The evolutionary debunking of morality
Richard Joyce

Penultimate draft of paper appearing in J. Feinberg & R. Shafer-Landau (eds.),

In The Descent of Man (1871) Darwin provides a detailed account of the origins of the human “moral sense”—a trait that he classes as the “most important” difference between humans and other animals. Readers were quick to worry that this evolutionary treatment of human morality in some manner undermined it. One fierce critic wrote that if Darwin’s view became widely adopted “the consequences would be disastrous indeed! We should be logically compelled to acquiesce in the vociferations of [those] who would banish altogether the senseless words ‘duty’ and ‘merit’” (Mivart 1871/2008: 204). Another called Darwin’s position “dangerous” and expressed concern that his views on the origins of morality “aims … a deadly blow at ethics” (Cobbe 1872: 10). Darwin himself remained unfazed; his general attitude to academic philosophy seems to have matched that of many contemporary scientists: slightly suspicious bafflement. But might these worries nevertheless be well-founded? Might a Darwinian account of moral genealogy somehow debunk morality? In order to answer this, we must first say something about what “a Darwinian account of moral genealogy” might involve, and also delineate different kinds of “debunking.”

Moral nativism

Sometimes “nativism” is used to denote a claim about individual development: that a trait is “inborn” rather than acquired. In evolutionary contexts, however, it denotes the claim that a trait is an adaptation: the trait exists and was transmitted from our ancestors because it improved their reproductive fitness relative to competitors. The two uses are clearly not equivalent. Genetic diseases (such as Down syndrome) are inborn but not adaptations, and adaptations (such as language use) may require environmental input to become manifest. Moral nativism in the Darwinian context is the thesis that the capacity to make moral judgments is a human adaptation: The reason we classify the world in moral terms (good, bad, right, wrong, etc.) is that doing so helped our ancestors make more babies than those competitors lacking the moralizing trait.2

Moral nativism is compatible with moral variation. The nativist may claim only that a moral faculty is innate, allowing that the socialization process leads people in different cultures and eras to hold different moral principles and judgments. If, for example, a Naga warrior judges that beheading a foreigner is a noble act, while a Nebraskan teenager judges such an action to be a heinous crime, both are manifesting the same trait: that of exercising a

---

1 Mivart in fact wrote these words prior to the publication of Descent; but he knew in advance the content of Darwin’s forthcoming book, and when the latter was published wrote a scathing anonymous critique in the Quarterly Review.

2 While it is perfectly legitimate to speak of Darwinian cultural evolution (and thus cultural adaptations), in this context, as a matter of fact, the process is generally limited to traits that are genetically transmitted.
moral faculty. (Analogy: A linguistic faculty may be a human adaptation, but which language one ends up speaking depends on which language one is exposed to as a child.) Moral nativism does not imply that the capacity to make moral judgments will develop irrespective of environmental conditions, nor that this capacity will be universally present, nor that it is essential for being human, nor that there is a “gene for moral judgment.” Nor should moral nativism be confused with the claim that humans are by nature good. To put the point provocatively: It is entirely possible that Hitler’s decisions were as much guided by the workings of an innate moral faculty as were Mother Teresa’s.

Whether moral nativism is true is much debated. It is an empirical hypothesis and must be tested as such: Predictions of the hypothesis must be identified, data must be gathered, and alternative hypotheses must be examined. Here our task is not to delve into that debate, but rather to examine what might follow if the nativist hypothesis is true. But there is one aspect of the nativist hypothesis that makes a big difference to that question, and so a little more must be said. The question is whether the most plausible version of moral nativism implies that moral judgment is a truth-tracking adaptation.

Most biological traits have nothing to do with truth. It makes no sense to say that one’s gall bladder, or any of its activities, is true or false. (Of course, that one has a gall bladder is true or false, but that’s a different matter entirely.) But judgments can be true or false, and thus an evolved faculty designed to produce some kind of judgment does have something to do with truth. It might be argued that the moral faculty governs only feelings and emotions, and that the “judgments” it produces are not really the right kind of thing to be assessed as true or false. But this extreme view is fairly implausible upon examination. It may be granted that the moral faculty has a great deal to do with emotions (like anger and guilt), but that is no reason to jump to the conclusion that truth-evaluable judgments (i.e., those that can be true or false) have no place in its operations. It is very hard to see, for example, how a mere feeling, absent any truth-evaluable judgments, could count literally as the emotion of guilt; guilt necessarily involves thoughts along the lines of “I have transgressed.” Let us assume then, that the moral faculty produces truth-evaluable judgments.

