

What follows is a close-to-final draft of a symposium in *Philosophy and Phenomenological Research* 77 (2008) on *The Evolution of Morality* by Richard Joyce (MIT Press, 2006). It consists of (1) a preçis by Joyce, then discussions by (2) Stich, (3) Carruthers and James, and (4) Prinz, and finally (5) replies by Joyce.

Preçis of *The Evolution of Morality*

Richard Joyce

The Evolution of Morality attempts to accomplish two tasks. The first is to clarify and provisionally advocate the thesis that human morality is a distinct adaptation wrought by biological natural selection. The second is to inquire whether this empirical thesis would, if true, have any metaethical implications.

Before the hypothesis that human morality is innate can be fruitfully investigated we must understand its content. In this context, in claiming that X is innate I mean that the present-day existence of the trait is to be explained by reference to a genotype having granted ancestors reproductive advantage, rather than by reference to psychological processes of acquisition. In claiming that *human morality* is innate, I do not mean that humans are innately social, or innately nice and friendly, or even that we innately have emotions that favor social cohesion; rather, I mean that humans have an innate tendency *to make moral judgments*. Thus, an evolutionary explanation of, say, human *altruism* (whether in a psychological or an evolutionary sense) or *sympathy* would not count as an evolutionary explanation of human morality. Having an inhibition against cheating one's fellows is to be distinguished from judging that cheating is prohibited.

Although *making a moral judgment* is a different phenomenon from *being helpful*, it is, nevertheless, natural to assume that the former typically works in the service of the latter: that the capacity to engage in moral judgment enhances in some manner a creature's social tendencies. This poses a prima facie puzzle for moral nativism, since, it would seem on the face of it, natural selection is a competitive race where the laurels always go to the self-serving egoist. Chapter 1 undertakes the task of combating this assumption, by outlining four processes whereby natural selection may favor traits of helpfulness: kin selection, mutualism, reciprocity, and group selection. (Also discussed is the role of cultural selection and niche construction in the special case of human ultra-sociality.) In so doing, no pretense is made that *morality* is being explained—for the organisms subject to these processes may be insects or plants—but the task is a prerequisite for explaining morality, for the hypothesis to be explored is that natural selection favored helpfulness in the human lineage, and that the capacity to engage in moral judgment is a proximate mechanism for regulating this behavior.

Chapter 2 addresses the question of just what a moral judgment is. I argue that moral judgments must be identified in two ways: in terms of a distinctive subject matter (the moral realm pertains largely to interpersonal relations), and in terms of what might be called the "normative form" of morality (a particularly authoritative kind of evaluation). The first claim is not intended to exclude the very idea of truly self-regarding moral prescriptions; it is, rather, a generalization about moral systems. I devote quite a bit of attention to the second quality, suggesting that moral prescriptions are thought of as categorical (i.e., they are not pieces of advice on how a subject's ends may be best achieved), inescapable, and not dependent on the decree of any human authority or institution. I also engage in a more idiosyncratically metaethical debate concerning what kind of mental state(s) moral judgments

(considered as speech acts) function to express. I argue for a view that combines aspects of traditional cognitivism and noncognitivism: Moral judgments are assertoric but also express the speaker's conative states. The analogy is pressed with pejorative terms like "kraut": To say that "X is a kraut" is both to assert something (that X is German) while also expressing a conative attitude (in this case, deprecation).

Chapter 3 is something of an aside to the central strategy of the book. I develop a novel argument to the conclusion that language use is a prerequisite to certain moral emotions. The argument has two steps, both of which are controversial. The first seeks to establish that certain emotions (some, but not necessarily all) involve the application of concepts. In the case of guilt, for example, I claim that one does not count as having the emotion unless one tokens the concept *transgression* (which is not to say that one must all-things-considered believe that one has transgressed). The second step is to argue that certain concepts—some of which are those involved in moral emotions—are language dependent. For example, to be granted the concept *German* one needs (roughly) to be able to discriminate Germans from non-Germans; by contrast, to be granted the pejorative concept *kraut* one must have some linguistic knowledge (albeit possibly know-how). The *OED* tells us that "kraut" is "derogatory"—a comment that doesn't seek to describe what the word denotes, but rather describes a convention of usage that requires "semantic ascent" in order to state. Were we to line up all our platitudes surrounding the concept *kraut* and all those surrounding our concept *German*, the former list would differ from the latter in its inclusion of platitudes necessarily involving semantic ascent; the former will have to say something about the conventions surrounding *the word* "kraut," such as "It is a derogatory term; it is a word used to insult people." Drawing on results from Chapter 2, I claim that what holds for "kraut" holds also for moral terms, since they too have an entrenched evaluative component the failure to understand which counts as a conceptual incompetence.

This argument occurs in the context of discussing the capacity of non-human animals to engage in moral judgment. Along with Frans de Waal and others, I agree that chimpanzees have some of the "building blocks" of morality; I am more interested, however, in what they *lack* that is necessary for the real McCoy. I claim that one necessary condition that they lack is language. If correct, this establishes a constraint on *when* we can speak legitimately and literally of a moral sense having emerged in our evolutionary lineage.

Chapter 4 returns to the main goals of the book. The early parts of the chapter address the crucial question of *why* the moral sense may have evolved in humans. What reproductive benefits might such a faculty bring? I focus on the question of what reproductive advantage there may be for *an individual* (as opposed to advantage for the group) if he or she makes *self-directed* moral judgments (as opposed to judgments concerning others' conduct). To some extent, this strategic decision reflects my interest in the emotion of guilt, which is a *prima facie* puzzling case for the nativist. There may be many complementary answers to this question; I discuss two. First, I suggest that self-directed moral thinking can advance an individual's welfare by acting as a kind of psychological bulwark against various kinds of motivational infirmity, such as weakness of will or the discounting of future profits. Thinking that a certain action "simply *must* be done" may, in some circumstances, engage motivational structures more resolutely than even an awareness that the action is to one's own advantage. The former may exclude certain possibilities (of cheating, say) from one's deliberative domain in a way that the latter does not; the latter is vulnerable to the agent choosing to "rationalize" a suboptimal choice. According to this view, moral judgment is a kind of

personal commitment device. Second, I emphasize the social aspect of morality, suggesting that the general conspicuous costliness of moral conformity makes it well suited to function as an *interpersonal* commitment device. Building on the work of Robert Frank, I argue that moral thinking can benefit the individual operating in the social sphere by foreclosing certain practical possibilities, thus bringing about desirable ends via altering interactants' choices.

As to the question of *how* natural selection may have gotten our ancestors thinking in moral terms, I advocate the thesis of moral projectivism: that what gives moral phenomenology its quality of “out-there-ness”—as if our moral evaluations are responses to a normative realm that precedes them—is the fact that we project aspects of our emotional lives onto our experience of the world. Projectivism sits comfortably with the empirical evidence indicating moral thinking to have both emotional and cognitive elements.

The latter parts of Chapter 4 outline the empirical evidence for moral nativism. No pretense is made that the case offered is comprehensive or compelling to a skeptic about nativism. The main objective is to counter the tired old objection that nativist thinking amounts to nothing more than an appeal to Just So stories. Rather, what we have is a coherent, plausible, and testable hypothesis—one that deserves to be taken seriously. The investigation into the truth of the hypothesis may involve data from numerous empirical disciplines: experimental economics, neuroscience, anthropology, primatology, and various fields of psychology. Of particular interest to me are results from developmental psychology, especially concerning childhood competence at distinguishing moral from conventional norms. I recognize that the data (as presented by Elliot Turiel and others) is not unproblematic, but I contend nevertheless that what evidence is available favors a nativist interpretation. I offer what I suppose will be construed as a “poverty of the stimulus” challenge: namely, that it is hard to imagine even what could *possibly* be done to teach a brain moral thinking (i.e., to get a child to internalize norms) if it is not initially set up with specific kinds of mechanisms geared for such learning. The moral nativist hypothesis that is advocated is not that morality *with a particular content* is innate—I accept that mechanisms of cultural transmission play an enormous and perhaps exhaustive role in determining the content of an individual's moral convictions. The hypothesis is, rather, that there is an innate faculty—deserving of the name “the moral sense”—designed precisely to make this particular kind of cultural transmission possible.

The second part of *The Evolution of Morality* investigates what metaethical implications may be drawn from the nativist hypothesis. There is a long tradition of *prescriptive* evolutionary ethics, stretching back to the Nineteenth Century, that argues that moral nativism in some way or another provides the basis for a *vindication* of morality. A standard objection to all such proposals is that one cannot validly derive an “ought” from an “is”—something that is often (erroneously) called “the naturalistic fallacy.” I refute this generic reason for rejecting prescriptive evolutionary ethics, but I nonetheless think that all such proposals are flawed. In Chapter 5 I critically examine four “vindicatory” projects: from Robert Richards, Richmond Campbell, Daniel Dennett, and William Casebeer. All are rejected, and I attempt to draw some general conclusions about the recurring faults of such strategies.

The sixth and final chapter advocates a kind of moral skepticism derived from moral nativism. Nativism offers us a genealogical explanation of moral judgments that apparently nowhere implies or presupposes that these beliefs are true. Compare this with a native sense of basic arithmetic. Any reasonable explanation for why it was to our ancestors' reproductive advantage to have a hardwired belief that $1 + 1 = 2$ (say) will depend on that belief's being

true: a false arithmetic belief just isn't going to be useful. But a *moral* belief may well be useful even if it is false. The plausibility of the adaptational account of moral genealogy isn't affected whether we hold the beliefs to be true or hold them to be false. My contention, then, is that moral nativism can have epistemological implications; it cannot show that moral beliefs are false, but it might well show them to be unjustified (or *render* them unjustified, depending on one's theoretical epistemological commitments). In particular, any epistemological benefit-of-the-doubt that might have been extended to moral beliefs—based upon some principle of conservatism, for example—will be neutralized by the availability of an empirically confirmed moral genealogy that nowhere implies or presupposes truth.

