The origins of moral judgment

Richard Joyce

Penultimate version of paper appearing in *Behaviour* 151 (2014): 261-278.
[Reprinted in F. de Waal et al. (eds.), *Evolved Morality: The Biology and Philosophy of Human Conscience* (Brill, 2014): 125-142]

Is human morality a biological adaptation? And, if so, should this fact have any substantial impact on the ethical inquiry of how we should live our lives? In this paper I will address both these questions, though will not attempt definitively to answer either. Regarding the former, my goal is to clarify the question and identify some serious challenges that arise for any attempt to settle the matter one way or the other. Regarding the latter, my ambitions here are restricted to some brief critical comments on one recent attempt to answer the question in the affirmative.

Let us start with Darwin:

> I fully subscribe to the judgment of those writers who maintain that of all the differences between man and the lower animals, the moral sense or conscience is by far the most important. … [A]ny animal whatever, endowed with well-marked social instincts, the parental and filial affections being here included, would inevitably acquire a moral sense or conscience, as soon as its intellectual powers had become as well, or nearly as well developed, as in man. (Darwin 1879/2004: 120-1)

There are several features of this passage worth highlighting. First, the trait that is under discussion is described as "the moral sense or conscience," which, it seems safe to claim, is a faculty that produces moral *judgments*. Darwin is not here wondering whether being morally *good* is the product of evolution, but rather whether the capacity to make self-directed moral judgments is the product of evolution. A moment's reflection on the myriad of ways in which morally appalling behavior may be motivated by a sense of moral duty should suffice to illuminate the distinction.

The second conspicuous feature of the passage is that Darwin sees the moral sense as emerging (inevitably) from other traits: "social instincts" combined with "intellectual powers." The latter powers he goes on to mention are memory, language, and habit. This raises the possibility that Darwin does not see the moral sense as a discrete psychological adaptation but rather as a byproduct of other evolved traits. In fact, he appears wisely to steer clear of adjudicating on this matter. When focused on the social instincts generally (rather than the moral sense in particular), he writes that "it is ... impossible to decide in many cases whether certain social instincts have been acquired through natural selection, or are the indirect result of other instincts and faculties" (1879/2004, 130).

Contemporary debate among philosophers (in particular) over whether the human moral sense is an adaptation has not always been so cautious. Several recent authors have developed arguments to the conclusion that human moral judgment is not a discrete adaptation but rather a byproduct of other psychological traits. (Nichols 2005; Prinz 2008; Ayala 2010; Machery & Mallon 2010.) Let us call these people "spandrel theorists" about morality. Others, myself included, have advocated the view that the

human moral sense is a biological adaptation. (Alexander 1987; Irons 1996; Krebs 2005; Dwyer 2006; Joyce 2006; Mikhail 2011.) We'll call these people "moral nativists." My first substantive goal in this paper is to reveal how difficult it is to resolve this matter.

Part 1: Adaptations versus spandrels

The spandrel theorist proceeds by offering "non-moral ingredients"—themselves quite possibly adaptations—which are sufficient to explain the emergence of moral judgment. We have seen Darwin mention such things as language use, social instincts, and memory. Francisco Ayala emphasizes "(i) the ability to anticipate the consequences of one's own actions; (ii) the ability to make value judgments; and (iii) the ability to choose between alternative courses of action" (2010: 9015). Jesse Prinz considers such non-moral ingredients as meta-emotions, perspective taking, and the capacity for abstraction (2008; 2014). Here I will take as my exemplar the view of Shaun Nichols (2005), but the general point I shall make could be leveled at any of the aforementioned (and, indeed, against any spandrel theorist).

The two non-moral ingredients that Nichols focuses on are a capacity to use non-hypothetical imperatives[1] and an affective mechanism that responds to others' suffering. He writes that:

> ... both of the mechanisms that I've suggested contribute to moral judgment might well be adaptations. However, it is distinctly less plausible that the capacity for core moral judgment itself is an adaptation. It's more likely that core moral judgment emerges as a kind of byproduct of (*inter alia*) the innate affective and innate rule comprehension mechanisms. (2005: 369)

An obvious way of critically assessing Nichols' claim would be to question whether these two mechanisms, working in tandem, really are sufficient to explain moral judgment.[2] This would involve describing the two mechanisms highlighted by Nichols in much more detail, searching for empirical evidence (e.g., can an individual have one of these mechanisms impaired and yet still make moral judgments?), and so forth. But the question I want to ask is much more general: What determines whether a trait (i) is a byproduct of other mechanisms x, y, and z or (ii) is an adaptation dependent upon pre-adaptational sub-mechanisms x, y and z? Answering this question in the abstract is fairly straightforward, but having a procedure for empirically determining whether a trait is one or the other is considerably more difficult. Let me explain.