However, this is not to say that the evolutionary function of the moral faculty is to track the truth. The evolutionary function of a trait is the reason that it was selected for; it reflects why it was reproductively useful to our ancestors. In many cases truth is useful. Having a true belief about whether there are lions around was more useful to our ancestors than having a false belief on the matter. Were humans to have evolved a faculty specifically for producing beliefs about the location of lions, then it is highly likely that the faculty would have generally favored the production of true beliefs. In other words, the faculty would have had the function of tracking the truth.

But it doesn’t always work out this way. Sometimes falsehood is useful. A fanciful example: Suppose that believing in the tooth fairy were an effective means of encouraging tooth-brushing. We can imagine that having clean teeth is such a benefit that those ancestors who believed in the tooth fairy out-competed their rivals, and that a faculty encouraging this

---

3 In cases of so-called survivor guilt, the person still has thoughts concerning “transgression,” though perhaps doesn’t really endorse them. The proposal that the survivor’s guilt is literally nothing more than a feeling without any accompanying cognitive content (like a headache) doesn’t withstand scrutiny.
belief gradually emerged as a prewired human adaptation. In this case, while the faculty’s outputs remain truth-evaluable entities (i.e., beliefs), it is not a truth-tracking faculty. The reason it evolved is not because beliefs about the tooth fairy gave anyone information about her whereabouts, but rather because they encouraged dental hygiene. In this case, someone’s belief that the tooth fairy exists can be assessed as true or false (hint: it’s false), but it would be a mistake to say that the evolutionary function of this belief is to correspond (even roughly) to reality.

Returning to moral nativism: Of course, we don’t know whether the moral faculty is an adaptation at all, but it is interesting to note that on all of the live versions of the nativist hypothesis, mention of truth-tracking is noticeably absent. Most nativist hypotheses suggest that morality plays a vital role in enhancing social cohesion. An individual who thinks that it is morally repugnant to cheat her comrades is less likely to do so—perhaps even less likely than a person who sees that cheating will harm her own long-term interests—and thus on the assumption that cheating is frequently maladaptive, the moral belief may be selected for. The plausibility of this claim seems independent of whether cheating one’s comrades (or anything else) actually is morally repugnant. Other moral nativists emphasize the role that morality can play in signaling one’s commitment to social projects. Perhaps by loudly proclaiming that cheating is repugnant (and acting accordingly) one advertises oneself to others as a worthy partner for mutually beneficial enterprises like hunting or rearing a family. Here the function of morality is its signaling role; again, whether one’s assertion that cheating is repugnant is actually true is irrelevant to its fulfilling this function well.

Debunking

The term “debunking” is intentionally vague, covering several distinct possibilities which I will here divide into three. The most extreme kind of debunking would show that all moral assertions are untrue. This view is known as the moral “error theory” (see Mackie 1977). The error theorist thinks that while moral concepts (goodness, evil, virtue, etc.) have the linguistic function of picking out real features of the world, the universe just doesn’t contain any such features. Thus any statement that something possesses one of these features (e.g., “Keeping promises is morally good”) is simply untrue. (Analogously, the atheist holds that religious terms like God, sin, afterlife, etc., function to pick out entities and features but that the universe simply doesn’t contain such stuff. Thus, the atheist holds that most or all religious claims are untrue.) A weaker view would be that a certain substantial subset of moral assertions are false. Perhaps, for example, all talk of moral rights is flawed, but talk of moral values and virtues is fine. We can call both the strong and weak versions examples of truth debunking.