The bulk of the chapter is devoted to the exercise of showing that the kind of evolutionary genealogy advocated does indeed not presuppose the truth of moral judgments. Certain versions of moral naturalism promise to rescue morality from skepticism, for they purport to show that moral facts are identical to, or supervene upon, the facts that explicitly figure in the genealogical account favored. Thus in Chapter 6 moral naturalism is attacked in general terms. I think there is much that could be said against moral naturalism, but the focus of my broadside is that naturalistic facts invariably fail to underwrite the essentially practical nature of the moral realm. I take on two different naturalistic strategies in turn: first, the kind that attempts to accommodate this special normative “oomph”; second, the kind that thinks that this characteristic is dispensable—that a set of properties with no special reason-giving force may nevertheless be identified with the moral realm. Regarding the latter, I try to draw out the unacceptable oddity of allowing that someone may be under a moral obligation with which he has no reason to comply. Any such theory seems to enfeeble our capacity to morally criticize wrongdoers, and whatever the properties are that the naturalist seeks to champion, they are surely too normatively wimpy to be mistaken for the ontological constituents of the moral realm. Chapter 4 argued that the “moralization” of our ancestors' practical lives contributed in various ways to the satisfaction of their long-term interests and made for more effective collective negotiation. (I think that this holds for *our* lives, too.) It is, I submit, precisely the special “practical oomph” with which moral prescriptions are imbued that enabled them to perform these functions. Any value system that fails to accommodate this feature could not so effectively play the social roles to which we have traditionally put morality, and thus we could not *use* it as we have used morality, indicating that it would not *be* morality.

The failure of moral naturalism means that we should not expect to locate moral facts surreptitiously buried in the apparently non-moral genealogy of morals. Thus moral nativism amounts to the discovery that our moral beliefs are the product of a process that is entirely independent of their truth, which forces the recognition that we have no grounds one way or the other for maintaining these beliefs. Moral nativism should undermine our confidence in moral thinking.

Some Questions About *The Evolution of Morality*¹

Stephen Stich
Rutgers University

Richard Joyce has written an admirable book, brimming over with fascinating findings, bold empirical hypotheses and philosophical arguments that are both innovative and provocative, all set out in a straightforward, engaging style. One of the virtues of this journal's book symposia that they give commentators an opportunity to ask questions that authors can address in their responses. But symposium articles must also be short, and by the time I had finished my second reading of Joyce's book, I had a list of questions that would fill many more pages than I am allowed. So, for want of a better strategy for narrowing down the list, I'll focus on questions that were suggested by apparent differences between Joyce's account of our "moral sense" and the account of the psychology of norms that Chandra Sripada and I have defended in a recent paper (Sripada & Stich, 2006). To fill in the necessary background, I'll begin with a very brief overview of the Sripada & Stich (S&S) model.

Figure 1 is a sketch of the psychological mechanisms which, Sripada and I argue, underlie the acquisition and implementation of norms. The job of the Acquisition Mechanism is to identify the norms in the surrounding culture whose violation is typically met with punishment, to infer the content of those norms, and to pass that information to the Execution Mechanism, where it is stored in the Norm Data Base. The Execution Mechanism has the job of inferring that some actual or contemplated behavior violates (or is required by) a norm, and generating intrinsic (i.e. non-instrumental) motivation to comply and to punish those who do not comply. There is good reason to believe that the emotion system is involved in punitive motivation and it may also play a role in compliance motivation, though the evidence for that is less persuasive. Influenced by the remarkable findings reported in Wheatley and Haidt (2005), Figure 1 portrays the making of moral judgments to be downstream from the emotion system. In Wheatley and Haidt's study, participants who were hypnotized to feel disgust when they read the word 'often' or 'take' made much more severe moral judgments about behavior described using one of these words than they made when the behavior was described without using the words. However, following Greene (2004, Greene et al. 2001), who has demonstrated very different patterns of brain activity in response to different sorts of moral dilemmas, we also included a second pathway leading to moral judgment which involves the explicit reasoning system and may not involve the norm and emotion systems at all. While Greene's account of the sorts of dilemmas that do not engage the emotion centers in the brain has been evolving steadily as new data become available, the rough idea is that they are relatively impersonal cases rather than those in which the interactions among agents are (as Greene used to say) "up close and personal." For present purposes, that's all we'll need about the S&S model, so let me turn to Joyce's book.

The central question in the first four chapters of *The Evolution of Morality* is "Is human morality innate?" (1)², and Joyce does an admirable job of saying how he will interpret the question. To ask whether morality is *innate* is to ask whether it "can be given an adaptive explanation in genetic terms: whether the present-day existence of the trait is to be explained

¹ I am grateful to Edouard Machery and Chandra Sripada for helpful comments on an earlier draft of this paper.

² All references to Joyce's book will be given in parentheses in the text. And, in case you were wondering, Joyce believes the answer to the question, as he interprets it, is *yes*.

by reference to a genotype having granted ancestors reproductive advantage.” (2) To ask whether *morality* is innate is to ask “whether the human capacity to make *moral judgments* is innate.” (4, emphasis added) As Joyce wisely notes, in order to address that question seriously, we need an account of what moral judgments *are*. Thus much of chapter 2 is devoted to a detailed account of the nature of moral judgments.

One crucial feature of moral judgments, on Joyce’s account, is that they are imbued with a kind of “practical clout” (or “oomph” as Joyce sometimes says) – they “draw attention to a deliberative consideration that cannot be legitimately be ignored or evaded.” (58) Moreover, this practical oomph “doesn’t have its source in internal or external sanctions, nor in some institution’s inviolable rules, nor in the desires or goals of the person to whom it is addressed. In this respect ordinary thought distinguishes moral requirements from conventional and prudential requirements.” Joyce goes on to note that “[t]here is a large body of empirical evidence ... demonstrating that even very young children make these distinctions.” (63) The empirical literature that Joyce is alluding to here is the work by Eliot Turiel and others that utilizes the “moral / conventional task”. (Turiel 1983; Nucci 2001)

I am inclined to think that the sort of architecture sketched in Figure 1 can go a long way toward explaining the “oomph” that looms large in Joyce’s account of moral judgment. For if a judgment is generated by the norm execution mechanism, then those who make the judgment are intrinsically motivated to comply with that judgment and to punish those who do not. Also, as Daniel Kelly and I have argued (Kelly & Stich, forthcoming), judgments generated by the norm execution mechanism will strike those who make them as “authority independent” in the sorts of experiments that Turiel and his associates typically employ. When participants in these experiments are asked to suppose that an authority figure has decreed that there is no rule prohibiting a transgressive action which violates a norm stored in the data base, this will have no impact on their motivation to comply with the rule and to punish the transgressions.

There are, however, other features of the S&S model which comport less well with Joyce’s account of moral judgment. One of these is the “second pathway” to moral judgment, the one which does not involve the norm execution mechanism or the emotion system. If there are moral judgments generated in this way, they pose a pair of problems for Joyce. First, it is far from clear where *these* judgments get their “practical clout” since there is no intrinsic motivation to comply with them or to punish those who don’t. Second, moral judgments generated in this way would pose a problem for Joyce’s projectivist account of moral phenomenology. According to Joyce, “moral attributes seem to be ‘in the world’” but “moral appearances are in fact caused largely by emotional activity. A corollary is that appearances are to some extent deceptive; though our judgments are in fact prompted by emotional activity, our phenomenology is one as of the emotional activity being a response to attributes instantiated in the world.” (128-9) There is much to commend in Joyce’s discussion of projectivism; it makes a promising start at analyzing and explaining important aspects of the phenomenology of those moral judgments that are “caused largely by emotional activity.” But, of course, the projectivist account does not apply to judgments generated via the second route – the one in which the emotion system plays little role.

There are a number of ways in which Joyce might respond to these problems. Perhaps the simplest and boldest way would be to deny that there is “second route” to moral judgment which does not involve the emotion system. Another option would be to offer some non-projectivist explanation of the objectivist phenomenology and practical clout of moral

judgments generated via the second route. These are not the only options, but rather than continuing to speculate, let me ask the author: What *is* your view about second route moral judgments? Do you think that they don't exist? If they do exist, what sort of account would you offer of their phenomenology and their clout?

These questions turn on a feature of the S&S model that seems to find no place in Joyce's account. Let me turn now to a feature of Joyce's account that plays no role in the S&S model. According to Joyce, the emotion of guilt "surely lies at the core of the moral conscience" (122-3), and conscience is unpacked as "a repertoire of judgments and emotions (most notably guilt) that motivate behavior in accordance with accepted standards of conduct even when external sanctions are absent." (120) So, for Joyce, guilt plays a central role in motivating moral behavior. On the S&S model, by contrast, guilt is accorded no special role. Since the model allows that the emotion system *might* be involved in compliance motivation, it is not incompatible with the claim that guilt is important in moral motivation. But I am rather skeptical of the proposal, since I find it hard to see how it is supposed to work. Guilt, after all, is an emotion one has typically has *after* one has committed some transgression. As Joyce puts it, "[g]uilt seems most naturally to associate with the judgment that the person *has performed* a wrongful action for which amends might be made." (102, emphasis added) But if guilt is an emotion one feels after one has performed a wrongful action, how, exactly, does it "motivate behavior in accordance with accepted standards of conduct?"

One familiar idea is that people believe that they will feel guilty if they violate one of the norms they have internalized, and that they are motivated not to violate the norm since they also believe that guilt is a very unpleasant emotion, and they want to avoid having that unpleasant experience. This is, however, a singularly implausible account of the phenomenology of *my* moral motivation when, for example, I decide to return a lost wallet or not to tell a convenient lie. And informal surveys among my students confirm that I am not unique. Indeed, these surveys suggest that concern about future guilt plays almost no role deciding what to do, except when the student has been raised in a religious family and the behavior being contemplated is sexual behavior. Even if I am quite wrong about the phenomenology of moral decision making – or if I am right about the phenomenology and the thoughts about future guilt are typically unconscious – it still would not support Joyce's contention that guilt plays a major role in moral motivation. For on the account we are considering, *the emotion of guilt* is playing no role in the generating compliance motivation. Rather it is the *belief that one will feel guilty* and the desire to avoid this feeling that are doing all the work. Joyce might, I suppose, suggest that the emotion of guilt plays a crucial role in *producing and sustaining* that belief, because people have learned via inductive inference that transgressions lead to guilt. But I know of no evidence that even begins to suggest that people learn the link between transgressing and feeling guilty in this way. Rather than speculating further, let me ask Joyce: Do you think that the emotion of guilt (rather than beliefs about the emotion) plays an important role in motivating people to act in accordance with prevailing norms even though guilt is typically experienced after a transgression has taken place? If so, can you provide some further details on how this works?