---

[1]A hypothetical imperative (e.g., "Go to bed now") recommends that the addressee pursue a certain means in order to achieve one of his/her ends (e.g., to get a good night's sleep). If it turns out that s/he lacks that end, then the imperative is withdrawn. A non-hypothetical imperative demands an action irrespective of the addressee's ends. For example, the imperative "Don't slaughter innocents" is not withdrawn upon discovery that the addressee loves slaughtering innocents, won't get caught, and doesn't give a fig for morality. Moral imperatives are a subset of non-hypothetical imperatives. Non-moral non-hypothetical imperatives include etiquette, road regulations, rules of games and sports, and the norms of institutions generally.

[2]For the sake of simplicity I'm ignoring Nichols' sensible "inter alia" in the previous quote.

No psychological faculty for producing a species of judgment is going to exist as a monolithic entity that takes inputs and magically produces outputs; all such faculties will depend on the operation of numerous psychological sub-mechanisms, which in turn depend on sub-sub-mechanisms, etc. Suppose that Nichols is correct that the two mechanisms he highlights are indeed sufficient to explain the phenomenon of moral judgment. One interpretation—the one Nichols favors—is that the capacity for moral judgment is a byproduct of the operation of these two mechanisms. But a second hypothesis is always available: that the capacity for moral judgment is a distinct adaptation of which these are two sub-mechanisms. The second hypothesis is true if (and only if) the manner in which these two mechanisms interact has been at all modified by natural selection because their interaction has some impact on reproductive fitness. Let us suppose first of all that these two mechanisms evolved for their own evolutionary purposes. But in certain circumstances they interacted, in such a way that the trait of moral judgment emerged as a byproduct. Suppose further, however, that this new trait (moral judgment) had some reproductive relevance, such that the process of natural selection began to "tinker"—perhaps strengthening the interaction of the two mechanisms in some circumstances, dampening it in others. If this has occurred, then the capacity for moral judgment is no longer a mere "byproduct" but rather an adaptation in its own right. (Of course, one can still maintain that it originally appeared as a byproduct, but this is true of virtually everything that counts as an adaptation.[3])

In sum, spandrel theorists about morality seem to think that it suffices to establish their view if they offer non-moral ingredients adequate to account for moral judgment. But the consideration just raised indicates that this matter is not so straightforward, for any spandrel hypothesis can be interpreted instead as a description of the sub-mechanisms of the nativist moral sense. (And if the ingredients mentioned are indeed adequate to explain moral judgment, then so much the better for the resulting nativist hypothesis.)

But how would one distinguish empirically between these two hypotheses? The difference between an adaptation and a byproduct cannot be discerned by consulting intrinsic features of the organism, no matter in what detail. Consider Stephen Jay Gould's architectural analogy that originally provided the term "spandrel" (Gould & Lewontin 1979). Renaissance architects faced the design challenge of mounting a dome upon a circle of arches; when this is accomplished, the spaces between the arches and dome produce roughly triangular areas of wall: spandrels. These areas of wall are not design features—they are byproducts of the design features. Yet one could not discern this by examining the intrinsic structural features of the building; one must know something about the purposes of the architects. It is, after all, conceivable that an architect may have a direct interest in creating spandrels, in which case the dome and arches would be byproducts. The resulting church would be intrinsically indistinguishable from the ordinary church for which the spandrels are byproducts.

---

[3]See Dennett 1995: 281.

In the same way, in order to know whether a trait is an adaptation as opposed to a byproduct one must understand something of the intentions of the architect—in this case, the forces of natural selection that existed during the period of the trait's emergence. Lacking, as we usually do, concrete evidence of the subtle evolutionary pressures operating upon our ancestors, our epistemic access to this information will always depend to some extent on intelligent inference. Consider, for example, Nichols' contention that the capacity to use non-hypothetical imperatives is an adaptation whereas the capacity to use moral imperatives is a byproduct. An alternative view is that the capacity to use moral imperatives is the adaptation while the more general capacity to use non-hypothetical imperatives is the byproduct. One could not decide between these hypotheses simply by examining the human organism; rather, the decision would have to involve comparing the plausibility of two conjectural hypotheses. On the one hand, one might hypothesize that the ancestral environment contained adaptive problems for which the specific capacity to use *moral judgments* would be a reasonable solution. Alternatively, one might hypothesize that the ancestral environment contained adaptive problems for which the specific capacity to use *non-hypothetical imperatives* would be a reasonable solution. In either case, the adaptive problems would need to be described in a manner supported by available evidence. To the extent that the former hypothesis turned out to be more plausible than the latter, moral nativism would be supported. But if the latter were more plausible than the former, then support would be provided for the spandrel view. A troubling possibility, of course, is that we may very well find ourselves lacking solid ground for favoring either kind of hypothesis over the other, in which case we'd lack ground for claiming with confidence which trait is the adaptation and which the byproduct. One can see now, perhaps, the wisdom of Darwin's quietism on this matter.

Part 2: What is the trait under investigation?

