists, consequentialists, deontologists, and believers in partial reasons hold that what we morally ought to do outstrips what would be best from an evolutionary perspective: the normative outstrips the natural. So, this is not a great problem for proponents of LDAs.
11. For a further response to Tersman's argument, see *infra* note 22.
12. *Good Reason to Believe Unreliable**—*supra* note 9—implies only slightly different principles, namely that the following third clauses be added to these three principles:

    Epistemic Premise . . . .and (iii) we have good reason to believe that whether X occurred is irrelevant to whether p that is independent of our beliefs about whether p . . .*

    Debunking Premise . . . .and (c) we have good reason to believe that whether X occurred is irrelevant to whether p that is independent of our beliefs about whether p . . .*

    Vindicatory Premise . . . .and that (c) we have good reason to believe that whether X occurred is irrelevant to whether q that is independent of our beliefs about whether q . . .*
13. Finlay and Schroeder (2017).
14. See the discussion below.
15. See also Williams (1995a: 36), Markovits (2014: 55), and Smith (1995).
16. See Williams (1995a: 39) and Raz (2011: ch. 2).
17. Stratton-Lake (2002: xxv-xxvi).
18. It has been put to me that this judgment might be debunked because there will be an evolutionary story about why we find analogies compelling. However, we lack *good reason to believe* a particular story and we lack *good reason to believe* that the process identified is irrelevant to the truth of whether practical and theoretical reasons are the same kind of thing; the process at work might, for instance, be the

process that allows us to understand truths about logic or mathematics. And that process may be truth-tracking; cf. Clarke-Doane (2012) and de Lazari-Radek and Singer (2012: 17-18).

19. See Cowie (2015) and the references therein.
20. See Reisner (2015: 475-476) and Schroeder (2015: §1.1).
21. Thanks to Guy Kahane for pushing me on these two issues.
22. In §3A I argued that we lack good reason to accept an evolutionary debunking explanation of internalist judgments. But do we lack good reason to accept a (non-evolutionary) *debunking explanation* of internalist judgments? Tersman (2006: 403) suggests that some may accept impartial consequentialism because they overvalue theoretical simplicity and coherence at the cost of ignoring relevant differences in cases that count against such simplicity and coherence; perhaps similarly it could be argued that internalists accept their view due to overemphasizing the importance of having a simple and coherent theory of normative reasons. It seems to me that although this *may* be true we lack good reason to accept it. Furthermore, as I explained in *supra* note 9, in order for a judgment that p to be debunked we need to have good reason to believe that the thing that it is the causal result of is irrelevant to whether p which is independent of our beliefs about whether p. But it does not seem that we have good reason to believe that the valuation of simplicity and coherence that underpins judgments about the truth of impartial consequentialism or internalism *are overvaluations* that is independent of our judgments about the truth of impartial consequentialism or internalism: if we have good reason to believe that Rossian pluralism/externalism is true, we have good reason to believe that impartial consequentialism/internalism is false and so good reason to believe that judgments that impartial consequentialism/internalism is true based on simplicity and coherence are the result of an overvaluation of simplicity and coherence. But these good reasons are not independent of our beliefs about the truth of impartial consequentialism/internalism.
23. Emphasis added.
24. Or at least that we lack good reason to believe the negation of *No Direct Stranger Concern*, which would be enough for the purposes of this paper—see §2. See also de Lazari-Radek and Singer's (2012: 26, n. 45) discussion of personal correspondence with Tersman.
25. Kahane (2014: 329 n. 9)
26. See, for instance, Pleasants (2009: esp. 677).
27. See e.g. Dworkin (1995).
28. Cf. de Lazari-Radek and Singer's case that Sidgwickian impartialist intuitions are not vulnerable to debunking in §1.
29. It might seem that bodily imperatives just consist in pain states. But beliefs that we have non-derivative reasons to avoid pain are vulnerable to evolutionary debunking (see Kahane's argument in §1B). However, Kahane argues that beliefs in reasons to avoid pain that derive from the judgment that we have non-derivative reason to avoid our own pain are vulnerable to evolutionary debunking. But in this section, I've been discussing beliefs that we have reasons to satisfy others' bodily imperatives that do not derive from the judgment that we have reasons to satisfy our own bodily imperatives.

30. If they are, won't they be debunkable for the reasons that de Lazari-Radek and Singer argue that egoist views are? (§1A) No. For they argue that the intuition that we have reasons to promote our own good can be debunked. I've argued that we lack good reason to believe that intuitions such as that epistemic and practical reasons have the same structure can be debunked. And that because of this beliefs in internalism on the basis of intuitions like this one cannot be debunked.
31. See Joyce (2006: 24-26) and the evidence cited therein.
32. See *ibid*. 41 and the evidence and references therein.
33. See Olson (2014: 142).
34. See Stratton-Lake (2002).
35. Buck-passing accounts of well-being are slightly more complicated; see Rowland (2019: ch. 5). But this does not matter for our current purposes.
36. Kahane grants this claim (§1B, §4).
37. I would like to thank audiences at the Australasian Association of Philosophy in Adelaide, the University of Melbourne, and the University of Utrecht, as well as Jessica Isserow, David Killoren, Guy Kahane, Ole Koksvik, and Pekka Väyrynen for comments on previous drafts of this paper.

References

Bogardus, Tomas (2016). 'Only All Naturalists Should Worry About Only One Evolutionary Debunking Argument'. *Ethics* 126: 636–661.

Brunero, John. (2017). "Recent Work on Internal and External Reasons." *American Philosophical Quarterly* 54: 99–117.

