Evolution and the Possibility of Moral Realism

PETER CARRUTHERS[1]
University of Maryland

SCOTT M. JAMES
University of Kentucky

Richard Joyce covers a great deal of ground in his well-informed, insightful, and provocative book (Joyce, 2006), much of which we can agree with. But he also argues that any adequate evolutionary understanding of morality, and of the innate "moral sense" that underlies it, will serve to *undermine morality*. Since we disagree with this claim, we propose to take it as the focus of our commentary. Joyce develops two main arguments, targeted on the *form* and the *content* of morality, respectively. The first is that no adequate evolutionary naturalism about morals can give an adequate account of what he calls the "practical clout" of morality. This is Joyce's term to cover both the *inescapability* and *authority* that (he thinks) form essential components of our beliefs about morality. In which case, whatever might be described by the evolutionary naturalist's account of our moral sense will fail to count as a system of *morality*. It will be too weak and watery for that. Joyce's second argument is that plausible theories of the evolution of our moral sense will fail to line up in the right sort of way with any adequate account of the *content* of our moral beliefs – in such a way, that is, as to give us the required confidence that our moral faculty has evolved to track moral truth. So it is much as if we were to discover that our belief that Napoleon lost the battle of Waterloo had actually been caused by taking some kind of pill, rather than by the facts and/or any sort of sensitivity to the evidence. Once we discover the historical origins of our belief in an episode of pill-taking, our belief that Napoleon lost is undermined, and should be suspended. So, too, Joyce claims, with morality, once we see its evolutionary origins. We will discuss these arguments in turn.

[1] The order of the authors' names is alphabetical. We are grateful to Leland Saunders and Stephen Stich for their comments on an earlier draft.

1 Practical Clout

Joyce recognizes, of course, that he can't insist on practical clout as a *defining* feature of morality. This is because, as he explains, the sort of authority that he has in mind as a property of genuine moral norms would imply that a subject is irrational in ignoring those norms, or at least has powerful reasons to comply with them independent of other goals (p.62). Yet it is hugely controversial to assert that morality and rationality must be connected in this sort of "internalist" way. What Joyce does insist, is that an evolutionary account should deliver something "sufficiently close" to practical clout (pp.200-1). It should explain why the idea of the authority of moral judgment should seem so natural and compelling to many people, as well as explaining how moral judgment (as construed by the evolutionary proposal in question) could have the sort of role in our lives that it does. So far, so good: we agree. But Joyce goes wrong in failing to consider the most plausible kind of account of the architecture of our moral sense, and he commits clear fallacies in the course of his argument. Let us elaborate.

We think (loosely following Sripada and Stich, 2006) that the human moral sense must include at least the following components: (1) A data-base of stored normative beliefs about what must, must not, or may be done. (Some of these might be innate or innately channeled; others will be acquired via cultural learning of various sorts; and yet others might result from individual or collective reasoning.) (2) An inferential system for figuring out which norms apply in a given circumstance, and judging accordingly. (3) A system for generating emotional and motivational reactions in response to the emerging judgments. This third system issues in indignation and punitive motivations in response to a judgment that someone else has done what they mustn't do, and guilt and regret in response to such a judgment where the subject is oneself.[2] (It should be stressed that the resulting motivations are intrinsic, not instrumental.)

In terms of this architecture one can smoothly explain the *phenomenology* of moral clout, at least. If the beliefs in the norms data-base are for the most part not conditional in form (hence specifying what must or must not be done in various circumstances in ways that are independent

[2] Whether there is any need for a separate pro-moral motivation is moot. It may be that anticipatory guilt, felt in response to the thought of doing something that one believes one must not do – or in response to the thought of failing to do what one believes one must – would be sufficient by itself, since guilt is experienced as strongly aversive.

of the goals of the agent), then moral judgments will be applied to agents irrespective of what those agents want. Moreover, if the normative beliefs underlying our moral judgments are deeply embedded ones, then the resulting judgments will strike subjects as obvious truths about the world, in this respect much like the truths of common-sense physics, or truths about one's own past. Hence the seeming *inescapability* of moral requirements is easily explained. And if the motivational side of the system works reliably, then the seeming *authority* of moral requirements can be explained as well. For as soon as agents find themselves making a moral judgment, they will inevitably experience a corresponding motivation, such as indignation or anticipatory guilt. These feelings will not seem in any way "optional". And if they are experienced as bound together with the normative judgment that causes them, it will be natural for subjects who reflect on the matter to come to the view that the mere act of judging *itself* provides sufficient reason (or at least strong reason) for action. But of course the connection between judgment and motivation is a contingent one (contingent on the proper functioning of our faculty of moral sense). So this isn't actually a form of moral internalism.