To establish a moral error theory would automatically debunk various rival moral theories (i.e., any theory according to which moral judgments are sometimes true). Yet it would be a slightly different enterprise to set out with the objective in mind of refuting particular moral theories. This can be classed as a second kind of debunking: theory debunking—aimed not at upsetting any accepted view concerning which moral judgments are true and which false, but rather aimed at showing that certain interpretations of moral judgments are false. For example, while it is widely accepted that “Keeping promises is
morally good” is true, there is much disagreement concerning what makes it true. One might say that it’s true because God commands promise-keeping, or because it maximizes happiness, or because it follows from the optimally coherent equilibrium of our intuitions, etc. To debunk one of these theories (but not all of them) would probably leave the truth value of the claim “Keeping promises is morally good” intact.4

A very different, third kind of debunking—justification debunking—would show that all moral judgments are unjustified. Precisely what it takes for a judgment to be justified or unjustified is something over which philosophers wrangle, but on all accounts it is different from truth and falsity. If I form beliefs about whether someone is coming to dinner on the basis of rolling dice, then my beliefs are not justified, even if I happen to get lucky and have a true belief. If I form beliefs about when Napoleon died on the basis of books, websites, teachers, etc., then my belief is justified; but it is possible that I have fallen victim to a grand hoax (involving fake websites, etc.), such that my justified belief is actually false.

Justification is a relative affair. One may be justified in believing something while someone else is unjustified in believing that very same thing. One may be unjustified in believing something but later become justified (if one gains evidence); or one may be justified but later become unjustified (if some new countervailing evidence comes to light which one chooses to ignore). Thus this kind of debunking can come in various strengths. A strong view is that everyone is unjustified in holding their moral beliefs; an even stronger view is that this situation is permanent.

So some arguments employ moral nativism to undermine moral truth, some use it to undermine specific moral theories, and some use it to undermine moral justification. How might such arguments run?

**Evolutionary debunking of moral truth**

It has sometimes been claimed that one cannot show a belief to be false on the basis of discoveries about its origin; the putative error committed in trying to do so has been dubbed “the genetic fallacy.” For a lot of cases this seems reasonable. For example, however ill-advised we may think it to form scientific ideas on the basis of dreams, it seems wrong to conclude that a scientific idea formed in this way must be false. (Friedrich Kekulé famously envisaged the ring-shaped structure of the benzene molecule after dreaming of a snake biting its own tail.) But for certain cases it doesn’t seem reasonable. If Fred believes “My dreams never influence me in any way,” then the discovery that he formed this belief on the basis of his dreams would allow us to infer that the belief is false. This, however, is clearly a carefully constructed example, since the belief in question implies something about its own origins.

---

4 The “theories” that can be attacked in *theory debunking* include both normative and metaethical theories. Normative ethics is the enterprise of building a general theory of moral action that is applicable across all or a large range of cases. An example of a normative theory is hedonic utilitarianism, which holds that one must always act so as to maximize overall happiness. Metaethics is concerned with a number of interrelated theoretical matters, such as the nature of moral properties, the nature of moral language, and the justification of moral judgments. An example of a metaethical theory is moral constructivism, which holds that moral facts exist but are in some manner constituted by our attitudes or practices.
and on the face of it nothing like that goes on with moral beliefs. What then are the prospects of using moral nativism to show that moral judgments are untrue?

On one interpretation, Michael Ruse advocates such an argument (1986, 2006, 2009). In order to discuss Ruse’s case, we must first introduce the concept of moral objectivity—a notion that gets used differently by different philosophers, and one so slippery that some have recommended its elimination. The basic idea is that some facts depend on us and some do not; but only certain kinds of dependence count. Chihuahuas depend on us (we bred them into existence, we keep them alive, etc.), but are clearly objective entities. The value of money, on the other hand, seems to depend on us in a different way. If we all ceased to believe in the value of money, it would lose all value (whereas if we all ceased to believe in chihuahuas, they wouldn’t pop out of existence).

Some philosophers believe that moral facts depend on us (our beliefs, practices, institutions) in a similar way to the value of money. Such a position allows that moral facts exist but denies that they are objective facts. We can call this position “constructivism.” Other moral philosophers think that it would be desirable to vindicate moral objectivity, if possible; they worry that a constructed morality lacks the kind of binding authority which we would like morality to have. If a criminal sees that the wrongness of his crimes is merely a matter of collective opinion, then what reason does he have to refrain (if he can evade punishment)? Some go so far as to think that objectivity is written into our very moral concepts, such that the idea of a “subjective morality” is not even coherent; for them, anything deserving the name “morality” is an objective morality. (Analogy: anything deserving the name “square” is a four-sided square.) Those who think that moral facts exist and are objective may be called “moral realists,” while those that think that moral facts do not exist at all, or exist only subjectively, can be called “anti-realists.”

