Accepted for publication at *Pacific Philosophical Quarterly*. Excerpt from penultimate draft. Please do not cite.
Wiegman, Isaac "The Evolution of Retribution: Intuitions Undermined" Pacific Philosophical Quarterly 98.2 (2017), 193-218

Excerpt from "THE EVOLUTION OF RETRIBUTION: INTUITIONS UNDERMINED"

Isaac Wiegman

Abstract: I argue that evolutionary influences on anti-consequentialist intuitions sever their evidential connection with retributive theories of punishment. I suggest that anger places value on actions of revenge and retribution, value not derived from the consequences of these actions. As a result, it contributes to the development of retributive intuitions. Moreover, if anger evolved to produce these retributive intuitions because of their biological consequences, then these intuitions are not a good indicator that punishment has value apart from its consequences. This severs the evidential connection between retributive intuitions and the retributive value of punishment. This argument may generalize to other deontological intuitions and theories.

Deontological moral intuitions are revealed by widespread tendencies to judge or act[1] contrary to an act-consequentialist evaluation of actions.[2] They are often taken to be a primary source of evidence against consequentialist theories. Work in the last decade of empirical moral psychology suggests that emotions are responsible for at least some of these deontological intuitions. A prominent criticism of deontological intuitions, offered by Joshua Greene and Peter Singer, is that plausible evolutionary explanations of altruism reveal epistemic defects in the emotions responsible for these intuitions (Greene, 2008; Singer, 2005). If they are right, then consequentialism is vindicated by undercutting a primary source of evidence against it.[3]

One of Greene's central criticisms is this: emotions were selected for their role in increasing fitness and deontological intuitions are a byproduct of this evolutionary function, thus emotions would have produced deontological intuitions whether or not these intuitions were true. The most common objection to this argument is that it proves too much. It threatens to undercut a wider set of evaluative intuitions (e.g. that pain is bad), ones that share the same kind of evolutionary explanation and some of which support consequentialism and evaluative realism (see e.g. Berker, 2009; Kahane, 2011; Mason, 2011).[4] This is a problematic feature of Greene's and Singer's evolutionary debunking argument, because it vitiates their aim of defending consequentialism. Moreover, there is a suspicion that similar problems will plague any evolutionary debunking argument pitched at the level of first-order moral discourse (e.g. Kahane, 2011; Mason, 2011; Vavova, 2014).

My purpose here is to give an evolutionary debunking argument along these lines but that is not subject to this criticism. Unlike Greene's argument, I focus specifically on the role of retributive intuitions in supporting retributive theories of punishment, according to which (deserved) punishment has positive value aside from its good consequences. And rather than undermining the epistemic value of these intuitions entirely, this argument severs the evidential connection between

retributive intuitions and retributive theories. Thus, the force of the argument is isolated to retributive theories rather than also undermining the wider range of intuitions that seem to support evaluative realism and the theories of value necessary to evaluate outcomes.

In section 1, I propose a friendly amendment to a prominent dual-process account of moral intuition in order to pick out the kind of process that is likely responsible for retributive intuitions. Rather than capturing the difference between retributive and consequentialist intuitions about punishment in terms of the difference between emotion and cognition (as does Greene), I capture the difference with a distinction between *prospective* and *non-prospective* processes. Non-prospective processes place *non-derivative* value on actions (or action types), value that does not derive from the action's consequences. Anger is an example of a non-prospective process, because it places non-derivative value on actions like revenge and retribution.

Once this distinction is in place, I present a novel debunking argument (in section 2). If anger produces retributive intuitions (which are one species of deontological intuition) because of the biological consequences (e.g. increased fitness) of those intuitions, then these intuitions are not good indicators of non-derivative value. This severs the putative evidential connection between retributive intuitions and the non-derivative value of punishment. Thus, retributive intuitions are not good evidence for retributive theories of punishment (according to which punishment has non-derivative value). This argument may generalize to other deontological intuitions, but it does not uniquely favor the consequentialist theories that Greene and Singer attempt to defend. In section 3, I address an objection that has been leveled at Greene and Singer, that the argument proves too much.

1. The dual-process framework: reaction versus prospection

A central part of Greene and Singer's debunking arguments is Greene's dual-process theory of moral intuition according to which consequentialist and deontological intuitions have distinct psychological underpinnings.[5] Moreover, Greene and Singer suppose that the distinct evolutionary etiologies of these psychological processes distinguish the epistemic value of the intuitions they produce. Specifically, the evolutionary history of emotional processes, is supposed to undercut deontological intuitions, whereas they do not undercut consequentialist intuitions, which are supposed to be produced by a cognitive process with a distinct evolutionary history (cf. Singer, 2005, p. 350). Nevertheless, as suggested in the introduction, a successful evolutionary debunking of deontological intuitions may also require distinguishing deontological intuitions from the broader

class of evaluative intuitions in terms of their epistemic value. Below, I make a friendly amendment to Greene's dual-process theory, one that identifies a distinguishing feature of processes responsible for deontological intuitions.

The main contrast Greene employs is between cognitive and emotional processes. Emotional processes have what Greene calls 'behavioral valence', meaning that these 'alarm like' emotions have the following properties. First, they include inclinations to behave in specific ways or to judge those behaviors as appropriate. Second, they are elicited in response to a limited range of factors (such as the presence/absence of personal force), and finally, once triggered they can override cognitive processes. By contrast, cognition involves slow, flexible, and controlled processes (see e.g. Evans, 2003; Stanovich, 2004). Cognition aligns with consequentialism because both are 'systematic and aggregative': '…the advantage of having such neutral representations is that they can be mixed and matched…without pulling the agent in multiple behavioral directions at once…' (Greene, 2008, p. 64) In other words, these representations are supposed to make possible the kind of systematicity characteristic of consequentialism, because they can take all of the consequences of an action into account. This is the most important contrast for Greene: cognitive processes can consider an indefinite number of different factors when deciding how to act, whereas emotional responses are triggered by only a few kinds of factors.