I have been outlining one way in which the dispute between the moral nativist and the spandrel theorist is likely to run aground. However, it might reasonably be responded that this problem is of little consequence, since the contrast that is of greater theoretical interest is whether the capacity to make moral judgments is the product of evolutionary forces (whether an adaptation or a byproduct) or is an acquired ability. Frans de Waal calls the latter position "veneer theory": the view that morality, along with cooperative and altruistic tendencies in general, is "a cultural overlay, a thin veneer hiding an otherwise selfish and brutish nature" (de Waal 2006: 6). I doubt that many people nowadays endorse the veneer theory; that humans have been designed by natural selection to be gregarious and cooperative seems beyond reasonable doubt. The devil lies in the details of *how* we are gregarious and cooperative. Note that declaring that we are by nature gregarious and cooperative is not to declare in favor of moral nativism, for it remains entirely possible that our social nature consists of biologically entrenched tendencies toward altruism, sympathy, love, and so forth,

while the capacity to make moral judgments is an acquired and relatively recent cultural characteristic.

This observation, however, focuses attention on the knotty question that lies at the heart of these debates: What is a moral judgment? There is little to be gained in arguing over whether a trait is an adaptation or a spandrel, innate or acquired, if we don't have a firm handle on the nature of the trait under investigation. It is a great inconvenience to these debates that the concept *moral judgment* is a slippery and highly contested idea even among those who are supposed to be experts on the topic—namely, metaethicists.

In order to approach this problem, let us pause to compare chimpanzee sociality with human sociality. De Waal has often claimed that chimpanzee life contains some of the "building blocks" of morality (1992; 2006). He focuses on such things as reciprocity, consolation behavior, inequity aversion, empathy, and the following of rules of conduct reinforced by others. At the same time, de Waal is positive that chimpanzees do not make moral judgments (1996: 209).[4] This raises the question of what additional building blocks need be added, or how the building blocks need be rearranged, in order to create something deserving of the name "a moral sense."The fact that the answer is not at all clear problematizes the whole dialectic concerning the evolution of morality. In what follows I will attempt to say something useful on the matter.

A striking feature of the chimpanzee building blocks is that they seem to require emotional arousal. A deviation from a social rule in chimpanzee society receives a negative response only because those giving the response get angry. Consolation behavior is provided only by those in whom sympathy has been stirred. A reciprocal act (grooming behavior, say) occurs because the reciprocator feels friendly and caring toward the recipient (or, perhaps, feels fearful of the reprisal that non-reciprocation might bring). What chimpanzees seem to lack is a psychological apparatus that could motivate such behaviors in the absence of emotional arousal. In humans, by contrast, a deviation from a social rule might receive a negative response because those giving the response judge that it is *deserved*; consolation behavior might be provided by those who considers it *right* to do so; reciprocation might be offered because one judges oneself to have a *duty* to repay a debt; and so forth.

There are many who would claim that because the prominent "building blocks of morality" seem to be affective phenomena, the fully-fledged moral faculty must also be an affective mechanism.[5] One might argue, for example, that what humans have (which other primates lack) is the capacity to have meta-conations. Perhaps if an individual not only dislikes a certain behavior, but likes the fact that she dislikes it (and perhaps also dislikes anyone who fails to dislike it, and likes anyone who does

[4]See also Boehm 2012: 113-131.

[5]Here I am using the terms "affective," "noncognitive," and "conative" synonymously. I am not shunning the term "emotional," but am treating it with care, for emotions—at least many of them—are mixtures of affective and cognitive components. (For this reason, I do not consider Christopher Boehm's claim that the internalization of norms requires that one "connect with these rules *emotionally*" (2012: 114) to be necessarily at odds with the cognitivist line I push in this paper.)

dislike it) then we may speak of her "morally disapproving" of the behavior. Perhaps if one's dislike of another's action prompts not only anger, but a disposition to feel anger at those who do not also feel anger at the action, then we may speak of one's judging that the anger is *merited*. (See Blackburn 1998: 9-13; Prinz 2007: 113-115; Mameli, forthcoming.)

I find this line of reasoning unpersuasive. The building blocks of morality found in chimpanzees (and, by presumption, our ancestors) may well be affective phenomena, but it is entirely possible that the crucial modification of these building blocks in the human lineage was the addition of certain cognitive aptitudes. After all, generally speaking, the explosion of cognitive abilities is surely the most striking aspect of recent human evolution. Moreover, it is far from obvious, just on conceptual grounds, that one can really build a moral judgment from these affective ingredients alone. The natural way of assessing the claim is to examine potential counterexamples, of which there are two types. First, can we imagine these noncognitive capacities being deployed without a moral judgment occurring? Second, can we imagine a moral judgment occurring without these noncognitive capacities being deployed? I'm inclined to think that the answer to both questions is "Yes."