Joyce argues against any naturalistic view that makes the connection between moral judgment and motivation a contingent one. In a case where someone happens to lack the relevant motivation, he points out how odd it would sound to say that the person did something wrong (committed a murder, say), but nevertheless did what they had no reason to refrain from doing (in light of their desires) (pp.203-4). But this is to conflate what one says from the first-order perspective of someone who possesses a normally-functioning moral sense, with what we might say as theorists *of* moral sense. From the first perspective of course we aren't going to allow that the agent had no reason to refrain from murdering. For our judgment that murder is wrong is categorical, not conditional on any particular set of goals. And as soon as we make that judgment we feel the appropriate indignation and punitive emotion. Consistently with this we might still, as theorists who maintain that the connection between moral norms and motivation is a contingent one, allow that the agent in this case (who is perhaps a psychopath) possessed no goals that provided him with reason to desist from his murderous course.

A similar mistake occurs a few pages later (207), where Joyce claims that if the connection between morality and motivation is merely contingent, then thinking in moral terms will be superfluous. Rather, all of the practical "oomph" will be equally achievable by thinking in terms of the relevant goals (the goals that contingently motivate moral action). But seen in

light of the model of moral sense sketched earlier, this is an obvious error. For the only way for an agent to *have* the relevant motivation (whether indignation or anticipatory guilt) is by having the appropriate belief activated from her system of moral norms – it is only *via* coming to believe that the act is wrong that the motivation to avoid it comes to exist. Likewise on the following page (208), Joyce asserts that if the connection between morals and motivation is contingent, then "moral deliberation just *is* deliberation about what is desired and how it might be achieved." Again, this is an obvious error. On the model of moral sense sketched earlier, moral deliberation will be deliberation about what is required of us by the norms that are stored in the norms data-base. The attachment of motivation to the results of that deliberation is automatic (but contingent). The resulting motivations don't themselves enter into our moral deliberations.

Joyce presents a related, but distinct, argument on page 207 which needs to be handled somewhat differently. He writes:

> If thinking and talking of the action as "morally wrong" adds something substantial that cannot be gotten from thinking and talking of the action's instantiating some natural property, then this counts as evidence against the adequacy of the moral naturalist's theory.

Suppose, for example, that the natural property in question is that the action would be forbidden by any set of rules that no one could reasonably reject who shared the aim of reaching free and unforced agreement (or some other variant on the constructivist accounts to be discussed in Section 2 below). If we endorse this theory, then we are committed to saying that all and only those beliefs stored in the norms data-base that possess this property will be true. But it doesn't follow that we can then dispense with normative concepts. On the contrary, our claim can be that it will only be if beliefs are stored in a certain canonical form (as beliefs about what is *wrong*, or *forbidden*, for example) that they will engage the motivational side of the system. What talking in moral terms adds is not some extra property in the world, but some extra motivation which, as a matter of contingent fact, wouldn't exist without it. But this is no problem for a moral naturalist who claims to be providing a constitutive, metaphysical, account of moral truth, rather than an a priori analysis of moral concepts.

2 Evolution and Truth Tracking

The second argument of Joyce's that we propose to consider is that evolutionary theorizing will undermine morality in the same sort of way that one's belief that Napoleon lost the battle of

Waterloo will be undermined if one discovers that one's belief was actually caused by taking some kind of pill (p.179). For Joyce thinks that the best evolutionary hypothesis will be some or other variant on the idea that nascent moral judgments among early hominids served to strengthen social commitments and encourage social cooperation. They didn't serve to register perception of an independent moral realm. He writes, "the function that natural selection had in mind for moral judgment was [nothing] remotely like *detecting a feature of the world*, but rather something more like *encouraging successful social behavior*" (p.131). He therefore thinks that the story of human evolution "debunks" moral realism. (He calls this "genealogical debunking".)