Joyce's account of moral judgment is rich and complex, and while most of the details are compatible with the S&S model, few of them would be predicted by that model. For example, according to Joyce, in order for an utterance, *S*, to count as a moral judgment there must be a "linguistic convention that decrees that when *S* is uttered [in an appropriate context] the speaker thereby expresses *two* mental states" (57, cf. 53); one of these mental states is a

belief, the other is “a connotative attitude” “such as approval, contempt, or, more generally, *subscription to standards*.” (70, emphasis added) Thus, he maintains, a pair of sentences like:

(1) The Elgin Marbles morally ought to be returned to Greece. But I subscribe to no moral standard that commends their return to Greece.

“would be challenged if uttered...”(56) There are, I suspect, many philosophers who would take issue with this (and other) features of Joyce’s account of moral judgment. Moral particularists, for example, might well balk at Joyce’s insistence that making moral judgments requires “subscription to standards”. (Dancy, 2005) But even if we grant that Joyce’s characterization of moral judgments is correct, its richness and complexity pose a problem. For if moral judgment requires *all of that*, what reason is there to think that people in cultures very different from ours *make* moral judgments? Why should we think that making moral judgments is a pan-cultural phenomenon? The question is an important one for a project like Joyce’s since, as Joyce recognizes, if he is right that human morality is innate, we should expect it to be present in all cultures, with the exception, perhaps, of those that are so stressed that normal psychological and social processes break down. Joyce clearly believes that “morality (by which I here mean *the tendency to make moral judgments*) exists in all human societies we have ever heard of.” (134, emphasis in the original). But once we realize how much Joyce has built into the notion of a moral judgment, the evidence he offers for this claim seems far from convincing. “Moral precepts,” he tells us, “are mentioned in the Egyptian Book of the Dead and in the Mesopotamian epic of Gilgamesh.... Moreover, morality exists in virtually every human individual. It develops without formal instruction, with no deliberate effort, and with no conscious awareness of its special features.” (134-5) And, lest we mistakenly interpret him as talking loosely here, Joyce adds: “When I talk here of ‘moral development’ I don’t just mean prosocial behavior or even simply prosocial emotions; I mean genuine cognitive ... moral judgments.” (135)

As I see it, these considerations (and those that Joyce offers in the next few pages) don’t come close to supporting his claim that the tendency to make the sort of rich and complex moral judgments that he has gone to such pains to characterize exists in all human societies. To the best of my knowledge, we have no serious information about the details of the linguistic conventions that prevailed in the communities that produced the epic of Gilgamesh or the Book of the Dead. To support his contention that “morality exists in virtually every human individual,” Joyce appeals to work in the Turiel tradition. Researchers in that tradition have maintained that the capacity to draw the moral/conventional distinction is pan-cultural and emerges early in development. But there is a growing body of literature indicating that it is simply false that there is a pan-cultural ability to draw the “moral/conventional” distinction as characterized by Turiel and his associates. Indeed, as I read that literature, the right conclusion to draw is that the moral/conventional distinction, as characterized by Turiel and his followers, is a myth.³ Moreover, I suspect that the practice of making moral judgments of the sort that Joyce describes is a culturally and temporally local one restricted to Western (and Western-influenced) cultural groups in relatively recent times. Of course, this suspicion would be substantially undermined if there was evidence that folks

³ For more on this admittedly controversial claim, see Kelly et al. (2007), Kelly and Stich (forthcoming), Nado, Kelly and Stich (forthcoming).

in a number of cultures very different from our own really do make Joyce-style moral judgments. Richard, do you know of any such evidence?

Though the S&S model says nothing about the evolution of the mechanisms it posits, the model does pose a puzzle for Joyce's account of the evolution of morality. Though that account is complex and nuanced, two ideas are quite central. The first is that the evolutionary function of moral judgment is to get people to *behave* in appropriate ways. "My thinking on this matter," Joyce tells us, "is dominated by the natural assumption that an individual sincerely judging some available action in a morally positive light increases the probability that the individual will perform that action..." (109) The second idea, and the one I propose to question, is that the primary sort of behavior moral judgment was selected to motivate is *cooperative or prosocial* behavior. Here is how Joyce makes the point.

[S]elf-directed moral judgment may enhance reproductive fitness so long as it is attached to the appropriate actions. We have already seen that the "appropriate actions" – that is, the fitness enhancing actions – will in many circumstances include helpful and cooperative behaviors. Therefore it may serve an individual's fitness to judge certain prosocial behaviors – *her own* prosocial behaviors – in moral terms. (109)

The benefits that may come from cooperation ... are typically long term values, and merely to be aware of and desire these long term desires does not guarantee that the goal will be effectively pursued.... The hypothesis, then, is that natural selection opted for a special motivational mechanism for this realm: moral conscience. (111)

On the S&S model, the norm acquisition system is designed to internalize whatever norms prevail in the surrounding environment. So if there are prosocial norms or norms of cooperation, they will be acquired. And, as Joyce rightly notes, "all human moral systems give a leading role to *reciprocal relations*." (140). But, as Sripada and I note, norms of cooperation are just one among many sorts of norms that are to be found in just about every culture.

[M]ost societies have rules that prohibit killing, physical assault and incest (or sexual activity with one's kin).... Most societies have rules regulating sexual behavior among various members of society, and especially among adolescents.... Examples like these could be multiplied easily in domains such as social justice, kinship [and] marriage [Most societies also have norms] governing what food can be eaten, how to dispose of the dead, how to show deference to high ranking people, and a host of other matters. (Sripada & Stich, 2006)

Since norms governing all of these matters are as ubiquitous as norms governing reciprocity, it strikes me as rather implausible that reciprocity and prosocial norms should have pride of place in an account of the evolution of morality. Moreover, there are other suggestions about the evolution of norms in which prosocial and cooperative norms play no special role. (Boyd forthcoming; Sripada forthcoming) Joyce does not deny that other processes may have played a role in the evolution of morality. Indeed, he suggests that "[g]roup selection – most probably at the cultural level – may well have been a major factor." His "hunch" however, "is that reciprocity, broadly construed, is what got the ball rolling." (141) Since Joyce offers no argument for his hunch, my last question is: Why does he think that an account which gives reciprocity a central role in the evolution of morality is a better bet than competing accounts in which reciprocity plays no special role?

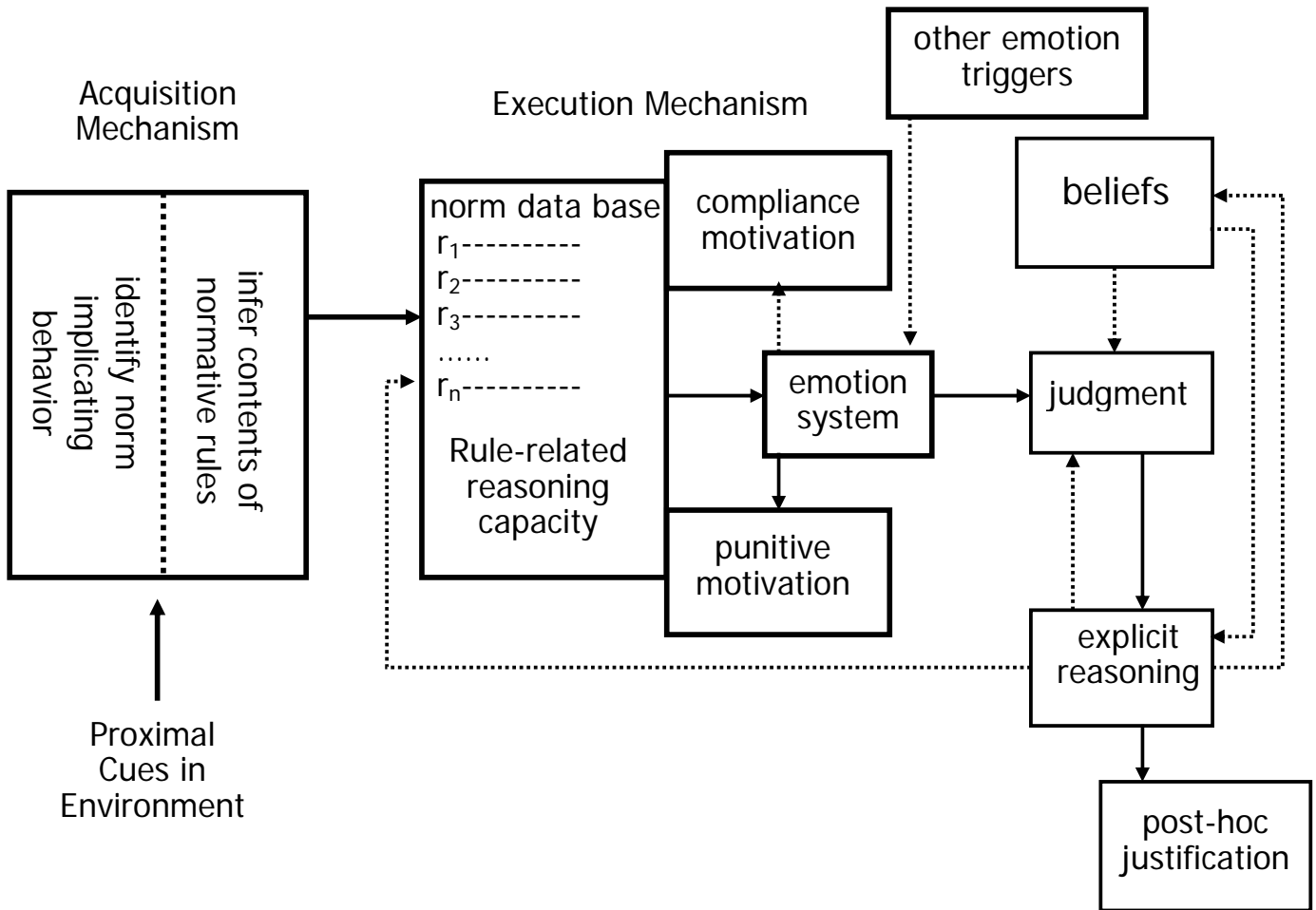


FIGURE 1

A sketch of the mechanisms underlying the acquisition and implementation of norms set out in Sripada & Stich (2006). Solid lines indicate links that we take to be well supported by evidence; dotted lines indicate more speculative links.