While Greene believes that there is a natural mapping between the content of consequentialism and the properties of cognitive processes and 'between the content of deontological philosophy and the functional properties of [emotional responses]' (2008, p. 63), these distinctions do not align. First, it is possible to imagine processes that consider multiple factors, but that are not fully consequentialist in their deliberations. For instance, retributivism about punishment is a deontological theory since a retributive justification of punishment refers to what a transgressor deserves based on what she did in the past rather than on the good outcomes that would attend punishment.[6] Nevertheless, moral agents can weigh consideration of desert against other considerations to yield an all-things-considered judgment. For instance, I might be motivated to punish a person because of what she deserves, but I might nonetheless adjust the severity of the punishment in relation to the consequences of punishing. In that case, I have a backward-looking, non-consequentialist motive for punishment that aggregates with other factors by a process that can consider whatever factors seem relevant to the question at hand.[7] Thus, what is distinctive about the psychological processes that consequentialism maps onto is not that they consider or weigh multiple factors (since they could easily consider non-consequentialist factors as well), but rather that they are

prospective, or outcome based. That is, they place value on actions according to their anticipated, internally represented, outcomes (relative to the agent) and decide what to do only based on (positively or negatively) valued outcomes.

Notably, work in animal behavior and computational neuroscience has revealed neural systems in human and non-human animals that place value on actions with reference to a *causal model* relating the action to its outcome.[8] In contingency learning experiments, a rat learns not (only) that pressing a bar is a good *kind of action* to perform, but that pressing the bar produces a specific *outcome*, namely the delivery of a certain kind of food (e.g. Balleine & Dickinson, 1998). If the rat is satiated with (or conditioned to aver) that kind of food, or if pressing the bar ceases to deliver it, the rat will diminish its bar pressing behavior. One of the best explanations of these patterns is that the rat develops a causal model of the outcome of pressing the bar and changes its actions when the hedonic outcome is less favorable or when the causal model updates to include a relevant change in contingency.

A second problem with Greene's alignment of distinctions is that we can imagine there being emotion-like responses that are only sensitive to one or two factors, but that nonetheless line up with consequentialist rationales. For instance, one could design a robot with an alarm-like response to the detection of a doomsday device. When triggered this response would do whatever is required to destroy the doomsday device. Moreover, the emotion-like response need not take anything else into account (besides the existence of an armed doomsday device) to accord with a purely consequentialist judgment about what to do. Thus, being triggered in response to a few kinds of factors is not a distinctive feature of processes that produce deontological intuitions. Rather, I think the kinds of processes that map onto deontological intuitions, if any, are *non-prospective* processes.

Non-prospective processes form a disjunctive class because there are many ways to motivate action and moral judgment that are not prospective. For instance, some processes place value on actions according to past experiences involving actions *of that kind*. In a recent experiment (Cushman, Gray, Gaffey, & Mendes, 2012), participants were asked to physically enact simulations of harmful actions (e.g. holding a fake gun to another person's head and pulling the trigger). Physiological indicators of aversion prior to performing these actions were greater than when participants performed nearly identical actions (e.g. holding a spray bottle in the air and pulling its lever) and greater than when they observed other people physically simulating the same harmful actions. The aversive reaction to these actions is not readily understood as an aversion to their

consequences (e.g. the pain or discomfort of the 'victim'). Rather the better explanation is that participants had a stronger aversion to performing actions of a certain *type*, namely ones that resemble taking someone's life by putting a gun to their head and pulling the trigger.

Other non-prospective processes are *reactive*, in the sense that the aim of action (or the reason for which it is selected) is represented in relation to past or present occurrences (as opposed to internally represented future outcomes). While such a process may be directed at a future outcome in some external sense (e.g. directed at the outcome by design), it is not guided by an internal representation of a future outcome. For such a process, '…the orientation towards a future state…can merely involve change from the present, – change from now: disappearance of pain, disappearance of the desired object being out of reach' (Frijda, 2010, p. 572). For instance, a heat-seeking missile need not be guided by an internal representation of its target or of the 'desired' outcome of hitting its target. Of the many ways one could design such a missile, one of the simplest would be for it to receive feedback signals that indicate whether a heat source is reducing or increasing its distance in a given direction. When appropriately connected to its controls, this feedback can guide the missile to its target. Again, one need not include in the missile's program any internal representation of its aim (hitting a moving object) or its physical target (such as the geometrical structure or distinctive color patterns of the target or its spatial location relative to other objects). It only requires feedback signals that adjust its path in reaction to the path of its physical target and that direct it toward the achievement of its aim.

There is evidence from animal behavior, neuroscience and psychology that prospective and non-prospective systems operate independently of each other (and sometimes antagonistically) in a range of animal species including humans.[9] Moreover, it is likely that in many of Greene's examples reactive processes are producing the empirical results that Greene attributes to alarm-like emotions.[10]

2. A Novel Debunking Argument

2.1 An adaptive function of anger

With the distinction between prospective and non-prospective processes in hand, we can now turn to the novel debunking argument. If some non-prospective processes were selected for their fitness enhancing consequences but also cause deontological intuitions because of these consequences, then the function of non-prospective processes (producing biologically favorable outcomes) disconnects them from the states of affairs that intuitions report (that actions have value aside from their outcomes). Thus, these deontological intuitions are not good evidence for the

principles, beliefs, and theories that they seem to support. In this section, I sketch out this argument with reference to a specific reactive process, namely anger.

Consider a suggestion about anger similar to those made by Greene, Singer and even Derek Parfit (2011, p. 429): anger is responsible for the intuitions that support an anti-consequentialist, *retributive* principle.

> R – The value (or justification) of an act of punishment is not (or not only) derived from the consequences of the act (or the practice) of punishment.[11]

Greene appeals to evolutionary accounts of altruism to explain why an emotional response like anger would lead to the intuitions that support R: '…we have a taste for retribution, not because wrongdoers truly deserve to be punished regardless of the costs and benefits, but because retributive dispositions are an efficient way of inducing behavior that allows individuals living in social groups to more effectively spread their genes' (Greene, 2008, p. 71). Here as elsewhere, Greene is appealing to evolutionary considerations of the sort that would also undermine the theories of value that consequentialism requires (cf. Street, 2006). For the argument to work in favor of consequentialism, we need a slightly different story about why the evolutionary function of anger makes it untrustworthy with respect to R.