One response to this objection is to challenge the logic of the genealogical debunking argument. For genealogical explanations that don't involve truth-tracking need only undermine our warrant for holding the explananda beliefs on a temporary basis. Having discovered that a pill-taking caused your belief that Napoleon lost, you should withhold your assent from such a belief. But by consulting a history book your warrant for the belief can be restored. Likewise, we suggest, in the moral case. Having discovered that our moral sense was designed to encourage cooperative behavior, not to track truth, each of our moral beliefs is thereby undermined. But if we accept an independently motivated account of moral truth, then our warrant for some of those beliefs can be restored by showing that they are entailed by the account in question. Provided that rational reflection of this sort can insert or remove moral beliefs from our norms data-base (as it surely can – see Saunders, forthcoming), then we can arrive at a set of warranted (and true) moral beliefs even if our moral sense didn't evolve to track truth.

We ourselves are inclined to think that the correct account of moral truth will be provided by some or other version of constructivist moral theory, which can be warranted independently of evolutionary theorizing by considerations of wide reflective equilibrium. (Such theories occupy a central place in contemporary moral theory.)[3] In which case we can restore confidence

[3] See, for example, Rawls (1972) and (1980), Scanlon (1982) and (1998), Copp (1995), Milo (1995), and Korsgaard (1996a) and (1996b). For our purposes here, it is inessential which exact constructivist view we champion, provided that it constitutes a form of moral realism, and so long as it can be rendered consistent with some kind of reductive metaphysical naturalism. While all constructivists are agreed that moral truth is the outcome of some sort of hypothetical agreement or justificatory process, not all think of themselves as realists about moral truth, and not all are motivated by naturalistic concerns. These are large and difficult issues. Here we note only that by virtue of depending on subjunctive hypotheticals (e.g. what rational agents *would* agree on under certain conditions), at

in at least a subset of our moral beliefs by showing that the corresponding norms couldn't reasonably be rejected by those who share the aim of reaching free and unforced general agreement (say).

In fact, however, we suspect that if some or other variant of constructivist approach to moral truth is correct, then there is the prospect of a non-debunking *alignment* between evolutionary explanation and the content of moral belief. For it is plain that the evolutionary pressures that created our innate moral faculty would have designed it to be intimately connected with the evaluative attitudes of others.[4] We think that Joyce is right to emphasize an early human's need for cooperation and social cohesion, as well as her need to conceive, as he puts it, "how others will receive her decisions, her confidence in to whom she can justify herself" (p.117). But on a constructivist account, moral facts just *are* facts about whether or not one's actions could be justified to others or, more generally, facts about the sorts of attitudes others could hypothetically take toward certain courses of action. In which case, there is some reason to think that evolution has built us to track moral truth, as characterized by constructivist lights.

There is an obvious objection to this line of thought, however. This is that what matters from the point of view of evolution is that individuals should identify and internalize the norms of their community *whatever those norms should happen to be*. Whether the norm is decidedly moral (as in, you shouldn't steal from your neighbor) or non-moral (as in, you shouldn't eat duiker meat when the moon is full), breaching it may have essentially the same negative consequences for your fitness. And the anthropological data demonstrate that a very wide range of norms around the world are counted as moral ones (in the sense of attracting indignation, punishment, and guilt), in addition to those that we in the liberal West would recognize as such (for example, norms dealing with harm or fairness).[5] To put the point differently: a great many of the normative beliefs that are stored in any given individual's norms data-base will be *false* by

---

least some moral truths can be strongly mind-independent, obtaining even when evaluated against worlds in which there are no rational agents. That seems to us a kind of realism worth the name.

[4] Indeed, if one is impressed by the idea that natural selection would have favored individuals capable of attributing mental states to their conspecifics in order to predict and explain their behavior, then much of the cognitive architecture required for constructivist reasoning would already have been in place. The next step would have been "projecting" oneself into the minds of one's conspecifics in order to predict the evaluative responses likely to issue from their standpoints.

[5] See e.g., Haidt et. al. (1993) and Nichols (2002, 2004).

the constructivist's lights. Yet the acquisition of these false beliefs may be just as important to the individual's fitness, provided that they are widely shared in the surrounding community. The upshot, then, is that our moral sense hasn't evolved to track *truth*, but to track the moral beliefs of one's community.