Ruse endorses moral nativism, arguing that having a faculty that issued moral judgments was adaptive because those judgments strengthened our ancestors’ motivation to cooperate, and did so by seeming to be objectively binding. “The Darwinian argues that morality simply does not work (from a biological perspective) unless we believe that it is objective” (1986: 253). Of course, this doesn’t mean that there are any objective moral facts, only that it was beneficial to our ancestors to believe that there are. Ruse goes on: “[M]orality is a collective illusion foisted upon us by our genes. Note, however, that the illusion lies not in the morality itself, but in its sense of objectivity” (ibid.). At this point it seems that Ruse thinks that only a certain aspect of morality is in error, not morality per se. But why is there any error at all?

Ruse’s answer is that once armed with (his version of) moral nativism there is simply no need to posit objective moral facts in the world. Nativism fully explains why we believe in moral objectivity, and there is no need to appeal to actual objective moral facts to explain any other phenomena. Ruse thus relies on a principle of parsimony, according to which one should not endorse unnecessarily complicated theories. An analogy he often uses concerns the popularity of spiritualism after World War I: In order to explain why a grieving mother believes that her son spoke to her at a séance, all we need appeal to are psychological facts

---

5 A word of warning: How the terms “moral realism” and “moral objectivity” should be used is itself disputed ground.
about her wishful thinking and vulnerable state of mind; there is no need to imagine that she was really in contact with a supernatural being.

One might, however, take issue with the analogy. In the case of the grieving mother, it is clear that in order to imagine her beliefs to be true we would have to posit a whole weird realm of facts that doesn’t fit with our best current science. But this is not obviously so in the moral case. It is easy enough to make “moral objectivity” sound spooky, but it need not be. A utilitarian, for example, identifies moral facts with whatever produces the most happiness; and the question of which action available to a person produces the most happiness is an objective matter (it’s something we could all be wrong about). Therefore it is not clear that accepting the existence of moral objectivity requires believing in categories of things above and beyond those we already accept. Thus it is not obvious that using a principle of parsimony to eliminate moral objectivity from the world is quite as straightforward as Ruse seems to think.

If Ruse aspired to nothing more than showing that moral objectivity is flawed, then we should classify him as a theory debunker (for he would be undermining any theory according to which morality is objective, leaving constructivist theories intact). But sometimes he seems to go further, accepting resources that allow him to go for a stronger conclusion: truth debunking. He writes: “Ethics is subjective, but its meaning is objective” (2006: 22; see also Ruse 2009: 507). If we take this claim seriously—that objectivity is an essential aspect of moral concepts—then showing that there is no such thing as moral objectivity would amount to showing that there are no such things as moral properties at all. (To recycle an earlier analogy: To show that there are no four-sided square shapes in a box is to show that there are no square shapes in the box.) This claim that morality is essentially objective (conceptually speaking) is by no means implausible, though it is certainly contentious. In any case, it is not something that Ruse can just help himself to; if the debunking argument is to get all the way to the error theory then this crucial connecting premise will need a lot of support.

**Evolutionary debunking of ethical theory**

Sharon Street (2006) develops a debunking argument but explicitly doesn’t endorse the aforementioned connecting premise. Her target is unambiguously moral realism, but since she doesn’t hold that morality is essentially objective, she thinks that versions of moral constructivism survive the argument intact. Thus she is no error theorist; she doesn’t think that morality itself is erroneous, but that a certain interpretation of morality (the realist’s) is false. However, if one agreed with the debunking part of Street’s argument, but disagreed with her tolerance of constructivism (because one thought morality essentially objective, say), then a moral error theory would beckon.