To understand this function, it helps to get oneself in the grip of a puzzle. It is obvious that humans cooperate in a wide range of circumstances. There is cooperation not only among people who are genetically similar (a phenomenon that the theory of kin selection explains) and not only among people that frequently interact (a phenomenon that theories of direct reciprocity explain) and not only among people who signal an intention to reciprocate (a phenomenon that indirect reciprocity and costly signaling explain) but also, puzzlingly, '...among genetically unrelated people, in non-repeated interactions, when gains from reputation are small or absent' (Fehr & Gächter, 2002, p. 137). This last kind of cooperation is beneficial because it may have allowed our ancestors to live in larger groups and receive the benefits of doing so (E.g. Richerson & Boyd, 2005). Nevertheless, it remains controversial how we evolved the tendency to do so. Punishment could help explain this phenomenon, because it can help to secure cooperation. Nevertheless, punishment is costly. Even granting that the group level benefits of punishment (the ones that accrue to individuals in large cooperatives) outweigh the immediate costs (e.g. Boyd, Gintis, & Bowles, 2010), what would motivate relatively short-sighted individuals to consistently punish in the conditions characteristic of life in large groups (non-iterative interactions amongst genetically heterogeneous

individuals where reputation is difficult to track)? What would make someone willing to commit to punishment in the face of its momentary costs?

Fehr and Gächter set out to answer these questions (among others) with an economic game (Fehr & Gächter, 2002). They set up the purest instance of the kind of interactions that we are concerned with (ones that are non-repeating and anonymous) by having groups of four people (anonymous to each other) play a 'public goods' game on computer consoles. They gave participants an endowment of 20 monetary units (MUs) and then gave them the opportunity to invest it in a group project. The group project would return .4 MUs to each group member for every one MU invested in the project, and at the end of the round, each player received information about how much the other group members donated. To see how participants changed their strategy over time and with changing conditions, Fehr and Gächter had participants play the game several times, but they told participants that they would never interact with anyone more than once.

The structure of this game creates incentives and costs that militate against donating much to the project. For instance, the best possible outcome for any one individual would be to invest none in a case where everyone else went all in. In that case, the free-rider would walk away with 44 MUs, whereas everyone else would gain a modest 4MUs, walking away with 24. Moreover, if someone is the only one to contribute all one's endowment, then that person will end up with less than half of the initial endowment, while everyone else turns a profit. These incentives influence how people actually play the game. After six iterations of one version of the game, Fehr and Gächter report, '...58.9% of the subjects contributed nothing and 75.6% contributed 5 MUs or less' (2002, p. 138).

In another version of the game, Fehr and Gächter introduced the possibility of punishment. They told participants that at the end of the round (after receiving information about how much others in the group invested) each participant had the opportunity to spend some of their MUs to punish another group member. For each MU contributed toward punishing an individual, that individual would lose three MUs (with the loss was capped at 30). With the possibility of punishment in place, the average investment immediately shot up to more than 12 MUs. On the sixth iteration of the punishment condition, Fehr and Gächter report, '...38.9% of the subjects contributed their whole endowment and 77.8% contributed 15 MUs or more' (2002, p. 138). Not only was the threat of punishment effective in securing cooperation, punishment also occurred quite frequently, with more than 80% of subjects punishing at least once across the six iterations of the punishment condition.

Conjecturing that emotions were responsible for this pattern of punishment, Fehr and Gächter followed up at the end of the game with questions about participants' feelings toward another player given how that player's investment compared to that of the rest of the group: 'You decide to invest [5] francs to the project. The second group member invests [3] and the third [7] francs. Suppose the fourth member invests 2 francs to the project. You now accidentally meet this member. Please indicate your feeling towards this person' (2002, p. 139). Subjects rated both their anger and annoyance at the free-rider on a seven-point scale (one being 'not at all' and seven being 'very much'). Even with this rather modest discrepancy, 17.4% of participants indicated 'very much' anger. Moreover, the higher the discrepancy between the investment of the free-rider and that of the other group members, the more anger participants reported (84% indicated five or greater in response to a vignette with a greater disparity between the free rider and others). Punishment in the public goods game followed a similar pattern. The higher the discrepancy was between the free-rider's investment and the average investment of the others, the greater the punishment (the more MUs that were lost). Moreover, the most common form of punishment in the game was when above-average investors punished below-average investors.

For my purposes, this study holds a pinch of evidence and a generous helping of illustration. It provides a pinch of evidence that anger is a reactive process that results in punishment and that punishment secures cooperation of the relevant kind. The study suggests that anger is a reactive process in that it leads to an impulse to punish in response to the *past* actions of a free rider even in cases *in which anticipated outcomes (such as maximizing profits) do not favor punishment*. This is plausible because participants knew that they would not encounter the free-rider again, yet participants punished free-riders proportionally to the severity of their free-riding. Moreover, they punished free-riders just as frequently on the last round of the game, in which no one in further iterations of the game would benefit from punishment. Finally, the threat of punishment in the game secured the benefits of cooperation even between people who were unrelated, who had no idea of each other's reputation, and who had reason to think they would not interact with each other again. Importantly for my purposes, the study also illustrates how anger could possibly play an adaptive role in securing cooperation across a broad range of conditions *because it is a reactive process*. It supports a 'how possibly' explanation for the evolution of altruism according to which a reactive process was selected for its role in securing cooperation, and that is all I really need if I want to show how evolutionary considerations could, in principle, undercut deontological intuitions without undercutting evaluative intuitions more broadly.

2.2 Severing the evidential connection between retributive intuitions and retributive theories

Let us suppose that this adaptationist story is correct and that anger is a reactive process. In order to secure the non-immediate consequences of cooperation, the reactivity of anger leads us to act based on the past action of the free-rider rather than because of immediate outcomes relevant to punishment (e.g. improved investment in a cooperative venture). In so acting, we manifest an inclination toward punishment that is not motivated by the consequences of punishment. If so, then in the context of punishment, anger includes a set of inclinations to act in accordance with R (to act as if the value of punishing did not depend on its consequences). However, if we are inclined to judge and act as if the value of punishment is not based on outcomes precisely because this feeling was adaptive for securing good outcomes, then the feeling is not trustworthy regarding R. There is a disconnect between, on the one hand, the value of punishment reported by intuition and, on the other hand, the manner in which these intuitions arose. According to R, punishment (whether an act or a policy) is supposed to have value apart from its consequences but these intuitions arose because of their consequences.