REFERENCES

Boyd, R. (forthcoming). Population structure, equilibrium selection and the evolution of norms. To appear in *Economics and Evolution*, Ugo Pagano ed., Cambridge University Press.

Dancy, J. (2005). Moral particularism. In *The Stanford Encyclopedia of Philosophy (Summer 2005 Edition)*, Edward N. Zalta (ed.), URL = <http://plato.stanford.edu/archives/sum2005/entries/moral-particularism/>.

Greene, G. (2004). fMRI studies of moral judgment. Unpublished lecture given at the Dartmouth College Conference on The Psychology & Biology of Morality.

Greene, J., Sommerville, R., Nystrom, L., Darley, J., & Cohen, J. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, Vol. 293, Sept. 14, 2001, 2105-2108.

Kelly, D., Stich, S., Haley, K., Eng, S. & Fessler, D. (2007). Harm, affect and the moral / conventional distinction. *Mind and Language*, 22, 2, 117-131.

Kelly, D. & Stich, S. (forthcoming). Two theories about the cognitive architecture underlying morality. To appear in P. Carruthers, S. Laurence & S. Stich, eds., *Innateness and the Structure of the Mind: Foundations and the Future*. (New York: Oxford University Press) 2007.

Nado, J., Kelly, D. & Stich, S. (forthcoming). Moral judgment. To appear in the *Routledge Companion to the Philosophy of Psychology*, ed. by John Symons & Paco Calvo.

Nucci, L. 2001. *Education in the Moral Domain*. Cambridge: Cambridge University Press.

Sripada, C. & Stich, S (2006). A framework for the psychology of norms. In P. Carruthers, S. Laurence & S. Stich, eds., *The Innate Mind: Culture and Cognition*. (New York: Oxford University Press) 2006. Pp. 280-301.

Sripada, C. (forthcoming). *Adaptationism, culture and the malleability of human nature*. To appear in P. Carruthers, S. Laurence & S. Stich, eds., *Innateness and the Structure of the Mind: Foundations and the Future*. (New York: Oxford University Press) 2007.

Turiel, E. 1983: *The Development of Social Knowledge*. Cambridge: Cambridge University Press.

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

Evolution and the Possibility of Moral Realism

PETER CARRUTHERS¹

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 form essential components of morality (or so we are inclined to believe). 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 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 follow Sripada and Stich (2006) in thinking that the human moral sense must include at least the following components: (1) A data-base of stored normative beliefs about

¹ The order of the authors’ names is alphabetical. We are grateful to Leland Saunders for comments on an earlier draft.

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 aren't conditional on 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

² Whether there is any need for a separate pro-moral motivation is moot. It may be that anticipatory guilt 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.

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”.)

We aren't convinced. We think that Joyce is right to emphasize an early human's need for cooperation and social cohesion. He is right to emphasize, further, 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). The alternative Joyce neglects to consider, however, is that 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. This is an alternative that occupies a central place in contemporary moral theorizing.³ If some or other variant of this constructivist approach to moral truth is correct, then we have the prospect of a non-debunking alignment between evolutionary explanation and the content of moral belief.

It is plain that the logic of the evolutionary pressures that created our innate moral faculty would have designed that faculty to be intimately connected with the evaluative attitudes of others. Only by identifying and internalizing the moral norms that are prevalent in one's community can one reliably avoid actions that will call forth the condemnation and punishment of others – with obvious consequences for one's fitness. If this is right, then we should expect to find that a central component of the human moral sense is *guilt*, since guilt essentially involves registering and responding to the justifiable disapproval of others. For then anticipatory guilt can be what motivates us to avoid actions that would call forth others' disapproval. Moreover, in cases where one has breached a moral norm, guilt can both send an honest signal to the community that one continues to share its values, and can motivate reparatory actions of various sorts, which can facilitate one's re-absorption into that community. It would appear, then, that there is some reason to think that evolution would have built us to track moral truth, as the latter would be 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).⁴ 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 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.

³ See, for example, Rawls (1972) and (1980), Scanlon (1982) and (1998), Copp (1995), Milo (1995), and Korsgaard (1996a) and (1996b). For our purposes 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 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 See e.g., Haidt et. al. (1993) and Nichols (2002, 2004).

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 temporarily undermine our warrant for holding the explananda beliefs. 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 easily be restored. Likewise, we suggest, in the moral case. Having discovered that our moral sense was designed to track community belief, not truth, each of one's moral beliefs is thereby undermined. But if one accepts a constructivist account of moral truth, then one's warrant for some of those beliefs can be restored by asking whether the corresponding norms could reasonably be rejected by those who share the aim of reaching free and unforced general agreement (or whatever). Provided that rational reflection of this sort can insert or remove moral beliefs from one's norms data-base (as it surely can – see Saunders, forthcoming), then one can arrive at a set of warranted (and true) moral beliefs even if our moral sense didn't evolve to track moral truth.

Suppose that this point is set to one side, however. We think that there is a further 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 would be to postulate 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 are 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 promising line to pursue.⁵ 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, proscribed, or 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 (which are generally strongly non-hierarchical in structure) take group stability, group cohesion, and the avoidance of conflict as explicit goals in their thinking and reasoning. 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

⁵ 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 neither of these things.

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.

Acquired Moral Truths

Jesse Prinz

The Evolution of Morality is a clear, highly provocative, and thoroughly enjoyable book. It beautifully integrates lessons from recent experimental psychology with work in evolutionary theory and philosophical ethics, making it essential reading for anyone interested in naturalistic approaches to morality. In the first part of the book, Joyce argues that our capacity to make moral judgments is innate, and, in the second half, he argues that a fully developed naturalistic account of morality can be used to debunk moral realism. I disagree with both of these claims, and I will take them up in turn. But it will also become clear that there is much I agree with in this book.

1. Moral Nativism

1.1 Joyce's Innateness Hypothesis

Admirably, Joyce distinguishes moral nativism from the view that we have an innate tendency to behave in morally admirable ways. Rather, he has in mind the view that we have an innate tendency to make moral judgments. To get clear on this claim, we need to see how Joyce characterizes moral judgments and how he characterizes innateness. He often uses the phrase "moral judgment" to refer to verbal reports, such as "killing is wrong." But, for clarity, I will use the phrase to refer to the attitudes expressed by such reports. On Joyce's analysis, moral judgments have several important features. First, they have both an emotional and descriptive component: they involve feelings such as guilt, and they also aim to designate real properties. Second, they purport to apply regardless of one's inclinations or extant social conventions, and those who endorse them take themselves to have reason to comply ("practical clout"). Third, they are intrinsically motivating. Fourth, they govern interpersonal relations and imply that violators deserve to be punished. I basically agree with this characterization. Indeed, I think that all the features that Joyce mentions can be derived from the supposition that moral judgments have an emotional basis. Emotions aim to designate properties (e.g., fear represents danger), emotions are intrinsically motivating, and emotions such as guilt and anger promote punishment behavior. I will argue later, contra Joyce, that emotions can also explain the practical clout of moral judgments. For now the main point is that I don't think Joyce's argument for nativism rests on implausible assumptions about the nature of moral judgment.

In calling moral judgments innate, Joyce means that the capacity to make moral judgments "can be given an adaptive explanation in genetic terms" (p. 2). I take the term "genetic" to imply that morality has a biological basis (though Joyce does not think there is a gene for morality). Joyce mentions the possibility that morality might result from cultural evolution, rather than biological evolution, but the bulk of his arguments are designed to show that morality is part of our bioprogram—children have domain-specific cognitive resources dedicated to the acquisition of moral rules.

Joyce speculates that morality is an outgrowth of systems that promote cooperative behavior, especially reciprocal altruism. But unlike the mere inclination to reciprocate, which we share with apes, he thinks that moral judgments serve as communicable commitments. Expressing a moral judgment is like signing a contract; moral testimony signals that you will

tend to act in accordance with a particular norm, that you will enforce that norm, and that you will tend to submit to punishment if you transgress.

Joyce thinks that morality could not have emerged before the evolution of language. As I understand that argument, it goes as follows. Moral terms both describe things and express attitudes towards those things. By analogy, consider the word “kraut” as used as a pejorative for a person from Germany. If you call a German a kraut, you describe them as being from Germany, and you express your negative feeling towards Germans. To acquire the concept expressed by “German” you need only acquire an ability to classify people by their national origins. In principle, that doesn’t require language. But, to acquire the concept expressed by “kraut,” you need to learn that the word “kraut” is derogatory. Thus, the concept requires understanding the role of the word used to express it. Likewise for moral concepts, because they too are both descriptive and expressive.

I am not convinced by this argument. Language certainly isn’t necessary for acquiring concepts that have an emotional component. For example, we can acquire the concept *disgusting* or *funny* or *sexy* without words, even though the words for these concepts are, arguably, both descriptive and expressive. We acquire such concepts easily because some categories in the world induce emotions in us, and the emotions they induce get incorporated into the category representations. For example, it may be that when we encounter scantily clad members of the opposite sex, we get aroused, and the concept *sexy* results. Likewise, if Germans were intrinsically off-putting, one could, in principle, acquire a concept like the one expressed by “kraut” without the help of language. Likewise, if (though evolution or conditioning) moral transgressions induce emotional responses, then we can acquire a concept of wrong that refers to those expressions and has an emotional component, even if we never learn the word “wrong.”

Fortunately, the claim that morality requires language is not essential to Joyce’s arguments for moral nativism. I turn to those arguments now.

1.2 Five Arguments For Moral Nativism

Joyce’s first argument for moral nativism is that morality exists in all known societies. I think this observation can be interpreted in two ways, and neither entails that morality is innate. If we take moral rules to be emotionally grounded beliefs about how one ought to behave, then the universality of morality is predicable on a non-nativist story. All societies need people to conform to rules, and the best way to get people to do so is to condition their emotions. If transgressors are punished or shunned, such emotional dispositions will arise quite naturally. But suppose Joyce’s claim is that all cultures have rules that are treated as something like categorical imperatives, such that people choose to conform because they think they are under a special, non-conventional, non-prudential, obligation to do so that would obtain regardless of our preferences. The anthropological record does not establish the universality of such rules. Indeed, small-scale societies might have little need for rules of this kind. In small-scale societies, people know each other, are often related, and share common beliefs and customs. We don’t need morality to avoid harming our near and dear; it is enough that we like them and that we have an obvious interest in treating them well. Arguably, categorical moral rules emerged only after human populations grew to large scales.