Suppose I am strolling among a group of normally docile animals when one bites me aggressively. Naturally, I dislike this; perhaps I smack the animal on the nose in order to make it release me. Perhaps, moreover, I judge that it's important that these animals don't form aggressive habits (maybe my children often play in their vicinity), so I would wish to see others smack the animal if bitten. Perhaps I go so far as to dislike anyone who wouldn't smack the animal if bitten. Yet these emotions and meta-emotions do not appear to amount to a moral judgment of the animal's behavior. It doesn't seem that I judge that the animal *deserves* to be smacked; indeed, I do not treat the animal's behavior as a *transgression* at all. I do not *disapprove* of its aggressive behavior; I simply dislike it in an elaborate way.

The reason we do not make moral judgments concerning animals is because they lack a certain kind of agency that we think of as a prerequisite for moral assessment. (It doesn't matter to our current purposes what the nature of this agency is.) Taking this into account, one might respond that the emotions that form the basis of moral judgment are a kind that can be coherently deployed only toward creatures that fulfill these criteria of agency. The "dislike" felt toward a violent animal just isn't the right sort of affective state to begin with (the response goes); perhaps talk of "disapproval" would be more apt than talk of "dislike."

The problem with this response is that *disapproval* is not a mere noncognitive response; it is a mental state permeated with conceptual content. Disapproval requires a concomitant judgment that the object of assessment has transgressed in a manner that warrants some sort of punitive response (if only treating with coolness). One therefore cannot appeal to disapproval as the basic noncognitive state to explain *meriting* (for example). The same problem would emerge if one tried to account for moral judgment in terms of the emotion of *guilt*—for this is an emotion with conceptually rich components (see Joyce 2006: 101-104). I therefore doubt that one can build moral judgments out of affective phenomena alone.

Not only are purely noncognitive building blocks insufficient for moral judgment, but they appear to be unnecessary. Consider a moral judgment voiced in circumstances of emotional fatigue. Perhaps one has just been exposed to a sequence of similar moral scenarios and one's capacity for emotional arousal has ebbed. (Maybe one is ticking the hundredth box on a psychology experiment designed to ascertain subjects' moral intuitions on a range of cases.) Or perhaps one is simply distracted. All too often those who claim that emotional arousal is necessary for moral judgment focus on extreme cases: our disgust at pedophilia, our horror at the thought of the trains discharging their passengers at Auschwitz. Mundane moral judgments—like thinking that the gold medalist *deserved* her win, or that a person's ownership of his shoes grants him certain *rights* to that footwear—don't get a look in. One can claim, of course, that even for these mundane cases emotional arousal is *possible* (imagine someone having his shoes stolen; picture his outrage; visualize his suffering as he walks home barefoot through the snow), but emotional arousal to *anything* is possible.

This is one problem with Prinz's view that even if someone making a moral judgment isn't emotionally aroused he or she is at least *disposed* to become emotionally aroused (2007: 84ff). Even if one could specify precisely what kind of emotion is relevant, there is simply no such thing as the disposition to have that emotion (occurrently) *period*; it must be a disposition to have that emotion (occurrently) *in such-and-such circumstances*. But while one may identify circumstances under which an individual might become emotionally aroused at the thought of someone's enjoying rights over his own shoes, so too one may think of circumstances under which an individual might become emotionally aroused at the thought that gold has atomic number 79 (or any other matter). It may be possible to find a principled distinction between such cases, but to my knowledge none has ever been articulated.

Highlighting the cognitive achievements inherent in moral judgment is not intended to exclude the affective components. As we have seen, affective mechanisms were probably central to the emergence of moral judgment—at least as pre-adaptations—and all the evidence indicates that emotions continue to play a central role in human moral life. (See Haidt 2001; Greene & Haidt 2002; Wheatley & Haidt 2005; Valdesolo & DeSteno 2006; Small & Lerner 2008; Horberg et al. 2011.) None of this, however, undermines the hypothesis that certain cognitive capacities are necessary for moral judgment, and that these capacities were the key development—the crucial additional building blocks—in the emergence of human morality.

The cognitive capacities I have in mind might be described as those necessary for the "moralization" of affective states. Consider the elaborate cluster of conations and meta-conations described earlier, which I doubted were sufficient for a moral judgment. What the cluster seemed unable to account for were ideas like *disapproval*, *transgression*, and *merited reaction* (i.e., *desert*). Without these, the fully-blown

moral conceptions of *obligation*, *prohibition* (and thus *permission*[6]) are unavailable. Without the concept of obligation, there is no possibility of judging anyone to have a *right*, and without rights there can be no idea of *ownership* (only the idea of possession).