Suppose for the sake of argument that there is some state of affairs that makes punishment valuable independently of its consequences (or a state of affairs that makes R true). The retributive intuitions produced by anger do not plausibly have any evidential connection to that state of affairs. Anger produces intuitions that support R because such intuitions deter freeriding, but the fact that these intuitions deter freeriding lacks an evidential connection to the states of affairs that R reports. More specifically, deterrence (or any other consequence of punishment) cannot be an *indicator* of any value that a punishment might have aside from its consequences (for reasons that I discuss in more detail in the following section). Thus, the putative evolutionary function of anger in the production of retributive intuitions shows that these intuitions are not good evidence for R.[12]

Perhaps an analogy will flesh out the line of reasoning. Suppose that Geppetto is designing the psychology of a cyborg that he calls Pinocchio. Geppetto wants to make Pinocchio very realistic, and his aesthetic sensibilities favor a slightly scrawny boy. He foresees that this design preference will result in real boys picking on Pinocchio. Thus, he programs into Pinocchio a strong drive to resist bullies. He reasons that the policy of resisting bullies, even in cases where immediate consequences militate against doing so, will lead Pinocchio to suffer less from bullies in the long run. Bullies will realize that it is less costly to pick on other scrawny boys who are less scrappy, and they will bother Pinocchio less as a result. Geppetto wants Pinocchio to have the capacity for

prospection, but Geppetto cannot guarantee that Pinocchio will be able to consistently anticipate the long-term value of resisting bullies.[13] Therefore, Geppetto designs Pinocchio with a drive to resist bullies that is not derived from the immediate prospective value of doing so. This drive gives Pinocchio an urge to react to the provocations of bullies rather than only to respond to the immediate prospects (largely negative) of doing so.[14] To Pinocchio, the urge to resist is there whether or not it will result in a good outcome, thus to him, the urge does not to derive from the anticipated outcome of resisting (nor from the anticipated outcome of a policy of resisting).[15] Once Geppetto completes his design, Pinocchio will tend to act and judge in accordance with the principle that resisting bullies has value not derived from its consequences or *non-derivative value*.[16] He might even discover that his intuitions about resisting bullies support the following principle and come to consciously believe it.

B - The value of an act of resistance toward bullies is not (or not only) derived from its consequences (or from the consequences of the policy of resisting).

From the third-person perspective, it seems obvious that Geppetto's design has distorted Pinocchio's axiological beliefs about resisting bullies. If someone were to tell Pinocchio of Geppetto's design choices, he should no longer believe B. With the right information, he should conclude that his inclinations to resist bullies are not good evidence for B. Since Pinocchio's inclinations to resist bullies are disconnected from any source of value that resistance might have aside from its consequences, the intuitions are not good evidence for the principle. Therefore, Pinocchio should not believe B on the basis of his intuitions.

If this argument is compelling in Pinocchio's case, then it should also be compelling in the case of R. If the adaptationist story about anger is correct, then the intuitions that support R secure good biological consequences in the long run just as they would if they were designed to do so. As such, they are similarly disconnected from any non-derivative value that punishment might have. Thus, insofar as anger influences our intuitions about punishment, we are not justified in believing that punishment has non-derivative value on the basis of intuition.

2.3 A more detailed explanation of the disconnect

The argument I have offered is supposed to sever the evidential connection between retributive intuitions and the retributive principle R by showing that retributive intuitions are not a good indicator of non-derivative value (reported by R). How does this work? To answer this question, let us take a closer look at the concept of indication. One state of affairs can indicate

another if the states are highly correlated with one another, either *because one state of affairs causes (or constitutes) the other or because both states of affairs have a common cause (or constitutive base)* (Dretske, 1999).[17] The requirement of a causal or constitutive dependency relation between two variables, is intended to rule out coincidental correlations between them. For instance, from 1999 to 2009 there was a strong correlation between the number of people who drowned in swimming pools and the number of films that Nicolas Cage appeared in (reported at www.tylervigen.com). However, this does not mean that Cage appearances are an indicator of drowning deaths (or vice versa), because the correlation could be entirely coincidental.

Now, even if there were a correlation between retributive intuitions (the inclinations that support R) and non-derivative value (for punishment), it would be entirely coincidental (notwithstanding a 'pre-established harmony' ordained by God). Thus, none of the relevant causal or constitutive dependency relations are likely to obtain. To see this, consider the three possible dependency relations that might explain such a correlation. First, suppose that retributive intuitions cause or constitute a state of affairs in which punishment has non-derivative value. If non-derivative value of punishment was constituted by the existence of retributive intuitions and if retributive intuitions were selected for their deterrent value, then the non-derivative value of punishment would depend on (either causally or constitutively) the deterrent value of retributive intuitions. Such a dependency seems impossible. Given the definition of non-derivative value, there should be cases in which punishment has non-derivative value but in which the deterrent value of retributive intuitions fails to obtain. For instance, there are possible contexts in which retributive inclinations do not deter freeriders (in public goods games like the one discussed above). When those with retributive inclinations represent only a small percentage of a population, punishment for freeriding becomes so unlikely as to obliterate the deterrent effect of these inclinations (see e.g. Bowles & Gintis, 2004). It seems that if punishment has non-derivative value, then it should retain its value even in those circumstances. But if this is correct, then any correlation between the deterrent value of the intuition and the non-derivative value of punishment will be a coincidence, due to the fact that most of the time, acts of punishment *just happen* to occur in a context in which retributive intuitions do have deterrent value.

Second, consider the contrary dependency relation: suppose that the non-derivative value of punishment causes or constitutes the states of affairs in which retributive intuitions exist or are manifested. This possibility seems to be almost entirely ruled out by the evolutionary explanation of retributive intuitions. If that explanation is correct, then we would have retributive intuitions

whether or not punishment has non-derivative value. So it also seems unlikely that the manifestation of retributive intuitions depends in any way on the non-derivative value of punishment. Moreover, if these intuitions were selected for their good outcomes, then there is no reason to think that their manifestation would depend on punishment having non-derivative value. Rather, their manifestation will depend on whatever conditions are necessary for them to have deterrent value. Of course, retributive intuitions might happen to be manifested in cases in which punishment has non-derivative value, but this would appear to be entirely coincidental.