The point is well taken. However, we think that there is a promising line of reply, which would involve building rather more into the structure of our innate moral sense than we have hitherto suggested. In a nutshell, the idea is that there is an innate disposition to engage in constructivist reasoning. If this were so, then it would of course be no accident that at least some of the beliefs in one's norms data-base would be true by the constructivist's lights, and the genealogical debunking argument would have been met head-on. For we would be able to claim that our moral sense has indeed evolved (in part) to track moral truth.

There are a number of considerations that suggest to us that this is a fruitful line to pursue.[6] One point is that it is very plausible that we possess an innate disposition to try to justify our actions to others in terms that they can freely accept (as well as to refrain from actions that cannot be so justified), as Joyce himself seems to acknowledge (p.117). For actions that can be justified to others will be immune from community punishment. Secondly, there are a variety of reasons why the process of justification cannot be a matter of mechanically applying existing norms. For moral norms (even more than the rules of formal legal systems) can be indeterminate, and can conflict. So establishing what the community norms require in any given case will be no easy matter. Moreover, any set of norms (again like the legal system) is likely to be radically incomplete. Many actions will neither be prohibited, prescribed, nor explicitly permitted. In which case individuals will need to be prepared to justify themselves to others in terms that don't just appeal to existing norms, but which rather presuppose that those others are in the market for reasonable agreement. Finally, there is some reason to think that many norms are actually formulated and modified via processes of the sort that constructivists envisage. For as Boehm (1999) demonstrates, people in hunter–gatherer societies take group stability, group cohesion, and the avoidance of conflict as explicit goals in their thinking and reasoning. (And note that

[6] Of course, following through on this strategy in detail would require both a worked-out constructivist moral theory (and one that is not only realist in nature but naturalistically acceptable) *and* a comprehensive theory of the innate structure of our moral sense together with an account of the evolutionary forces that shaped it. Needless to say, we are actually in a position to provide none of these things.

such societies are generally strongly non-hierarchical in structure, involving freely cooperating groups of individuals.) Hence much debate about what is, or isn't, acceptable will take the form of reasoning about rules that others could reasonably accept, on the assumption that those others, too, share the aim of reaching free and unforced agreement. If the disposition to reason thus is innate, then that will at least approximate to establishing that the disposition to engage in constructivist reasoning is innate also.

In conclusion, we believe that certain forms of moral realism are likely to be fully consistent with evolutionary accounts of the origins of morality, and with the postulation of an innate moral sense.

References

Boehm, C. (1999). *Hierarchy in the Forest.* Harvard University Press.

Copp, D. (1995). *Morality, Normativity, and Society*. Oxford University Press.

Haidt, J., Koller, S., and Dias, M. (1993). Affect, culture, and morality, or is it wrong to eat your dog? *Journal of Personality and Social Psychology*, 65, 613-628.

Joyce, R. (2006). *The Evolution of Morality*. MIT Press.

Korsgaard, C. (1996a). *The Sources of Normativity.* Cambridge University Press.

Korsgaard, C. (1996b). *Creating the Kingdom of Ends*. Cambridge University Press.

Milo, R. (1995). Contractarian Constructivism. *The Journal of Philosophy*, 92, 181-204.

Nichols, S. (2002). Norms with feelings: Toward a psychological account of moral judgment. *Cognition*, 84, 223-236.

Nichols, S. (2004). *Sentimental Rules: On the natural foundations of moral judgment.* Oxford University Press.

Rawls, J. (1972). *A Theory of Justice*. Oxford University Press.

Rawls, J. (1980). Kantian constructivism in moral theory. *Journal of Philosophy*, 77, 515-572.

Saunders, L. (forthcoming). Reason and intuition in the moral life. In J. Evans and K. Frankish (eds.), *In Two Minds: Dual Processes and Beyond*. Oxford University Press,

Scanlon, T. (1982). Contractualism and Utilitarianism. In A. Sen and B. Williams (eds.), *Utilitarianism and Beyond*. Cambridge University Press.

Scanlon, T. (1998). *What We Owe to Each Other*. Harvard University Press.

Sripada, C. and Stich, S. (2006). A framework for the psychology of norms. In P. Carruthers, S.

Laurence, and S. Stich (eds.), *The Innate Mind: Culture and Cognition*. Oxford University Press.