The chimpanzee brain lacks the mechanisms necessary to access this conceptual framework probably as surely as the human brain lacks the mechanisms for navigating the world using echo-location. Even if we could ramp up the chimpanzee's capacity for meta-conations (allowing them, say, the capacity to get angry at those who don't get angry at anyone who fails to get angry at someone who does so-and-so), we still would not thereby grant them the capability for judging a punitive response to be *deserved*. Nor would we grant them this capability if we could boost their abilities to discriminate factual data in their environment (allowing them, say, the capacity to infer that if X desires Y's welfare, and X believes that Z will get angry at Y if Y performs action ϕ, then X will want Y to refrain from ϕing). It cannot be the mere "abstract" quality of moral concepts that places them beyond the chimpanzee's grasp, for in other ways chimpanzees wield abstract concepts smoothly.[7]De Waal rightly claims that humans have a greater capacity to internalize norms than other primates (Flack & de Waal 2001: 23; see also Boehm 2012: 113-131), but the puzzle remains: What mechanisms does a brain need in order to have the capacity to internalize a norm? It is natural to answer by saying something about the fear of punishment becoming assimilated, such that the individual self-regulates behavior by administering his/her own emotional punishment system. But the puzzle reiterates. To *fear* punishment is not to have internalized a norm (since one can fear punishment for a crime that one does not believe really is a crime); for internalization, one must believe that punishment would be *merited* and thus be disposed to dispense a kind of punitive self-reproach to oneself even in the absence of witnesses. But what accounts for this concept of "meriting"? Again I would answer that it is challenging to see how a creature could form such a thought using only purely conative and more general data-processing mechanisms (no matter how elaborate). I propose that norm internalization requires cognitive resources dedicated to normative thinking in particular.

The suggested hypothesis is that the human brain comes prepared to produce normative cognitions in a similar way that it comes prepared to encounter faces, other minds, and linguistic stimuli. This is not to say that it comes prepared for any particular normative system—that is, one with a particular normative content. The conspicuous phenomenon of moral disagreement demonstrates that moral content is learned and to some extent flexible, in the same way that the abundance of natural languages demonstrates that languages are learned and to some extent flexible. And to

---

[6] If one lacks the concepts of *obligation* and *prohibition*, then one lacks the concept of *permission*. Contra Camus' claim that "if we can assert no value whatsoever, everything is permissible" (1951), if there are no moral values then *nothing* is morally permissible.

[7]Consider a chimpanzee's postponing a vengeful act against a rival until a good opportunity arises. Perhaps we grant it deliberations about plans it will execute "later"—but *later* is an abstract concept. Or consider the way that chimpanzees can play "spot-the-odd-one-out"-type games (Garcha & Ettlinger 1979). *Sameness* and *difference* are abstract concepts.

restate an earlier point: The hypothesis that the human brain comes prepared for normative thinking is a more general proposition than the moral nativist hypothesis. Perhaps Nichols is correct that non-hypothetical normative thinking is an adaptation while specifically moral thinking is a spin-off capacity. Or perhaps it's the other way round. Deciding whether something is an adaptation involves a large dose of inference and speculation concerning what we suppose were the relevant adaptive problems placing pressure upon our ancestors in the distant past.

Insisting on the cognitive components of moral judgment still leaves much undecided about the exact nature of these judgments. Some have argued, for example, that one characteristic of moral judgments is a particular kind of practical authority: moral rules (unlike those of most other normative systems) are those with which one *must* comply whether one likes it or not. Others have doubted this, allowing that a person with sufficiently aberrant goals and desires (and appropriately situated) may well have no reason to care about moral imperatives.[8] The cognitive quality of moral judgment is consistent with either view; it is silent of the subject. A disquieting possibility is that the notion of *moral judgment* is in fact not as determinate on this matter (or on other matters) as we generally presuppose. Perhaps there is simply no fact of the matter as to whether moral rules have or lack this authoritative quality. Certainly people seem to generally imbue their moral prescriptions with this kind of strong authority, so maybe having a theory that provides this authority is a theoretical desideratum. But perhaps this authority is not an indispensable component of morality; maybe if we can make no sense of this authority and have to settle for a normative system lacking it, the system would still deserve the name "morality." One way of diagnosing this situation would be to say that *strictly speaking* morality has this authoritative quality, but *loosely speaking* it need not.

Something similar has been said about language by Marc Hauser, Noam Chomsky, and W. Tecumseh Fitch, who argue that one can speak of language in a broad sense or a narrow sense (2002). The former consists of linguistic capacities that we share with other animals, whereas the latter includes the uniquely human trait of linguistic recursion. There is no answer to the question of which idea captures what is "*really*" language; our vernacular concept of *language* is simply not so fine-grained as to license one answer while excluding the other. Faced with the query of whether vervet monkeys, say, have a language, the only sensible answer is "In one sense yes and in one sense no."