Finally, suppose that retributive intuitions (either their existence or manifestation) and the non-derivative value of punishment have a common cause.[18] This too seems impossible if the evolutionary explanation is correct. Again, the existence of retributive intuitions depends only on their deterrent effect. So it is hard to see how some other factor could cause retributive intuitions to exist and also cause punishment to have non-derivative value. Likewise, it is hard to see how a third factor could cause *instances* (as opposed to punishment generally) of punishment to have non-derivative value and cause retributive intuitions to be manifested. If these intuitions were selected for their deterrent function and if this function depends on the frequent manifestation of retributive intuitions, then it appears unlikely that non-derivative value would be caused by any of the conditions that elicit retributive intuitions. This is because retributive intuitions would tend to be elicited when punishment has deterrent value, and it is unclear why those conditions would cause instances of punishment to have non-derivative value. If the two happen to co-occur then this would be entirely coincidental.

The problem with all of these possibilities is that non-derivative value and deterrent value are, by their definition, independent sources of value. As a result, they are constituted by different facts, and any connection between the two sources of value (aside from divine intervention) will be coincidental. Non-derivative value is constituted by facts about an action aside from its consequences, whereas deterrent value is constituted by facts about the action's outcome aside from the intrinsic features of an action or what came before the action. Any overlap between these sources of value is likely to be coincidental.

2.4 What the argument does not show

This argument might tempt someone to conclude that retributive inclinations are untrustworthy in the sweeping sense that we are never justified in acting or judging on their basis. Additionally, one might think that the debunking argument also undermines the influence that

retributive intuitions might have on consequentialist theories of punishment, especially those that justify retributive practices of punishment based on the consequences of such practices. Neither of these conclusions follow from the debunking argument as I understand it.

A key feature of the argument is that it only severs the evidential connection between retributive intuitions and the retributive principle R. It does not undercut their evidential support of other beliefs or principles (ones that do not posit non-derivative value). The Pinocchio example helps to illustrate this. Even if he stops believing that resisting bullies has non-derivative value, Pinocchio may be warranted in acting on his urge to resist. Geppetto's forethought and beneficent design may give Pinocchio some warrant for acting on his urge to resist bullies. They have practical value, even though Pinocchio must prune their apparent axiological implications (the existence of non-derivative value). Likewise, my debunking argument provides no additional reason to mistrust our retributive intuitions when applied to practical questions about when to punish as opposed to axiological questions about whether it has non-derivative value. However, we cannot guarantee that natural selection gives us retributive intuitions that are morally valuable.

Put in a slightly different way, the argument debunks retributive intuitions as evidence for retributive standards of rightness, but does not present evidence against the value of a retributive decision procedure (which might be adopted on a consequentialist theory of punishment). A retributive standard of rightness might say that punishment is right if and only if it is deserved (given the nature of a past offense). If my debunking argument is sound, then retributive intuitions do not provide good evidence for such a standard, at least insofar as this standard implies that punishment has non-derivative value (e.g. value that attaches to giving someone what they deserve). Whereas acceptance of a retributive standard commits one to non-derivative value, acceptance of a retributive decision procedure makes no such commitment.[19] This is because such a procedure might very well be valuable or warranted because of its good consequences (e.g. deterrence). As a result, this debunking argument also leaves room for consequentialist theories to accommodate the retributive intuitions that support R.

By analogy, Pinocchio may very well be justified in accepting the following decision procedure: if you have an urge to resist the bully, then do so. Pinocchio can follow such a decision procedure while keeping in mind that there is really no non-derivative value that attaches to resisting bullies. Similarly, one could choose to follow a retributive decision procedure because one believes that the consequences of doing so align with one's moral or non-moral aims while also believing that punishment lacks non-derivative value (or that it does not have any value aside from the

consequences of the decision procedure). On either way of putting the point, the only severed evidential connection is that between our retributive intuitions and R. I cannot think of any reason to think that they are untrustworthy in the more sweeping sense.

This is a key difference between my debunking argument and those of Greene and Singer. Rather than only vindicating act-consequentialism over and above theories that accommodate deontological intuitions (including certain consequentialist theories), it favors consequentialist theories quite generally. This is because the main force of my argument is to undercut the evidential support of deontological intuitions for deontological theories. These theories conflict with consequentialist theories (in general) because they have opposing standards of rightness, ones that posit non-derivative value for various actions. In fact, my argument fits nicely with many indirect forms of consequentialism (e.g. rule-consequentialism, virtue-consequentialism), because it may leave room for these theories (but not deontological theories) to legitimately rely on (or conform to) deontological intuitions. For instance, one might think that given their evolutionary function, retributive intuitions are a reliable indicator that a retributive decision procedure can be justified by its consequences. However, I emphasize that my argument does not necessarily provide any positive support for this consequentialist theory of punishment.

2.5 Generalizing the argument

Importantly, the conclusion of the debunking argument only pertains to retributive moral intuitions, not to all deontological intuitions. Nevertheless, if the evolution of other moral emotions follows this same pattern (placing non-derivative value on action in order to improve fitness), then a similar argument can be given concerning deontological intuitions more broadly. Greene provides reasons to generalize (see esp. Greene, 2008, pp. 59–60, 72). The idea is that moral emotions are domain-specific adaptations, where the specific domain of each moral emotion is a recurring situation that constitutes one of the '…demands and opportunities created by social life' (2008, p. 60). In programming us with emotions, nature declines to '…leave it to our powers of reasoning to figure out that saving a drowning child is a good thing to do' (2008, p. 60) or that hurting others and lying are bad things to do. In other words, many moral emotions lead us to react to specific kinds of situations (ones that involve *inter alia* assistance, punishment, promises, testimony, and incentives to harm) in specific ways rather than responding only to the prospective value of acting. This is because the immediate prospects these situations present lean in favor of declining assistance, avoiding confrontation and punishment, breaking promises, lying to others, and doing physical harm

to get what one wants. Moreover, acting against these inclinations is supposed to be adaptive for its role in supporting human cooperation. If this is right, then we can generalize the argument to undercut intuitions that support a broader range of deontological principles, specifically, any deontological principle that the value of a certain action is not (or not only) derived from its consequences. The case of anger and punishment allows us to see the kind of problem that would be raised for other deontological intuitions. Nevertheless, from a methodological perspective, the case needs to be made one moral emotion and one deontological principle at a time.[20]

3. Objections and replies

3.1 Does the argument prove too much?

While there is room for optimism about how the argument might be extended, it seems to face a serious objection. One might think that the argument also severs the evidential relation between evaluative intuitions more broadly and the sources of value they report. For instance, nociceptive processes produce intuitions that seem to support the claim that that pain is bad, and they were selected to do so because of their tendency to aid survival. The worry is that my debunking argument may force the acceptance of this more general debunking argument. Intuitions about pain report the objective badness of pain, whereas nociceptive processes produce those intuitions because of the biological badness of bodily insult and injury. If there is a disconnect between derivative and non-derivative forms of value, this may commit one to a similar disconnect between moral badness and biological badness. If so, then the evidential relation is also severed between these intuitions and the principles they support. Therefore, the objection goes, the case against non-prospective processes seems to again prove too much, since it also undercuts the evaluative intuitions that support a theory of value (necessary for consequentialism). In other words, the worry is that it does not undermine deontology any more than it undermines consequentialism.