Clarke-Doane, Justin (2012). 'Morality and Mathematics: The Evolutionary Challenge'. *Ethics* 122: 313–340.

Cowie, Christopher (2015). 'Conservatism in Metaethics: A Case Study'. *Metaphilosophy* 46: 605–619.

Crisp, Roger (2006). *Reasons and the Good*. Oxford: Oxford University Press.

Darwall, Stephen (2006). *The Second-Person Standpoint*. Cambridge, MA: Harvard.

Dworkin, Gerald (1995). 'Unprincipled Ethics'. *Midwest Studies in Philosophy* 20, 1: 224–239.

Finlay, Stephen (2008). 'The Error in the Error Theory'. *Australasian Journal of Philosophy* 86: 347–369.

Finlay, Stephen and Mark Schroeder (2017). 'Reasons for Action: Internal vs. External'. *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), Edward N. Zalta (ed.), URL = https://plato.stanford.edu/archives/fall2017/entries/reasons-internal-external/

Goldman, Alan (2009). *Reasons from Within: Desires and Values*. Oxford: Oxford University Press.

Greene, Joshua (2008). 'The Secret Joke of Kant's Soul'. In Walter Sinnott-Armstrong (ed.), *Moral Psychology Volume* 3. MIT Press.

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

Kahane Guy (2011). 'Evolutionary Debunking Arguments'. *Noûs* 45: 103–125.

———. (2014). 'Evolution and Impartiality'. *Ethics* 124: 327–341.

Keller, Simon (2013). *Partiality*. Princeton, NJ: Princeton University Press.

Korsgaard, Christine (1986). 'Skepticism about Practical Reason'. *The Journal of Philosophy* 83: 5–25.

———. (1996). *The Sources of Normativity*. Cambridge: Cambridge University Press.

de Lazari-Radek, Katarzyna and Peter Singer (2012). 'The Objectivity of Ethics and the Unity of Practical Reason'. *Ethics* 123: 9–31.
Manne, Kate (2014). 'Internalism about reasons: sad but true?' *Philosophical Studies* 167: 89–117.
———. (2017). 'Locating Morality: Moral Imperatives as Bodily Imperatives'. In Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics Volume*12. Oxford: Oxford University Press.
Markovits, Julia (2014). *Moral Reason*. Oxford: Oxford University Press.
Olson, Jonas (2014). *Moral Error Theory: History, Critique, Defence*. Oxford: Oxford University Press.
Pleasants, Nigel (2009). 'Wittgenstein and Basic Moral Certainty'. *Philosophia* 37: 669–679.
Portmore, Douglas (2011). *Commonsense Consequentialism: Wherein Morality Meets Rationality*. Oxford: Oxford University Press.
Raz, Joseph (2011). *From Normativity to Responsibility*. Oxford: Oxford University Press.
Reisner, Andrew (2015). 'Fittingness, Value, and Trans-World Attitudes'. *The Philosophical Quarterly* 65, 260: 464–485.
Rini, Regina (2016). 'Debunking Debunking: A Regress Challenge for Pscyhological Threats to Moral Judgment'. *Philosophical Studies* 173: 675–697.
Ross, W.D. (1930). *The Right and the Good*. Oxford: Clarendon.
Rowland, Richard (2019). *The Normative and the Evaluative*. Oxford: Oxford University Press.
Scanlon, T.M. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.
———. (2014). *Being Realistic about Reasons*. Oxford: Oxford University Press.
Schroeder, Mark (2007). *Slaves of the Passions*. Oxford: Oxford University Press.
Schroeder, Tim (2015). 'Desire'. In Ed Zalta (ed.), *The Stanford Encylopedia of Philosophy*. https://plato.stanford.edu/entries/desire/#Bib
Sidgwick, Henry (1907). *The Methods of Ethics*, 7th ed. Macmillan.
Singer, Peter (2005). 'Ethics and Intuitions'. *The Journal of Ethics* 9: 331–52.
Smith, Michael (1995). 'Internal Reasons'. *Philosophy and Phenomenological Research* 55: 109–131.
Stratton-Lake, Philip. (2002) 'Introduction', in Philip. Stratton-Lake (ed.) W.D. Ross, *The Right and The Good*. Oxford: Oxford University Press.
———. (2016). 'Intuitionism in Ethics'. *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/intuitionism-ethics/>.
Street, Sharon (2006). 'A Darwinian Dilemma for Realist Theories of Value'. *Philosophical Studies* 127: 109–166.
Tersman, Folke (2008). 'The Reliability of Moral Intuitions: A Challenge from Neuroscience'. *Australasian Journal of Philosophy* 86: 389–405.
Vavova, Katia (2014). 'Debunking Evolutionary Debunking'. In Russ Shafer-Landau (ed.), *Oxford Studies in Metaethics Volume*9. Oxford: Oxford University Press.
———. (2018). 'Irrelevant Influences'. *Philosophy and Phenomenological Research* 96: 134–152.
White, Roger (2010). 'You Just Believe That Because . . . ' *Philosophical Perspectives* 24: 573–615.
Williams, Bernard (1981). 'Moral Luck'. In *Moral Luck*. Cambridge: Cambridge University Press.
———. (1995a). 'Internal Reasons and the obscurity of blame'. In *Making Sense of Humanity*. Cambridge: Cambridge University Press.
———. (1995b). 'Replies'. In J. Altham and Ross Harrison (eds.), *World, Mind, and Ethics: Essays on the Ethical Philosophy of Bernard Williams*. Cambridge: Cambridge University Press.