Joyce’s second argument for moral nativism is that the content of morality is similar across cultures. Every society has rules proscribing harm, prescribing reciprocity, sustaining

status hierarchies, and regulating bodily matters such as sex. The universality of such norms suggests that morality is heavily constrained by biology. Against this inference, I would offer two observations. First, the variation within these domains is absolutely dizzying. Some societies engage in headhunting, cannibalism and slavery; some societies tolerate grotesque inequity; some societies are nearly egalitarian, while others have rigid class hierarchies; some societies have strict moral rules governing the body and others are extraordinarily permissive. As Joyce notes, every moral value we endorse has been rejected by some other culture. Thus, learning clearly plays a large role in norm acquisition. Second, there are straightforward cultural explanations of these rule categories. A stable society must have rules protecting some people against harm, distributing access to power, mandating some degree of cooperation, and regulating sexual access. Thus, cross-cultural comparison offers little evidence for nativism and strong evidence for learning.

Joyce's third argument is that morality emerges in all children without formal instruction. This claim is questionable. Children are a bit like little psychopaths: they lie, steal, and cheat. Caregivers engage in nearly constant norm enforcement. Hoffman (2000) estimates that children under ten experience behavioral correction every 6 to 9 minutes. In addition, children hear public discourse about norms and observe the punishment of others. They may not get textbooks on moral conduct, but there is plenty of negative data.

Joyce also points to work by Cummins (1996), which purports to establish that three-year-olds are more competent with deontic rules than descriptive rules. Children do better in a game that requires testing to see which mice violate a queen's orders, than in a game that requires testing to see which mice violate a queen's descriptive claim. This finding can be explained without assuming moral nativism. Moral rules are behavioral injunctions. Children can learn the consequences of rule violation by getting punished and they can acquire the disposition to enforce rules by imitation. If a linguistically competent child is told something of the form, "If something is X, then it had better be Y," the child will understand that this is a conditional injunction to punish any individual that is X and not Y. In contrast, when a child is simply given the description, "If something is an X, then it is a Y" there is no behavioral injunction. Children may represent such conditionals using mental models in which arbitrarily chosen Xs are also Y. To confirm such a model, one should look for an arbitrary X and see if it is Y. The way we model descriptive conditionals carries no practical implications for entities that fail to satisfy the consequent (do we look at non-black things to confirm that all ravens are black?). The fact that children reason differently about such semantically different categories is not in and of itself evidence for innateness, especially given the fact that adults use these deontic and descriptive statement in different ways. Deontic conditionals are typically used to forewarn punishments of those that violate the consequent, and descriptive conditional are used to generate expectations about those that conform to the antecedent. I don't see why young children wouldn't pick up on this.

Joyce's fifth argument is based on the observation that young children can distinguish between moral and conventional rules, even though that distinction is not explicitly taught. Moral rules are treated as more serious and less dependent on authority; they also tend to be justified by appeal to harms and rights. In response, I would first point out that parents tend to use different socialization techniques for moral and conventional rules (Grusec and Goodnow, 1994). In the case of moral rules, parents may use harsher punishments, and they may draw children's attention to the suffering of others. These forms of intervention condition children to have negative emotions when they consider violations of moral rules. Those emotions can

make the rules seem authority independent, because the very thought of, say, hitting induces a negative feeling. Consistent with this explanation, it should be noted that the distinction between moral and conventional rules is highly flexible. Nisan (1987) has shown that in traditional societies there is a tendency for children to treat seemingly conventional rules moralistically (e.g., children in a traditional Palestinian village judge that there should be universal prescriptions against calling teachers by their first names). Thus, the distinction between moral and conventional rules is difficult to draw on the basis of content. Rather, it seems to be that some rules are punished in a more severe and evocative way, and these rules get treated as more serious, less authority contingent, and less likely to be justified by appeal to a mere cultural practice.

Joyce counters that it is difficult to see how any general purpose learning mechanism could be used to learn that conventional rules are less dependent on authority than moral rules. He says that authority independence is not an observable feature, and indeed, when it comes to social conventions children rarely experience authority figures saying that violations are acceptable. If children were learning from observation, they should say it's never permissible to violate a conventional rule.

Here I think Joyce underestimates the impact of emotional learning. If conventions are punished less harshly, then emotions play less of a role in their acquisition. Learning a convention is a bit like learning the rule of a game. While playing the game, it's important to obey the rules, but no one will spank you if you don't. One obeys the rule because that's how the game is played, but the rules could be changed. By analogy, if you asked a child whether it's okay for two people to play a special version of checkers in which the pieces move on a straight line instead of diagonally, she would probably say that's fine, even if she had never observed the game played that way. Suppose however, the child got an electric shock every time she moved a checker piece in a straight line, and then someone asked her whether it would be okay to play this alternate version of checkers. She might anticipate the pain, and judge that such behavior would be unacceptable. She might ultimately come to realize that there is nothing intrinsically wrong with moving checker piece in a straight line, but prior to such explicit instruction she would (as Joyce acknowledges elsewhere) project her bad feeling onto the behavior itself. When rule learning is not grounded in emotions, children are more flexible. They learn that such rules are regularities (this is what we do around here), but they do not learn anything that gives the rules special normative force (this is how things should be). In sum, I think we bootstrap into the moral/conventional distinction by emotional conditioning. Moral norms are those that are conditioned via intense emotions, and they seem serious and authority independent as a result. Conventional norms seem like regularities that can be violated without serious emotional cost. No domain-specific mechanisms are required to explain this distinction.

I have been suggesting that morality emerge through the course of emotional conditioning. This is not a complete account of where morality comes from (see Prinz, 2007), but I hope to have shown that Joyce's arguments for moral nativism are insufficient as they stand. For each of his observations, there are promising explanations that do not presuppose a domain-specific moral capacity.

2. Moral Facts

Joyce described the second part of his book as an implication of the first, but the arguments he apply to any theory that accounts for the existence of moral judgments in naturalistic terms. A naturalistic genealogy, whether nativist or not, would account for our moral judgments without presupposing that they are true, and this would render moral truths explanatorily dispensable. Moreover, Joyce argues that there are independent reasons to think that moral judgments presuppose a class of facts that are difficult to accommodate on a naturalistic theory of the world. Therefore, naturalists should abandon moral realism and conclude that moral facts do not exist. Joyce's argument improves on similar moves that have been made by Harman, Mackie, and others in the meta-ethics literature, but it is ultimately unsuccessful.

With Joyce, I think that moral judgments are emotionally based, and in this respect they are like judgments about what is funny, disgusting, delicious, loathsome, and sexy. When it comes to such judgments, there are three possibilities. The first is that they don't aim to assert anything. This seems like an unpromising suggestion. When we say that something is funny, we take ourselves to be saying something true; we have debates about what's funny; we make mistakes about funniness (a bad joke may seem funny when in a giddy mood); the word "funny" is a predicate; and we can sort the world into things that strike us as funny and things that do not. All this suggests that judgments of funniness aim at the truth. The second possibility is that judgments of funniness aim at the truth but fail; nothing is actually funny. One might be tempted by such an error theory because there seems to be no common denominator—no intrinsic essence—uniting all funny things. This discovery would entail an error theory if the concept of funniness logically entailed that funniness is a mind-independent property. But I don't think there is any such entailment. The concept of funniness is compatible with the possibility that being funny is a subjective or response-dependent property. Roughly, something is funny just in case it is disposed to amuse us. This third option allows us to say that judgments of funniness are true when they ascribe that property to things that tend to amuse.

Following John McDowell and David Wiggins, I think moral concepts refer to response-dependent properties too. Roughly, something is wrong just in case it is disposed to cause disapprobation in the judge (Prinz, 2007). This captures the intuition that moral facts are both real and motivating. Joyce would reject this account of moral facts. Recall that, on his view, moral judgments aim to designate facts that have normative force independent of our inclinations ("inescapability") and that provide sufficient reasons for action ("authority"). Joyce thinks that response-dependent properties cannot have this kind of practical clout. Let's see if he is right.

Suppose that moral facts are response-dependent (e.g., dishonesty is wrong because it is disposed to engage me). If so, there are still two ways to accommodate the intuition that moral truths are independent of inclinations. First, there may be actions that are disposed to enrage me, but happen not to on a given occasion. Second, I can use the predicate "wrong" rigidly to pick out things that enrage me in this world now, while recognizing that those things could exist in worlds where my sentiments are different. There is also a way to accommodate the intuition that moral facts provide sufficient reasons for action. If dishonesty is wrong, then it has the disposition to cause an emotion in me, and, in one sense of reason, that gives me reason to be honest. Just as it makes sense for me to say I watch Monty Python because it is funny, I can say I condemn condemn because it is infuriating.

Joyce would say that this is not the right kind of reason. If the authority of morality stems solely from the contingent fact that certain things evoke emotions in me, then I cannot criticize others, I will lose my motivation to be moral when I have a strong desire to transgress, and I will render appeals to moral truths redundant because my behavior can be explained by appeal to my emotions. If these implications follow, the norms that I follow are too wimpy to call moral.

I think this argument can be answered. First consider criticism. If you and I share the overlapping moral values, then I can certainly criticize you for making a moral mistake by your own standards. If we have radically different moral values, I may lose the authority to criticize you, but this is not a counter-intuitive result; when we discover morally divergent societies, we call them exotic not wrong. Now consider motivation. Suppose I really want a book on your bookshelf, and I know that my resistance to filching it is nothing more than an inculcated disposition to feel guilty. Still, that guilt is enough to stop me (compare: if I learn that my taste in music is conditioned, I still have reason to listen to my stereo). Finally, consider the charge of redundancy. Joyce complains that if moral authority consists in emotional dispositions, then appeal to moral facts adds nothing beyond appeal to how we feel. I think Joyce is right about this (it follows trivially from the naturalist reduction), but it does not undermine morality. When we say that dishonesty is wrong, we are asserting that it is outrageous, and that fact can guide behavior, evoke punitive attitudes, and convey commitments. Far from undermining morality, the response-dependent account helps to explain why moral facts have practical implications.