3.2 Reply

While evaluative beliefs and intuitions may be on shaky footing, their footing is independent of deontological intuitions. This is because my argument against deontological intuitions severs their evidential relation with non-derivative value by showing that deontological intuitions cannot indicate non-derivative value. The argument severs this specific evidential connection not only because deontological intuitions were shaped by evolution but also because of the content of the deontological principles in question: that certain actions have non-derivative value. If an inclination

was selected for good outcomes like deterrence, then it cannot indicate that acting in accordance with that inclination has non-derivative value (as deontological principles state). Thus, I am giving a reason to doubt the evidential value of deontological intuitions (with respect to deontological principles) that we do not have for doubting other evaluative intuitions (ones that do not posit non-derivative value).

It is an entirely different question whether an inclination that produces good biological outcomes can indicate that certain actions have moral value (regardless of whether this value is non-derivative). This is the question that more global evolutionary debunking arguments attempt to answer (e.g. Joyce, 2007; Street, 2006). Moreover, it is a question that does not hinge on whether non-derivative value enters into the content of evaluative beliefs (as does the question I attempt to answer). My argument focuses on the relation between derivative and non-derivative value, whereas these debunking arguments focus on the relation between biological and moral value. There are important differences between these relations. As I argued above (in section 3.3), non-derivative value and deterrent value are, by their definition, independent sources of value. By contrast, there is room in conceptual space for biological goods like survival to (partially) constitute moral goods like flourishing (see e.g. Enoch, 2010).

The point is that the argument that I have given severs the evidential connection between deontological intuitions and deontological principles in a way that does not apply to evaluative intuitions more broadly (since other evaluative intuitions do not posit non-derivative value), so the debunking argument does not prove too much. Thus, we have a clear case in which evolutionary considerations have implications for first order moral discourse, somewhat independently of metaethical concerns raised by the evolution of our moral faculties.

4. Conclusion

In their efforts to undercut some of the primary evidence against their theories, some consequentialists have employed evolutionary debunking arguments against deontological intuitions. So far, these arguments have met with little success, because the evolutionary considerations offered may undercut a much wider range of intuitions. I presented a new argument that overcomes these challenges. I argued that non-prospective processes – processes that motivate action for reasons aside from the consequences of action – might explain a range of deontological moral intuitions. These intuitions seem to support anti-consequentialist principles, according to which the value of various actions does not derive from the action's outcome (among other things). Nevertheless, non-

prospective processes cause deontological intuitions because they were selected to do so for their biological outcomes. If so, then the evidential connection is severed between deontological intuitions and non-derivative value.

Of central importance for the argument I have just given, the defect in non-prospective processes is not just that they aim at reproductive fitness, but more specifically that they produce deontological intuitions *by influencing organisms to react to certain types of situation* rather than approach them prospectively. The adaptiveness of this tendency reveals a disconnect between the intuitions that these processes produce and the states of affairs they seem to report (through their apparent support of principles like R). This specific defect only undermines deontological intuitions. The argument does not apply to evaluative intuitions more broadly because biological values and objective sources of value do not necessarily have an independent constitutive base as do derivative and non-derivative sources of value.

---

[1] I intend intuitions to include non-inferential inclinations to judge a proposition only by considering its content and non-inferential inclinations (not) to perform an action only by considering some representation of an action or situation (cf. Sinnott-Armstrong, 2008, p. 209; Sosa, 2007, p. 233). The non-inferential nature of intuitions refers to the fact that one can have an intuition that something is right or wrong, good or bad without any accompanying justificatory explanation for the feeling (of the sort from which the content of the intuition might have been inferred). On my view, 'intuition' is a theoretical term capturing a set of phenomena with common explanatory elements (perhaps they are all caused by some psychological process or another) that are explananda or objects of study in fields like philosophy, moral psychology, behavioral economics and social psychology. Accordingly, one can have an intuition without judging its content true or its practical conclusion prudent or morally right, but I do not think anything hinges on this terminological decision. Some philosophers argue that intuitions are properly understood as judgments. Whether or not there exists phenomena involving *inclinations* to judge or act (regardless of whether they in fact lead to judgment or action) does not depend on how philosophical debate proceeds. If the phenomena exist, then these phenomena are a proper object of study regardless of whether they consistently give rise to actual judgments.

[2] More specifically, they are tendencies to judge or act contrary to an act-consequentialist decision procedure: 'On each occasion, an agent should decide what to do by calculating which act would produce the most good.' (Hooker, 2003). Henceforth, I use 'consequentialism' to refer to act-consequentialism (unless otherwise indicated), and I use 'deontological intuition' and 'anti-consequentialist intuition' interchangeably. Thanks to Julia Driver for pointing out the tendency of empirical moral psychologists to conflate consequentialist decision procedures and consequentialist standards of rightness.

[3] For the sake of simplicity, I follow Singer and Greene in focusing on the contrast between act-consequentialism and deontology, rather than making finer-grained distinctions between consequentialist theories (e.g. between act-consequentialism and virtue-consequentialism). However, the debunking argument I give in section 2 also favors other forms of consequentialism (e.g. rule-consequentialism, virtue-consequentialism and global-consequentialism) over deontological theories, for reasons that I discuss in section 2.4. See esp. n. 11 and 19 for discussion of related issues.