Street argues that moral nativism presents the moral realist with a dilemma. Either (i) there is no causal connection between the putative realm of moral facts and the moral judgments that we have evolved to make, or (ii) there is a connection—namely, that our moral judgments are in some manner responses to the features of that realm. Either horn of the dilemma, Street thinks, is unacceptable. The first horn seems to imply the unpalatable result that either our moral beliefs spectacularly fail to match the moral facts or their matching is a massive coincidence. The problem with the second horn is that, as we saw
earlier, all serious versions of the nativist hypothesis appear to endorse a non-truth-tracking account of the evolutionary function of the moral faculty.

But even if moral judgments do not have the evolutionary function to track anything, they might nevertheless be reliably linked with the presence of some relevant feature of the world. For example, suppose that the adaptational function of the human blushing response is not to track the presence of anything, but rather to signal to others one’s embarrassment over a social faux pas (thus advertising one’s allegiance to social norms). Nevertheless, the blushing response presumably will correlate reliably with cases in which people violate social norms. Something similar could be happening in morality. Moral judgments may not have the evolutionary function to track anything, but nevertheless negative moral judgments (e.g., about what is bad) might correlate with actions that cause harm (say). It could then be argued (as indeed a certain kind of utilitarian does argue) that harm just is moral badness; thus negative moral judgments tend to occur in the presence of badness, after all.

An important part of Street’s response to this is to invoke a principle of parsimony: “The tracking account obviously posits something extra…, namely independent evaluative truths” (2006: 129). Presumably she will say the same thing regarding any account that allows that moral judgments correlate reliably with independent evaluative facts (even though their evolutionary function is not to track them). But we have already seen in the discussion of Ruse that such a principle of parsimony is not to be used hastily. The “independent evaluative truths” that Street speaks of may very well be comprised of kinds of things (such as facts about whether a given action causes harm) that all parties to these debates already accept.

Evolutionary debunking of moral justification

A third kind of genealogical debunking, which I have advocated myself (Joyce 2006), is weaker than the other forms in that it doesn’t purport to show anything false but rather aims to undermine the justification of moral claims. The argument is best approached via an imaginary thought experiment.

Suppose there were such things as different kinds of belief pill, each of which (somehow) caused one to start believing something (presumably also causing amnesia about having taken the pill). Pick one of your beliefs at random; say, your belief that Napoleon lost the battle of Waterloo. Imagine being unexpectedly given indubitable evidence that the only reason you believe this is that someone slipped you a Napoleon-lost-Waterloo belief pill last week. We’ll suppose that this knowledge breaks the spell of the pill, so you now have control over what choices lie before you: You are free to carry on believing that Napoleon was defeated, or you can disbelieve it, or take some other kind of attitude. What should you do?

It would seem wrong to just carry on believing it without concern; but it would also seem wrong to assume that the belief is false. I suspect you’d be rather confused about what to believe, and should probably suspend judgment about whether Napoleon lost Waterloo (just

---

6 Other evolutionary theory debunkers include Joshua Greene (2003) and Peter Singer (2005), both of whom argue that nativism debunks certain non-consequentialist normative theories. (See footnote 4.)

7 It may cause some confusion that in other contexts I have argued for a full-blown moral error theory (Joyce 2001), but not on evolutionary grounds.
as you currently suspend judgment about whether there is an odd or even number of footprints on the moon). Of course, you’re not necessarily stuck in that position; you could do some research which could provide concrete grounds for reinstating your belief that Napoleon lost Waterloo. Or maybe you’re happy to live with the indecision. (Even if the data about the footprints on the moon were available, you might just not care.)

Things would stand differently if you were warranted in thinking that the person who slipped you the pill was benign, in that he or she tended to deliberately give people pills that would produce true beliefs. That would be grounds for thinking that appropriate causal links are intact between the facts (Napoleon’s defeat at Waterloo) and your beliefs on the matter. But absent evidence for any such thought, the discovery that you’ve taken a belief pill appears to remove justification from your belief, because it’s the discovery that the belief was formed by a process that has no sensitivity to the relevant facts. (Whether it removes justification or reveals that the belief was unjustified all along depends on which general theory of justification is true—a matter we cannot discuss here.)