The same may be true of morality. The vernacular notion of a moral judgment may simply be indeterminate in various respects, allowing of a variety of precisifications, with no particular one commanding acceptance. This raises the possibility that the capacity to make moral judgments construed in one sense may be an adaptation, while the capacity to make moral judgments construed in another (equally legitimate) sense is not. One might even go so far as to say that chimpanzees satisfy the criteria for

[8]Philosophers who advocate the thesis that moral prescriptions enjoy some kind of special authority include Immanuel Kant, J. L. Mackie, Michael Smith, and Christine Korsgaard. Those who allow the possibility that one may have no reason to act morally include David Hume, Philippa Foot, David Brink, and Peter Railton.

making moral judgments *very loosely construed*—though I would urge against liberality taken so far. A less excessive and not implausible possibility is that on some broad construal of what a moral judgment is, the capacity to make them is a deeply entrenched part of evolved human psychology, while on a stricter construal the capacity is a recent cultural overlay: a veneer.[9]

Part 3: Implications of cognitivism

Whether on any reasonable precisification of the concept *moral judgment* the cognitive element is necessary is something on which I won't attempt to adjudicate (though earlier arguments reveal my inclination to think so). Certainly I maintain that this element is necessary at least for moral judgments *strictly construed*. I will close by considering some of the implications of moral judgments being cognitive in nature.

To claim that moral judgments essentially involve a cognitive component is basically to claim that they essentially involve *beliefs*. For example, if one holds (as one should) that a judgment that a punitive response is *deserved* must involve something more than just elaborate conative attitudes, then one holds that it involves (possibly inter alia) *the belief* that the punitive response is deserved. Once beliefs are in the picture, then certain distinctive ways of assessing moral judgments must be permitted, meaning that human morality can be interrogated in ways that, say, chimpanzee social systems cannot be. A chimpanzee group may enforce a rule that is in fact practically sub-optimal; so too may a human group. An individual chimpanzee may become affectively aroused at another in a way that harms its own interests (or furthers its own interests); so too may a human individual. But the fact that the human moral faculty involves normative *beliefs* means that human moral judgments can be evaluated in additional ways for which evaluating the chimpanzee response would make no sense. Beliefs can be assessed for truth or falsity in a way that purely noncognitive states cannot be. Beliefs can be assessed for justification or non-justification in a way that purely noncognitive states cannot be. (This is not to claim that all talk of justification is misplaced for noncognitive attitudes, but that it must be of a very different type.[10]) Therefore a human moral response may be probed with the questions "Is it true?" and "Is it justified?" And if one can do this for a token judgment, there seems nothing to stop one posing these questions on a grand philosophical scale: inquiring of human moral judgments in general "Are they true?" and "Are they justified?"

---

[9]This, clearly, would not be what de Waal means by "veneer theory," since, on the view just described, morality (strictly construed) would be a veneer over a core of social and altruistic tendencies, not (as de Waal envisages) over a core of nasty asocial selfishness.

[10] A basic distinction here is between instrumental justification and epistemic justification. Something is instrumentally justified if it furthers one's ends. Mary's belief that the famine in Africa is really not so bad may be instrumentally justified (for her) if her knowing the truth would cast her into a depression. A belief is epistemically justified if it is formed in a way that is sensitive to the evidence. Mary's belief that the famine in Africa is not so bad, though it makes her happier, is epistemically unjustified if she has been exposed to sufficient evidence of its falsehood (which she ignores). When I say that noncognitive attitudes cannot be assessed as justified or unjustified, I mean *epistemic* justification.

Some may say that asking these epistemological questions of morality is somehow off the mark—that the more important question regarding human morality *is* the one that can also be asked of chimpanzee social regulation: namely, "Does it work?" I find that I have nothing to say about which kind of question is more urgent or more interesting; it's a matter of what one's theoretical concerns are. I do think, however, that the epistemological questions can be legitimately asked of any belief, and it is the job of the metaethicist to press these questions hard regarding moral beliefs.

My approach to these matters puts me somewhat at odds with that of Philip Kitcher (2011; this issue). Kitcher sees moral judgment as having emerged for a purpose, allowing one to speak of its fulfilling its function well or poorly. This in turn allows one to make sense of moral progress, but not in the manner of scientific progress—that is, the attainment of improving approximations of the truth—but in the manner of refining a tool to better accomplish its task. Moral truth, for Kitcher, can enter the picture later: defined derivatively from the account of moral progress, not vice versa.