In conclusion, I am not persuaded by the two central theses in Joyce's excellent book. I think that morality is learned through emotional conditioning, and I accept a form of moral realism according to which moral facts are response-dependent. But there is much I agree with in Joyce, including the contention that emotions are essential to morality and the methodological maxim that empirical findings can have metaethical implications.

References

Cummins, D. 1996. Evidence for the innateness of deontic reasoning. *Mind and Language*, 11, 160-190.

Grusec, J. E. and Goodnow, J. J. (1994). Impact of parental discipline methods on the child's internalization of values: A reconceptualization of current points of view. *Developmental Psychology*, 30, 4-19.

Hoffman, M. L. (2000). *Empathy and moral development: Implications for caring and justice*. Cambridge: Cambridge University Press.

Nisan, M. (1987). Moral norms and social conventions: A cross-cultural comparison. *Developmental Psychology*, 23, 719-725.

Prinz, J. J. (2007). *The emotional construction of morals*. Oxford: Oxford University Press.

Replies

Richard Joyce

The Evolution of Morality (TEOM) is intended to be a short book with a lot of content. I chose to articulate an argumentative thread that moved from evolutionary biology, through psychology, and all the way to metaethics—and I wanted to do so in an economical and nimble manner. I would still like to think that on this occasion the virtues of this approach ultimately outweighed the vices, but the pitfalls are, nevertheless, unavoidable and undeniable. Many important arguments are too swift, significant points are left undeveloped, and large tracts of relevant literature pass by with just a quick wave of the hand. It's helpful to encounter four such astute commentators, whose critical remarks reveal to me where the gaps in my case were most yawning; I appreciate their careful criticism even more than their praise.

Broadly speaking, *TEOM* has two parts: one discussing and advocating an empirical thesis, and one drawing metaethical conclusions. Stephen Stich concentrates solely on the descriptive component, Peter Carruthers and Scott James are concerned only with the metaethical implications, while Jesse Prinz divides his critique between the two. It seems best to structure my rejoinder into two parts accordingly: the first answering Stich and part 1 of Prinz, the second answering Carruthers and James and part 2 of Prinz. Even so, there are more questions raised than I can possibly answer here; the necessary triaging of my commentators' queries leaves a number of criticisms untreated.

I The empirical case for nativism

Stich and Prinz are skeptical of my confidence in the evidence supporting moral nativism. It has to be stressed that I did not take myself in *TEOM* to be presenting compelling evidence confirming the nativist case; I was satisfied with the more modest ambition of sketching out a clear, coherent, productive, plausible, and testable hypothesis. Of course, there is no denying that in *TEOM* I took on the task of advocating this hypothesis in the interests of exploring where one might look for confirming evidence, but I did not suppose myself to be presenting a comprehensive case. Indeed, I did not mean to suggest that there even necessarily is a comprehensive case to be made: My all-things-considered judgment is simply that nobody knows whether moral nativism (under any of its disambiguations) is true, and nobody should be either asserting or denying the hypothesis with any confidence.

Thus, I do not intend to take on Prinz's non-nativist interpretations of the various data point by point, though there would no doubt be much interest in that exercise. Interpreting such evidence in such a way as to convincingly support either nativism or non-nativism would take far more care, time, and patience than either Prinz or I have yet brought to the matter (and I certainly don't have the space for the task on this occasion). I concede that his non-nativist interpretations of the data may well turn out to be correct; much more work needs to be done.

However, there is one general clarification worth making concerning the way I have characterized moral judgment. In Chapter 2 of *TEOM* I went to some trouble to specify the features of the phenomenon, because it seemed to me utterly fruitless to embark on an investigation of moral nativism if it were not made clear what the phenomenon *is* for which nativism is affirmed or denied. This need seems particularly pressing when it comes to

making (or denying) claims of cross-culturality. I asserted that moral judgments are a cross-cultural phenomenon; both Stich and Prinz have their doubts. In Stich's case, it almost seems that he thinks that the complexity of my characterization *per se* counts against nativism—which surely cannot be right. (I don't think he intends to endorse the principle that only phenomena that can be characterized in simple terms can count as adaptations.) In any case, all parties will agree that it is a vital preliminary to investigating any claim of cross-culturality that we understand the nature of the target trait sufficiently to identify it. My sense is that both Stich and Prinz have misunderstood my intentions on this matter. I suspect them of thinking that because I have provided a fairly complex account of what a moral judgment is, it is more likely that moral judgments of this sort ("Joyce-style moral judgments" Stich calls them) are cognitively sophisticated and therefore will more plausibly be construed as cultural elaborations of a more fundamental suite of native capacities. But although my attempt to articulate the features of moral judgment in a comprehensive manner leads to a relatively complex description (involving reference to categorical imperatives, practical inescapability versus authority, etc.), I deny that the experience of *making* such judgments is a terribly complicated one. What I am really trying to pin down is a kind of blunt conviction that certain actions just "must be done"—a sense of what is "fitting" that is not tied to the subject's ends and is not ultimately contingent on the decrees of any authority figure. Such a sense is, I maintain, at the phenomenological level a brutish and basic kind of thought. One might usefully compare this with the concept of *logical inference*: The idea that from "*p*" and "*If p then q*" the conclusion "*q*" simply *must* follow seems to me a phenomenologically primitive thought (and without consulting any anthropological literature, I'm confident that it's a universal human thought), and yet it is extremely challenging to articulate the content of the thought—even for those who are completely competent at employing it—and the philosophical debate about how to understand conditional inference is dauntingly complex.¹

It might also clarify my view of moral normativity if we compare it with prudence. Suppose someone says "You ought not do that" and the answer comes back "Why not?" If it is a prudential "ought," then the speaker should be able to provide some kind of answer that makes reference to the subject's welfare ("Because you'll hurt yourself," "Because you risk punishment," etc.). But if it is a *moral* "ought," I maintain, then the speaker may find herself feeling somewhat dumbfounded, answering "Well, you *just mustn't*; it's a rule; I shouldn't have to explain this to you." I submit, in short, that when we look for cross-cultural evidence of moral judgments, it is a certain kind of psychological naivety, not a sophistication, that we should be seeking.

Prinz thinks that the anthropological record does not support the universality of these Joyce-style moral judgments: "All societies need people to conform to rules, and the best way to get people to do so is to condition their emotions. ... In small-scale societies, people know each other, are often related, and share common beliefs and customs. We don't need morality to avoid harming our near and dear; it is enough that we like them and that we have an obvious interest in treating them well." This argument is puzzling to me. On the one hand, Prinz admits that all societies have need for rules; on the other, he suggests that social cohesion may be maintained in a small-scale society simply if all members like each other. But if merely *liking* each other were sufficient to maintain cooperation, then what need is there for rules? The fact that a society does need rules—ones for which "transgressors are

¹ One can parody Stich: "If inference requires *all of that*, what reason is there to think that people in cultures very different from ours *make* inferences?"

punished or shunned”—indicates that warm and fuzzy fellow feelings are in fact not doing the whole job. Moreover, the very idea of a punishable *transgression* already indicates a kind of categoricity—the categoricity that Prinz suspects “emerged only after human populations grew to large scales.” After all, to fail to conform to a *hypothetical* imperative (i.e., to fail to heed advice on how to achieve one’s ends) hardly counts as a *transgression*—let alone a punishable one. What is bad about failing to conform to a hypothetical imperative is that one has put at risk the satisfaction of one’s own ends. (Even if the imperative is “Don’t punch others,” if it has hypothetical status then it is not the harm to the others that ultimately undergirds the advice, but the fact that in harming others one will somehow be undermining one’s own ends.) I am confident that no culture employs only hypothetical imperatives as its principal normative framework. That a society thinks of nonconformity to a set of norms as a type of *transgression*, that it *punishes* noncompliance, that noncompliers feel, or are expected to feel, *guilt* (as opposed to *foolishness* at having sabotaged their own projects), that members of the society are likely to feel *punitive anger* towards noncompliers (as opposed to the *pity* that is usually reserved for those who thwart themselves)—are all factors that count as evidence against the normative framework in question being classified as hypothetical or prudential. (See Joyce 2007 for more discussion.) Prinz and Stich are not alone in failing to appreciate this fact; I think it’s a common oversight. Undergraduate encounters with Kant can leave one with the impression that categorical imperatives are deeply complicated and mysterious—the result of an idiosyncratic modern cultural cocktail that mixes Western Christianity, Enlightenment Humanism, and the convoluted reflections of a certain Prussian genius. Such assumptions make it easy to overlook how primitive and ubiquitous categorical norms really are.

Debates about cross-culturality will not get very far if the opposing parties are in fact discussing different phenomena. Assuming that consensus can be achieved, we then face the question of who bears the burden of proof. Usually I find burden of proof arguments tiresome and pointless, but here it seems to me legitimate to raise the matter. Of course the advocate of cross-culturality (with respect to some trait) would love to provide a catalog of all known cultures and demonstrate the presence of that trait in each. But is it really fair to demand this of her? It seems to me not unreasonable to suppose that the burden really lies on the shoulders of the skeptic about universality: Let him bring forth the counterexamples.

Prinz and Stich both raise questions concerning another aspect of my characterization of moral judgments—namely, that I often categorize them as a species of speech act. Prinz is bothered by this tendency, and prefers to use the phrase “moral judgment” to denote not verbal reports, but “the attitudes expressed by such reports.” The problem with this is that knowing what attitudes are expressed by the verbal reports is no straightforward matter. The question of whether moral judgments (as speech acts) express beliefs, or desires, or some combination of the two, or something else, is one about which metaethicists have argued for almost a century. It is hard to see how one could *start* by treating moral judgments as a certain kind of attitude without begging a big fat metaethical question. To my mind, the perspicacious approach is to begin by treating moral judgment as a species of speech act (for then we at least have a definite public object to discuss), and then, on the basis of cogent argument, *determine* what attitudes are expressed. Only then will one have earned the right to use the phrase “moral judgment” to denote some determinate attitude(s).