[4] One way of putting this point is that consequentialism is vacuous without a theory of value. Without some idea of what outcomes should be valued or disvalued, there would be no way to determine which action would be right by the lights of a consequentialist moral theory. One would be unable to evaluate the consequences of actions. Moreover, intuitions about what things are valuable will inevitably influence any

theory of value. This makes consequentialism vulnerable to a well known evolutionary critique. For instance, if the arguments of Sharon Street are correct, then we hold many if not all of our evaluative intuitions not because they are true (in the sense required by realist theories of value) but because they allowed our ancestors to more effectively spread their genes (Street, 2006). This applies to even the most uncontroversial of evaluative intuitions: that pain is bad, that it is bad to hurt others, and that it is good to help them. These arguments are supposed to result in a more global form of evaluative skepticism than Greene wants; one that vitiates any value theory on which his consequentialism might draw. Nevertheless, it is hard to see how he could reject this argument without denying premises on which his own argument depends.

[5] The dual-process view has actually been the most successful aspect of Greene's research program. There is a wide range of evidence that there is indeed more than one (though perhaps also more than two) processes responsible for moral intuitions. However, there remains some debate about what distinguishes the two processes and whether a division of the processes aligns with the distinction between consequentialist and deontological intuitions. See (Cushman, Young, & Greene, 2010; Cushman, 2013; Kahane, 2012; Kahane et al., 2012; Paxton, Bruni, & Greene, 2014).

[6] This remains true even if there is a valued outcome, perhaps justice, which attends deserved punishment. The world may be a better place because I punish a transgressor, but it is not a better place only because of the consequences of my action. The thought is that a consequentialist evaluation of action (as right or wrong) '…depends only on [the action's] consequences (as opposed to the circumstances or the intrinsic nature of the act or anything that happens before the act)' (Sinnott-Armstrong, 2003, emphasis mine). So long as consequences are understood in isolation from what happened before the act of punishment, the difference between a world in which the transgressor is punished and one in which she is not cannot only be a difference in the consequences of my action but also in their relation to what came before my action (the transgression). Moreover, one cannot understand the stated justification for punishment (as an act that promotes justice) in isolation from what came before the act. Thus, one cannot understand this justification as a consequentialist justification of punishment. I suspect that recent discussion of retributivism hinge on a different understanding of consequentialism (e.g. Berman, 2011).

[7] See Kahane (2012, pp. 531–533) for a similar argument that deontological reasoning often requires weighing different duties against one another. This too seems like a process that can take many factors into account. Though he does not draw the same conclusions that I do concerning prospective processes.

[8] Though it is not clear to me that these models are essentially causal. Perhaps in humans they can represent other assymetrical dependency relations such as the *in virtue of* relation or other non-causal explanatory relations. Relations of these kinds seem crucial for comparing the value of different possible worlds.

[9] I cannot review this evidence here, and in any case, much of it has been thoroughly reviewed elsewhere, see Cushman (2013) and Crockett (2013). While Cushman focuses on the distinction between two *learning* systems, he does not mean to exclude other kinds of non-prospective action selection mechanisms (personal communication). The psychologist Nico Frijda has long emphasized the importance of impulsive motivation, which has precisely the characteristics of the non-prospective processes that I discuss. See (Frijda, 1986, 2010)

[10] While space does not permit a thorough demonstration of this claim, I will consider Greene's example concerning retributive punishment in more detail below.

[11] There is some psychological evidence for intuitions that support this principle (Carlsmith, Darley, & Robinson, 2002; Carlsmith & Darley, 2008; Carlsmith, 2006). On my view, considerations of parsimony suggest that retributive intuitions should be characterized in opposition to an act-consequentialist decision procedure. Even so, in this case such intuitions seem to support a principle that opposes consequentialist theories more broadly (not just act-consequentialist theories). For instance, Michael Moore argues that intuitions about punishment in specific cases uniquely support principles like R (Moore, 2010, Chapter 4). That is, judgments about individual cases can lead one to believe that one should give a transgressor what she deserves even if neither the act nor the practice will have good consequences. Nevertheless, this is consistent with saying that the simplest way to characterize these inclinations (taken collectively) is in their pattern of opposition to an act-consequentialist decision procedure.

[12] I do not think this argument hinges on any conflation of biological and moral goods. It seems to me that if the intuitions were shaped to bring about good consequences of any kind (e.g. biological or moral), then they cannot indicate non-derivative value of any kind, whether moral or biological.

[13] In any single encounter with a bully, Pinocchio would anticipate suffering immediate losses that he would not suffer if he did not resist; losses that favor giving in over resisting.

[14] Notice that when a desire is characterized as non-derivative in this way, it need not be indefeasible by consequentialist considerations. That is, overturning or defeating such a desire with a competing desire to maximize consequences does not make the defeated desire any more derivative. For instance, if a bully threatens lethal force, Pinocchio might overcome his urge to resist because of the catastrophic consequences of doing so. However, notice that this would not mean that the urge is derived from anticipated outcomes. That is, when Pinocchio has this urge prior to the threat of lethal force, it is not an urge to bring about an outcome. Rather, the right way to describe the situation is that Pinocchio feels an urge to resist that does not derive from the consequences of doing so, but he does not give in to that urge because of the catastrophic consequences of doing so. Retributive intuitions are similarly defeasible by consequentialist considerations. For instance, I suspect that most retributivists would say that even if punishment were good in itself, it would be reasonable and right not to punish someone because doing so would have catastrophic consequences. That is, retributive intuitions seem defeasible in just the way Pinocchio's non-consequentialist urge could be.

[15] Essentially, Geppetto programs Pinocchio with a subjective commitment device, a concept owing to the work of several economists, (e.g. Frank, 1988).

[16] It is tempting to think that non-derivative value is identical to non-instrumental value. However, I think these concepts are distinct. For instance, a personal insult can be understood as instrumental for 'getting even', but this is not to say that the value of the insult is derived from its consequences. Rather a successful insult is constitutive of 'getting even'. To me, this looks like an example in which an insult has non-derivative value (from the perspective of the insulter), but in which it is understood as instrumental for another aim, namely getting even.

[17] Dretske cashes out the dependency relation in terms of causation. I suspect that the dependency relation that explains a correlation need not be causal; it could also be constitutive. See Berker (forthcoming) for a detailed discussion of the grounding relation in connection with debunking arguments.

[18] We need not consider the possibility that they have a common constitutive base, because as a matter of definition, they do not.