Kitcher and I agree that a "moralization" of affective attitudes occurred at some point in our ancestry. In this paper I have advocated the view that what allowed this moralization were new building blocks of a cognitive nature: essentially, *beliefs* about behaviors being forbidden, punishments being just, and so forth. Instead of their large-scale cooperative projects being at the mercy of capricious conative states, our ancestors became able to think of cooperation (in certain circumstances) as absolutely required, of defection meriting penalty, etc., which supported a more robust motivation to participate. I'm inclined to think that Kitcher is correct in holding that the purpose of morality is, broadly, to augment social cohesion, but I would place more focus on *how* moral thinking accomplishes this end: by providing people with *beliefs* concerning actions having moral qualities. Kitcher calls the view that moral judgments track pre-existing moral properties "a bad philosophical idea" (this issue: #). He may be correct, and yet exploiting this "bad idea" might be exactly how ordinary human moral thinking actually functions. After all, how, one might ask, does moral thinking augment social cohesion better than altruistic sentiments? Why is the motivation to cooperate often more reliable when governed by thoughts like "It is my duty to help him" than when governed by thoughts like "Gee, I really like him"? The former, it is tempting to answer, gains motivational traction by exploiting the idea (however vaguely) of externally binding rules of conduct—imperatives that are inescapable because they do not depend upon us for their authority—moral truths to which our judgments must conform, not vice versa. (Kitcher refers to the idea of a "transcendent policeman" (this issue: #).) But such ideas will typically accomplish this social role only if they are *believed*.[11] And so long as they are beliefs, we can immediately ask "Are they true?" and "Are they (epistemically) justified?"

It is the job of the philosopher to investigate whether any sense can be made of this idea of "inescapable authority." If it cannot, then human moral beliefs may be systematically false (or, less aggressively: human moral beliefs *strictly construed* may

[11]On other occasions I have explored the idea that merely *thinking* such thoughts, without believing them, might have motivational impact (Joyce 2001; 2005); but here such complications are bracketed off.

be systematically false). The fact that one may nevertheless be able to speak of some moral systems serving their evolutionary function better than others—that is, to speak of moral progress—wouldn't cut against this skepticism. To use a crude analogy: Religion may have evolved to serve some social function, and some religions may do so better than others, but for all this atheism may be true.

Conclusion

In conclusion let me summarize what this paper has attempted via a quick clarification of two potentially misleading pieces of terminology: "naturalization" and "value."

Most of us seek a naturalization of human morality. We want to understand morality as a non-mysterious phenomenon, with a history that possibly stretches deep into our evolutionary past, accessible to empirical scrutiny. Parts 1 and 2 of this paper sought to contribute (modestly) to this goal by drawing attention to some fairly deep challenges for this program. My intention wasn't to scupper the project, but to face up honestly to some difficulties confronting it. But there's another kind of "naturalization" which is a whole new ball game. When metaethicists talk of "moral naturalism" they typically mean the provision of a naturalization of moral *properties*. This very different kind of naturalization was the concern of Part 3.

The former kind of naturalization seeks to understand how *moral judgment* fits into the scientific worldview; the latter kind seeks to understand how *moral goodness* (etc.) fits into the scientific worldview. Obviously, one can be optimistic about the prospects of the former while highly dubious of the latter. Compare, again, the analogous two ways of understanding what it takes to provide a "naturalization of religion": one seeks to place religious practices and belief within a scientific worldview; the other would seek to place God within a scientific worldview.

A matching ambiguity emerges when we talk of "values." It's helpful to bear in mind that "value" is both a verb and a noun. We can investigate what is going on in a human brain when its bearer *values* something; it is far from obvious that doing so contributes anything to our wondering what things *have value*.[12] Of course, one might think that the latter is some sort of function of the former (in the same way that the monetary value of things depends on what pecuniary value we are collectively willing to assign them)—but this is a substantive and controversial position in moral philosophy requiring argumentative support. Many metaethicists (and folk!) think, by contrast, that "value" as a noun is the primary notion, while our valuing activity has the derivative goal of discovering and matching what values exist.

Parts 1 and 2 focused on the origins of moral valuing as an activity: worrying that it will be hard to discern whether the human trait of morally valuing things is an adaptation or a byproduct (Part 1), and concerned that the trait is not, in any case, well-defined (Part 2). Part 3 argued that if moral valuing involves beliefs (as I

[12]Cf. Patricia Churchland's guiding questions: "Where do values come from? How did brains come to care about others?" (2011: 12).

maintain it does), then it is always reasonable to inquire whether these beliefs are true. To do so is to focus on "moral value" as a noun—asking whether the facts that are necessary to render moral beliefs true (facts about which actions are forbidden, which are morally good, and so forth) actually obtain. Though on this occasion I have lacked the time to present any arguments, my notes of pessimism have probably been apparent.

Victoria University of Wellington

References

Alexander R. 1987. *The Biology of Moral Systems*. Hawthorne, NY.: Aldine de Gruyter.

Ayala F. 2010. "The difference of being human: Morality."*Proceedings of the National Academy of Sciences* 107: 9015-9022.

Blackburn, S. 1998. *Ruling Passions*. Oxford: Oxford University Press.

Boehm, C. 2012. *Moral Origins: The Evolution of Virtue, Altruism, and Shame*. NY: Basic Books.

Camus, Albert 1951. *L'Homme révolté [The Rebel]*. Paris: Gallimard.

Churchland, P. 2011. *Braintrust: What Neuroscience Tells Us About Morality*. Princeton: Princeton University Press.

Darwin, C. 1879/2004. *The Descent of Man*. London: Penguin Books.