Stich is concerned that in treating moral judgment as a specific kind of speech act I have undermined my case for nativism. I argue that moral judgments are ways of both expressing

beliefs and expressing conative states (e.g., subscription to a normative framework).² Stich objects: “To the best of my knowledge, we have no serious information about the details of the linguistic conventions that prevailed in the communities that produced the epic of Gilgamesh or the Book of the Dead.” But of course we do have such information, else we could not translate the works! To know that in Sumer a certain cuneiform pattern denoted *dog* and another pattern denoted *cat* is to know a linguistic convention. Moreover, we appear to know something of their *normative* linguistic conventions: A respectable translation of the epic of Gilgamesh contains words like “evil,” “duty,” and “transgression” (George 1999). We are told that the Egyptians of the Old Kingdom customarily referred to the Nubians as “vile” and “wretched” (Yamauchi 2001: 3). In Ancient Rome, “pagan” (“paganus”) was often used as a pejorative term meaning something like *country bumpkin; hick*—a word, note, with both a descriptive and a derogatory role (i.e., with which one would simultaneously express both a belief and a conative attitude). If the scholars who decipher such works are not inferring the normative conventions dominating the original language—and hence the conventions employed by the original language users themselves—then what business do they have in offering such translations? (Does anyone really imagine that an Ancient Egyptian would not have been perplexed by the utterance “The Nile is blue, but I don’t believe that the Nile is blue”?—but from this, I maintain, we can conclude something about Ancient Egyptian speech acts.)

Stich also puts to me a number of explicit clarificatory questions, to which I now turn. First, he asks about moral judgments that occur with no emotional arousal. The model developed by him and Chandra Sripada postulates a “second pathway” to moral judgment—one that bypasses the emotional systems altogether. They postulate this in order to accommodate certain data uncovered by Joshua Greene (2004; Green et al. 2001). Stich thinks that my model focuses on the role of emotion too exclusively, and cannot, without revision, be reconciled with Greene’s data.

In response, let me say first that I never purported in *TEOM* to provide a model with the architectural specificity apparent in Stich and Sripada’s diagram. Theirs is an admirable effort, and I did not intend my thoughts to be in serious competition. Although I certainly emphasized the role of emotion—especially in making the case for moral projectivism—I did not mean to exclude other possible pathways. Indeed, I explicitly declared “I don’t mean to suggest that every moral judgment humans make is the product of an emotional episode” (132). I did not even suggest that *most* moral judgments are the product of an emotional episode. The proposal that I sketched in briefest form on p.132, but which I intend to develop in future work (Joyce forthcoming *a*), is that for moral projectivism to be true as a general thesis (as opposed to holding true of a token moral judgment) moral judgments that are the result of emotional projection must be in the appropriate sense “paradigmatic”—where this is not a statistical notion but rather derives from an asymmetry between those episodes of moral judgment that involve emotion and those that do not. By analogy, the color projectivist need not claim that every single color judgment is the product of perceptual projection; in some

² Stich thinks that many philosophers would take issue with this. That may be so, but it should be noted that it is something for which I offer arguments, not something I claim dogmatically. Further, I am puzzled by Stich’s singling out of particularists as among those who might balk. A particularist will deny that in making a moral judgment one is subscribing to a *principle*, but this denial is consistent with affirming that in making a moral judgment one is expressing one’s subscription to a normative framework. Jonathan Dancy often describes the particularist as holding that moral judgment involves a kind of “sensitivity” to reasons—and this sensitivity may be categorized as a form of subscription to a normative framework.

circumstances one can work out the color of something by inference, without ever laying eyes on the item. But such an inferential color judgment would be parasitic upon color judgments that are the result of perceptual experience. One way to try to illuminate the asymmetry is by observing that a person blind from birth, who has never had the qualitative visual experience of an object, lacks the color concepts (or so it has often been claimed, and so there is at least some temptation to say). A person born blind may still know that ripe tomatoes, raspberries, and blood fall within the extension of the predicate "...is red," but it is tempting to say that when she utters the sentence "Ripe tomatoes are red" she doesn't *entirely understand* what she is claiming. By parity, the moral projectivist need not claim that the *only* way of coming to have a moral judgment is via projecting one's emotions onto one's experience of reality. Perhaps one can come to believe that someone's actions were morally wrong via inference or some other broadly cognitive, non-emotional process. The projectivist can limit his thesis to one concerning *paradigmatic* moral judgments. What this strategy would involve would be the location of a kind of emotional impairment (equivalent to color blindness) along with motivating a doubt that those who are so impaired really have moral concepts.

For example, suppose someone suffers from some kind of localized brain damage that leaves him utterly unable to experience the conative side of moral judgment—he has no capacity for guilt, never feels moralistic anger or moralistic disgust, never feels approval or disapproval. Such a person might be able to latch on to the knowledge that stealing, promise-breaking, and pedophilia fall within the extension of the predicate "...is morally wrong," but there is a defensible inclination, I think, to say that when such a person utters the sentence "Stealing is morally wrong" he doesn't *entirely understand* what he is claiming. One might say the same of psychopaths, who have trouble recognizing emotional cues in others (Blair et al. 2001a, 2002), and themselves experience very little in the way of fear, sadness, empathy, or guilt. It is almost certain that these emotional deficiencies are the cause of the psychopath's antisocial tendencies. But psychopaths do not merely make *poor* moral choices—what is important to my present argument is the possibility that they are unable to make full-blooded moral judgments at all. Certainly they can utter moral sentences that we would ordinarily consider true—"Stealing is morally wrong," etc.—so they can give a superficial impression of making moral judgments. But careful testing reveals that in fact they fail to make basic conceptual distinctions that just about everybody else makes (Blair 1995; Blair et al. 2001b). It is tempting to conclude that psychopaths are conceptually incompetent when it comes to moral concepts; they do not mean by "morally wrong" what the rest of us mean.³ Making this argument solid would lay the foundation for establishing a category of "paradigmatic" moral judgments which the moral projectivist may then privilege.

Stich also wonders where judgments arrived at via the "second pathway" get their "practical clout"—a quality I claim is an important element of moral assessment. Contrary to how Stich seems to interpret my argument, I do not assert that someone who makes a

³ In a classic study of psychopaths in the mid-20th century, Hervey Cleckley explicitly likened psychopathy to colorblindness: "The [psychopath] is unfamiliar with the primary facts or data of what might be called personal values and is altogether incapable of understanding such matters. It is impossible for him to take even a slight interest in the tragedy or joy or the striving of humanity as presented in serious literature or art. He is also indifferent to all these matters in life itself. Beauty and ugliness, except in a very superficial sense, goodness, evil, love, horror, and humour have no actual meaning, no power to move him. He is, furthermore, lacking in the ability to see that others are moved. It is as though he were colour-blind, despite his sharp intelligence, to this aspect of human existence. It cannot be explained to him because there is nothing in his orbit of awareness that can bridge the gap with comparison. He can repeat the words and say glibly that he understands, and there is no way for him to realize that he does not understand" (1941, p. 90).

judgment imbued with “clout” is thereby intrinsically motivated. If someone judges “Sally really mustn’t do that, irrespective of whether she enjoys it; Sally cannot legitimately ignore this consideration,” then that person is making a judgment with “clout.” But whether the person making this judgment is thereby motivationally aroused to act against Sally is another matter; the mere “cloutishness” of the judgment doesn’t require it. That said, I certainly also hold that there is a fairly reliable *contingent* link from moral judgment to motivation. A judgment imbued with clout can play a role in motivating compliant behavior (for example, when self-directed it can act as a buttress against weakness of will); and indeed I’ve speculated that this may have been an important aspect of what made moral judgment adaptive for our ancestors. In Stich and Sripada’s model there is a dotted arrow feeding back from “judgment” to “compliance motivation” via “norm data base.” Whether this is the correct feedback route is something I’m unsure of, but I do agree that there is some road from “judgment” to “compliance motivation,” though it may run through other intermediary boxes that are not at present represented in the diagram. In sum: Stich’s worry here seems to be that it is difficult to see how judgments arrived at via the non-emotional pathway could have clout, since clout is a motivationally loaded quality. It is the latter point I deny: clout can be represented in purely cognitive terms (whatever exactly that means), and I am satisfied if the model has motivation lying downstream from judgment (though, like Stich and Sripada, I want to see it potentially upstream as well).

Stich also worries about the role of guilt in motivating compliance. He wonders how guilt could motivate one to comply with a given norm at a given time, since typically guilt will arise only *after* one has already violated that norm. Certainly I never meant to suggest that one deliberates as follows: “If I violate this norm then I will feel guilty, which will be unpleasant, therefore I shall comply.” Nevertheless, there is in fact plenty of strong empirical evidence that guilt does affect one’s tendency to comply, though the exact mechanisms of this relationship remain obscure. One experiment manipulated subjects’ levels of guilt while they were engaged in bargaining games, and found that guilty-feeling individuals would, after pursuing a non-cooperative strategy in the first round of play, display considerably higher levels of cooperation in subsequent rounds, even a week later (Ketelaar & Au 2003). Another study revealed that guilt-prone fifth-graders were, as adolescents, less likely to engage in crime and more likely to be involved in community service (Tangney & Dearing 2002: ch. 8). In a study of college undergraduates, guilt-proneness (to be distinguished from shame-proneness) was associated with endorsing such claims as “I would not steal something I needed, even if I were sure I could get away with it” (Tangney 1994; see also Tibbetts 2003). In a longitudinal study of jail inmates, guilt-proneness assessed shortly after incarceration negatively predicted recidivism and substance abuse during the first year after release (Tangney et al. 2007). In sum, there is an enormous amount of evidence confirming the relation between the emotion of guilt and norm compliance. (See also Carlsmith & Gross 1969; Freedman 1970; Regan 1971; Zhong and Liljenquist 2006.) Stich asks me to elaborate on “how this works,” but this is something I must leave to the psychologists.