Might moral nativism be analogous to belief pills? One disanalogy is that (as we saw earlier) on many versions of moral nativism evolution doesn’t hardwire whole beliefs, but rather provides a moral faculty of concepts, such as right and wrong. But this is more of a distraction than a disanalogy, because the belief-pill example can be modified accordingly. Suppose instead of belief pills we had concept pills, such that the only reasons you have the concept C is that you were slipped a pill. Your concept C now figures in whole beliefs, such as “X is C” and “Y is not C,” and so forth. By what process this comes about is something we can leave unspecified; the point is that had you never been given the pill you’d have no beliefs involving C at all. It seems that, as before, the discovery of the pill origins of C undermines the justification of all beliefs essentially involving it. Might the discovery that moral nativism is true similarly undermine the justification of our moral judgments?

One thing that’s clear is that if the best versions of nativism show the moral faculty to be truth-tracking then this result won’t follow. That would be analogous to having grounds for thinking that the person who slipped you the belief pill was benign. This shows why this debunking argument isn’t going to generalize to any evolved psychological faculty designed to produce certain kinds of judgment. Perhaps humans have an evolved faculty for doing simple arithmetic; yet the only plausible account of why it might have been beneficial to our ancestors to have these beliefs (like 3+1=4) is that they are true. When fleeing four lions and seeing three quit the chase, having a false belief about whether there are any lions left is likely to be harmful. As we have seen, however, the corresponding supposition is doubtful in the moral case: Moral judgments might fulfill their adaptational function (e.g., motivating cooperation) even if such judgments are false.

But might the moral faculty, while not designed for truth-tracking, nevertheless produce judgments that reliably correlate with the moral facts? Here, I think, there is nothing to exclude a positive answer out of hand, but the burden really lies on the opponent of debunking to confirm the existence of this putative correlation. It’s a bit like discovering that your beliefs about Napoleon are the product of a pill but choosing to maintain them on the grounds that the pill-pushers might be benign. That’s not good enough to provide your belief with justification; you really need some positive reason for thinking that he or she is benign. In the moral case, the mere fact that moral judgments might correlate reliably with moral
facts doesn’t provide one iota of justification for those judgments; it’s up to the opponent of debunking to provide a plausible concrete theory.

One possible reply to this argument is to adopt a theory of justification known as reliabilism, according to which a belief is justified if it is the product of a reliable process of formation—even if the believer is unaware of that process. A reliabilist might object to the preceding argument by saying that so long as our moral judgments reliably correlate with the moral facts (regardless of whether it is their evolutionary function to track them; regardless of whether we have access to a plausible concrete explanation of the relation), then our moral judgments are justified.

However, it seems to me that this just pushes the real problem back a level. The kind of debunker under consideration started out arguing that although our moral judgments might be true, they lack justification. The reliabilist responds by providing grounds for thinking that they might be justified after all. But in the absence of solid evidence that the relevant belief-forming process is reliable, the debunker can retort that although moral judgments might be justified, we have no grounds for thinking that they are justified. (And never mind that there are serious objections to reliabilism.) In the context of an argument about moral epistemology, that seems a significant skeptical conclusion.

Conclusion

No evolutionary discovery lands a knock-out blow to morality; all such findings require supplementation with substantive philosophical argument. In the case of Ruse and Street, the crucial use of a principle of parsimony raises delicate metaphysical issues concerning the nature of moral properties. In the case of the evolutionary debunking of moral justification, things are complicated in different ways. First, whether the conclusion of the argument is (A) that moral judgments are unjustified or (B) that we have no evidence that moral judgments are justified, depends on what kind of theory of justification is preferred. Either way, the argument doesn’t purport to show that no justification can be provided; it represents a challenge for the moralist to do so. The debunker about justification argues that confirmation of a hypothesis about the origins of moral judgment—one that doesn’t imply or presuppose that any such judgments are true—would shift a burden of proof onto the moralist’s shoulders. It is open for the moralist to meet the challenge by retorting that while the moral faculty may not be evolutionarily designed for truth-tracking, it nevertheless produces judgments that reliably correlate with the moral facts. At this point, the moralist must advocate an account of the moral facts and the debunker may criticize the adequacy of this account. Although this may seem like a disappointing (though all too familiar) kind of philosophical impasse, the debunker may continue to insist that the presence of non-truth-tracking moral nativism places the burden of proof on the moralist to win the argument.

References