De Waal, Frans. 1992. "The chimpanzee's sense of social regularity and its relation to the human sense of justice." In R. Masters & M. Gruter (eds.), *The Sense of Justice: Biological Foundations of Law.* Newbury Park, CA: Sage Publications. 241-55.

De Waal, Frans. 1996. *Good Natured: The Origins of Right and Wrong in Primates and Other Animals*. Cambridge, MA: Harvard University Press.

De Waal, Frans. 2006. *Primates and Philosophers*. Princeton: Princeton University Press.

Dennett, D. 1995. *Darwin's Dangerous Idea*. NY: Simon & Schuster.

Dwyer, S. 2006. "How good is the linguistic analogy?" In P. Carruthers, S. Laurence, & S. Stich (eds.), *The Innate Mind, Volume 2: Culture and Cognition*. Oxford: Oxford University Press. 237–55.

Flack, J. & de Waal, F. 2001. "'Any animal whatever': Darwinian building blocks of morality in monkeys and apes." In L. Katz (ed.), *Evolutionary Origins of Morality: Cross-Disciplinary Perspectives*. Thorverton, UK: Imprint Academic. 1-29.

Garcha, H. & Ettlinger, G. 1979. "Object sorting by chimpanzees and monkeys." *Cortex* 15: 213-224.

Gould, S.J. & Lewontin, R.C. 1979. "The spandrels of San Marco and the Panglossion paradigm: A critique of the adaptationist programme." *Proceedings of the Royal Society: Biological Sciences* 205: 581-598.

Greene, J. & Haidt, J. 2002. “How (and where) does moral judgment work?” *Trends in Cognitive Sciences* 6: 517-523.

Haidt, J. 2001. “The emotional dog and its rational tail: A social intuitionist approach to moral judgment.” *Psychological Review* 108: 814-834.

Hauser, M., Chomsky, N., & Fitch, W. T. 2002. “The faculty of language: What is it, who has it, and how did it evolve?” *Science* 298: 1569-1579.

Horberg, E., Oveis, C., & Keltner, D. 2011. “Emotions as moral amplifiers: An appraisal tendency approach to the influences of distinct emotions upon moral judgment.” *Emotion Review* 3: 237-244.

Irons, W. 1996.“Morality as an evolved adaptation.”In J. Hurd (ed.), *Investigating the Biological Foundations of Human Morality*. Lewiston, NY: Edwin Mellen Press. 1-34.

Joyce, R. 2001. *The Myth of Morality*. Cambridge: Cambridge University Press.

Joyce, R. 2005. “Moral fictionalism.” In M. Kalderon (ed.), *Fictionalism in Metaphysics*. Oxford: Oxford University Press. 287-313.

Joyce, R. 2006. *The Evolution of Morality*. Cambridge, MA.: MIT Press.

Kitcher, P. 2011. *The Ethical Project*. Cambridge, MA.: Harvard University Press.

Krebs, D. 2005. “The evolution of morality.” In D. Buss (ed.), *The Handbook of Evolutionary Psychology*. NJ.: John Wiley & Sons. 747-771.

Machery, E. & Mallon, R. 2010.“The evolution of morality.” In J. Doris, G. Harman, S. Nichols, et al. (eds.), *The Moral Psychology Handbook*. Oxford: Oxford University Press. 3-46.

Mameli, M. Forthcoming. “Moral judgment: A hypothesis on its nature and evolution.”

Mikhail, J. 2011. *Elements of Moral Cognition: Rawls’ Linguistic Analogy and the Cognitive Science of Moral and Legal Judgment.* Cambridge: Cambridge University Press.

Nichols, S. 2005. “Innateness and moral psychology.” In P. Carruthers, S. Laurence, & S. Stich (eds.), *The Innate Mind: Structure and Contents*. New York: Oxford University Press. 353-430.

Prinz, J. 2007. *The Emotional Construction of Morals*. Oxford: Oxford University Press.

Prinz, J. 2008. “Is morality innate?” In W. Sinnott-Armstrong (ed.), *Moral Psychology, Volume 1: The Evolution of Morality: Adaptations and Innateness*. Cambridge, MA.: MIT Press. 367-406.

Prinz, J. 2014. “Where do morals come from? A plea for a cultural approach.” In M. Christen, J. Fischer, M. Huppenbauer, C. Tanner, & C. van Schaik (eds.), *Empirically Informed Ethics*. Springer Press. ##-###.

Small, D. & Lerner, J. 2008. “Emotional policy: Personal sadness and anger shape judgments about a welfare case.” *Political Psychology* 29:149-168.

Valdesolo, P. & DeSteno, D. 2006. “Manipulations of emotional context shape moral judgment.” *Psychological Science* 17: 476-477.

Wheatley, T. & Haidt, J. 2005. “Hypnotically induced disgust makes moral judgments more severe.”*Psychological Science* 16: 780-784.