[19] At least, this is true so long as non-derivative value is understood directly in terms of the value of acting and not in terms of value that is derived from the acceptance of a certain ethical decision procedure or practice (which might have a consequentialist justification). Consider an example. Rawls (1955) distinguishes between justifying actions *within a practice* of promising or punishing (both of which involve ignoring certain reasons one might have not to promise or not to punish) and justifying the practice itself (perhaps in terms of its consequences). Reasons that one accepts from within a practice (and which are indirectly supported by the consequences or aims of the practice) do not count as non-derivative reasons if they are ultimately grounded in the consequences of the practice. As a result, indirect consequentialist theories need not be committed to the existence of non-derivative value.

[20] For one, I am not sure that the moral emotions can all be explained in this way. For another, it is not yet clear to me that all cases of non-derivative valuing derive from domain-specific adaptations of this kind.

Bibliography

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, *37*, 407–419.

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

Bowles, S., & Gintis, H. (2004). The evolution of strong reciprocity: cooperation in heterogeneous populations. *Theoretical Population Biology*, *65*(1), 17–28.

Boyd, R., Gintis, H., & Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, *328*(5978), 617–20. doi:10.1126/science.1183665

Carlsmith, K. M. (2006). The roles of retribution and utility in determining punishment. *Journal of Experimental Social Psychology*, *42*(4), 437–451. doi:10.1016/j.jesp.2005.06.007

Carlsmith, K. M., & Darley, J. M. (2008). Psychological aspects of retributive justice. *Advances in Experimental Social Psychology*, *40*(07), 193–236. doi:10.1016/S0065-2601(07)00004-4

Carlsmith, K. M., Darley, J. M., & Robinson, P. H. (2002). Why do we punish?: Deterrence and just deserts as motives for punishment. *Journal of Personality and Social Psychology*, *83*(2), 284–299. doi:10.1037//0022-3514.83.2.284

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences*, *17*(8), 363–6. doi:10.1016/j.tics.2013.06.005

Cushman, F. (2013). Action, outcome, and value: a dual-system framework for morality. *Personality and Social Psychology Review*, *17*(3), 273–92. doi:10.1177/1088868313495594

Cushman, F., Gray, K., Gaffey, A., & Mendes, W. B. (2012). Simulating murder: the aversion to harmful action. *Emotion*, *12*(1), 2–7. doi:10.1037/a0025071

Cushman, F., Young, L., & Greene, J. D. (2010). Our multi-system moral psychology: Towards a consensus view. In *The Oxford handbook of moral psychology* (pp. 47–71). Oxford: Oxford University Press.

Dretske, F. I. (1999). *Knowledge and the Flow of Information* (p. 273). Cambridge: MIT Press.

Enoch, D. (2010). The epistemological challenge to metanormative realism: how best to understand it, and how to cope with it. *Philosophical Studies*, *148*(3), 413–438. doi:10.1007/sl

Evans, J. S. B. T. (2003). In two minds: dual-process accounts of reasoning. *Trends in Cognitive Sciences*, *7*(10), 454–459. doi:10.1016/j.tics.2003.08.012

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*(6868), 137–40. doi:10.1038/415137a

Frank, R. H. (1988). *Passions within reason: The strategic role of the emotions*. New York: Norton.

Frijda, N. H. (1986). *The Emotions*. *The Emotions* (Vol. 1, p. 564). Cambridge: Cambridge University Press. doi:10.1093/0199253048.001.0001

Frijda, N. H. (2010). Impulsive action and motivation. *Biological Psychology*, *84*(3), 570–9. doi:10.1016/j.biopsycho.2010.01.005

Greene, J. D. (2008). The secret joke of Kant's soul. In W. Sinnott-Armstrong (Ed.), *Moral Psychology, Vol. 3, The Neuroscience of Morality: Emotion, Disease, and Development* (pp. 35–80). Cambridge: MIT Press.

Hooker, B. (2003, December 31). Rule Consequentialism. In *Stanford Encyclopedia of Philosophy*.

Joyce, R. (2007). *The Evolution of Morality*. Cambridge: MIT Press.

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

Kahane, G. (2012). On the Wrong Track: Process and Content in Moral Psychology. *Mind & Language*, *27*(5), 519–545. doi:10.1111/mila.12001

Kahane, G., Wiech, K., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2012). The neural basis of intuitive and counterintuitive moral judgment. *Social Cognitive and Affective Neuroscience*, *7*(4), 393–402. doi:10.1093/scan/nsr005

Mason, K. (2011). Moral Psychology And Moral Intuition: A Pox On All Your Houses. *Australasian Journal of Philosophy*, *89*(3), 441–458. doi:10.1080/00048402.2010.506515

Moore, M. (2010). *Placing blame: A theory of the criminal law*. Oxford: Oxford University Press.

Parfit, D. (2011). *On What Matters, vol. 2*. Oxford: Oxford University Press.

Paxton, J. M., Bruni, T., & Greene, J. D. (2014). Are "counter-intuitive" deontological judgments really counter-intuitive?: An empirical reply to Kahane et al. (2012). *Social Cognitive and Affective Neuroscience*, *9*, 1368–1371.

Rawls, J. (1955). Two Concepts of Rules. *The Philosophical Review*, *64*(1), 3. doi:10.2307/2182230

Richerson, P., & Boyd, R. (2005). *Not by genes alone: How culture transformed human evolution*. Chicago: University of Chicago Press.

Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics*, *9*(3), 331–352.

Sinnott-Armstrong, W. (2008). Framing Moral Intuitions. In W. Sinnott-armstrong (Ed.), *Moral Psychology, Vol. 2, The Cognitive Science of Morality: Intuition and Diversity*. Cambridge: MIT Press.

Sosa, E. (2007). Experimental Philosophy and Philosophical Intuition. *Philosophical Studies*, *132*(1), 99–107. doi:10.1007/sll098-006-9050-3

Stanovich, K. E. (2004). *The Robot's Rebellion* (p. 358). Chicago: University of Chicago Press.

Street, S. (2006). A Darwinian dilemma for realist theories of value. *Philosophical Studies*, *127*(1), 109–166. doi:10.1007/sl

Vavova, K. (2014). Debunking Evolutionary Debunking. In *Oxford Studies in Metaethics 9* (pp. 76–101). Oxford: Oxford University Press.