The final question that Stich puts to me is why I think an account that gives reciprocity a central role in the evolution of morality is a better bet than competing accounts. I will make several brief comments in reply. First, Stich seems to conflate reciprocity with the more general notion of cooperation. The claim that all cultures have norms pertaining to cooperation (“prosociality”) is not the same as the claim that all cultures have norms pertaining to reciprocity. (Both claims, however, happen to be true.) Second, the notion of

reciprocity that I employ is broader than one might assume: It includes indirect reciprocity (thus including the general benefits of a good reputation and the costs of punishment); it includes *sexual access* as a form of “currency” (thus including forms of sexual selection). Drawing attention to the breadth and richness of the category of *reciprocity* was actually one of the sub-ambitions of Chapter 1. Third, I would reaffirm the dangers of assuming that what we in the West might count as a purely self-regarding action should be categorized as such in another culture. (Food taboos are often cited as an instance of self-regarding moral norms; but if one thinks that transgressions are likely to anger the ancestral spirits, thus provoking their injurious influence upon one’s family and community, then it is not being considered as a self-regarding action at all.⁴) I think that once these second and third points are properly digested, the claim that in all cultures norms of reciprocity are *dominant* is much more plausible. Finally, I would underline that my advocacy of the thesis that reciprocity was the leading evolutionary driving force in the emergence of human morality was explicitly given the status of a “hunch,” and not much of consequence in *TEOM* depends on its truth. My primary objective was to show that here we have a perfectly adequate hypothesis, and there is no need to go beyond *individual* selection in advocating moral nativism. If I were prepared to argue that there is a persuasive body of evidence in favor of the reciprocity hypothesis, I should not have used the word “hunch” (though see Joyce 2006, where I argue the case in a bit more detail.⁵)

II The metaethical implications of moral nativism

The final chapters of *TEOM* argue that moral naturalism is inadequate. But, of course, moral naturalism—in one or another of its myriad forms—is a popular position, and thus it comes as no surprise that my skeptical conclusions should bring forth piqued moral naturalists peddling their well-thumbed wares. In *TEOM* I admitted that “perhaps the best that we can do is to examine contender reductive moral naturalisms case by case” (190), but I rejected this protracted strategy as inappropriate for such a book. Yet two of the commentary papers—by Prinz and by Carruthers and James (C&J)—champion forms of moral naturalism, forcing me now to specify the flaws of particular versions.

Both Prinz’s and C&J’s favored naturalisms can be seen as broadly of the same family: They both identify moral facts with naturalistic facts about certain agents’ responses; they see moral facts as response-dependent facts. I am not sure that there is a cogent *generic* criticism of all such theories; they are sufficiently varied that they face different obstacles. Unfortunately, both of the naturalistic programs that my commentators are cheering on are expressed in too indeterminate a way for me to be entirely sure which criticisms apply. Nevertheless, there is enough here to raise powerful objections. I identify three problems that both theories face: the incompleteness problem, the practical relevance problem, and the content problem.

⁴ Note that it would not follow that such a norm is merely a prudential one. One may think of some action both as something that “simply must done” as well as recognizing that certain harms would befall the perpetrator. The crucial matter is whether the norm would continue to be affirmed for circumstances where the self-harm is (if only in imagination) subtracted from the equation. (I take Plato’s discussion of the Ring of Gyges to be the first comprehensive examination of this matter.)

⁵ Joyce 2006 is little more than a condensed version of certain parts of *TEOM*. An expansion of the discussion of the reciprocity hypothesis is just about the only different material in Joyce 2006.

Prinz expresses his preferred naturalism in several different ways: Something is morally wrong just in case it is disposed to cause disapprobation ... (i) in us, (ii) in the judge, (iii) in me. The first actually arises when he is discussing *funniness*, so perhaps it is best put aside; but it is nonetheless worth highlighting the question: “Who is *us*?” Regarding (iii), I assume that Prinz isn’t advocating the megalomaniacal view that everyone’s moral judgments should concern what *Jesse Prinz* (“me”) would approve of, so I guess he means that moral facts are relative to each judge—that is, (ii) and (iii) are equivalent. One thing that confuses me is that there’s no mention of *the circumstances* in which the judgment is made, rendering the specification of the disposition incomplete. It’s as if the property of *fragility* were analyzed as “the disposition to break.” This, clearly, is unfinished; one needs to hear about the conditions under which breakage would occur—e.g., “when dropped” (though even this would be too unspecified). Similarly, there is simply no fact of the matter about what I would approve of *simpliciter*. In some circumstances I might approve of such-and-such, but given different circumstances (e.g., had I been raised in Maoist China, or had I just received news that my family had been killed, or had I taken LSD, or were I to live on Mars in the 23rd Century, etc., etc.) I might have disapproved of the very same things. So Prinz needs to constrain the circumstances of the judge in some manner. What is he going to say? Moral wrongness (for X) is whatever would cause X disapprobation in circumstances of *full information*? of *impartial attention*? of *calm reflection*? or what? Note that even if Prinz plumps for one of these orthodox options, this “incompleteness problem” doesn’t quickly disappear. The description “*what Richard Joyce would approve of in circumstances of full information*,” for example, remains too unspecified to denote any dispositional property. Fully informed RJ raised in Maoist China approves of one thing, while fully informed RJ as he actually was as an angst-ridden teenager approves of something else; fully informed RJ in a grumpy mood on Monday may approve of different things than fully informed RJ in cheerful spirits on Tuesday.

Prinz doesn’t hint how (or whether) he intends to constrain the specification of the judge’s circumstances (or the idealized qualities of the judge). But supposing he does in some manner: Wrongness (for X) = whatever X would disapprove of if she had qualities Q and were in circumstances C. The challenge is to specify Q and C in such a way that wrongness is still *practically relevant* for X. If these aspects are idealized too far, then they will denote a state that the actual X rarely, if ever, attains. We could imagine X then reasonably complaining: “I acknowledge that if I *were* in auspicious circumstances C, and if I *were* to have the fine qualities Q, then I would feel thus-and-so about this action, but given that I’m *not* in C and I *don’t* have Q, what relevance does this counterfactual hold for me? (i.e., why is the fact that the action is *morally wrong* of any practical relevance to me?)” Even if C and Q are specified as states that X would find highly desirable to attain, his question is no less pertinent. I might find lying on a Tahitian beach a desirable state to attain, but it doesn’t give *me now*—in the middle of winter—a reason to bask shirtless outside, nor motivate me to do so.

In any case, the monster looming over Prinz’s version of naturalism is relativism of the most radical and rampant rank. What I will approve of will diverge from what you will approve of, and where we both may agree Genghis Khan will beg to differ. It seems that there must be moral facts *for me*, and moral facts *for you*, and moral facts *for Genghis Khan*, and moral facts for every other person. (There may even be moral facts *for me on Monday* and different moral facts *for me on Tuesday*.) The extent of the relativism alone is cause for

concern, but, moreover, it seems utterly unconstrained: there are no guaranteed checks on the *content* of what might be approved of. If Jack the Ripper approves of slaughtering women, then, according to Prinz, this is morally good *for him*. There is no perspective-transcendent point of view from which we can criticize Jack's perspective. Of course, we can criticize him *from our point of view* (because from our perspective slaughtering innocents is wrong), but Jack can just as legitimately criticize us *from his point of view*. It hardly seems adequate that all that we can say from an "objective" perspective against an outlook that glories in unspeakable violence is that it is statistically unusual.

Prinz suggests that X may criticize Y if (i) X and Y share overlapping moral values, and (ii) Y is making a mistake by Y's own standards. I am not sure how or why these two conditions are combined. If Y is making a mistake by her own standards, then this seems sufficient grounds for some sort of criticism, regardless of whether the one doing the criticizing shares Y's values. Both criteria are, besides, unclear. First, what is it to "share overlapping values"? If X and Y disagree over any moral matter, then trivially they do not share all values. If X and Y morally agree over anything (e.g., that eating broccoli on Tuesdays is permissible), then trivially their values overlap. How much disagreement must there be before we can speak of "radically" different moral outlooks, such that parties "lose the authority to criticize" each other? Second, what is it to make a "moral mistake by your own standards"? I suppose Prinz has in mind (inter alia) something like a person endorsing certain general values but failing to apply those values to a particular case (or range of cases) while lacking adequate ground for making this exception. But there is no reason to assume that those people whom we would wish to criticize (who occupy a perspective somewhat different, but not "radically" different, from our own)—e.g., liberals versus conservatives?—are making this kind of mistake, or, indeed, any kind of "internal" mistake. Though no doubt people sometimes endorse inconsistent moral frameworks (or apply them inconsistently) and are, as a result, subject to warranted criticism (though why only from those who share their perspective I don't know), the idea that this observation might somehow legitimate *all* cases of moral criticism that it is pre-theoretically desirable to accommodate—or even come within spitting distance of doing so—is a vain hope. Was Jack the Ripper making a moral mistake by his own standards? We don't know. Does it follow, though, that we must withhold passing moral judgment on him? Surely not. Was Jack's moral outlook *radically* different from ours? Maybe; perhaps he reveled in the violence and saw his actions as those of a misunderstood *ubermensch*. Or maybe not; it's possible that he shared our moral outlook but was subject to pathological compulsions for which he despised himself. Again, let's say that we don't know. Does it follow, though, that we must remain proportionally uncommitted about whether to morally criticize him or treat him as "exotic not wrong"? Surely not.

In defending moral relativism in his 2007 book, Prinz counters some technical concerns (that the theory is incoherent) and reacts to accusations of insidiousness. My objections here fall into neither category. Rather, I am observing that the kind of dispositional natural properties that are being offered as the ontological constituents of the moral realm *don't come close* to satisfying the pretheoretical desiderata of what moral properties should look like.

Carruthers and James also offer a version of moral naturalism. But before they do so they voice a specific objection to my case against naturalism—an argument they accuse of involving "obvious errors" and "clear fallacies." I will clarify my stance against moral naturalism before explaining what is inadequate about their preferred version.

I claim that moral normativity has a distinctive kind of practical “oomph.” Despite appearances, I choose the word “oomph” carefully, since it is indeterminate, non-theoretical, and metaphorical—and thus, I maintain, does a decent job of capturing certain aspects of the phenomenology of moral judgments made by ordinary thinkers in everyday contexts. Ordinary thinkers probably have a thoroughly inchoate idea of what the “must-be-doneness” of moral rules consists in, but this is not to say that it is a peripheral or negotiable aspect of morality. On the contrary, I argue that it is of the utmost importance. (Analogy: Ordinary speakers will be adamant that there is a distinction between accidental behaviors and intentional actions, but press them to articulate what this *intentionality* (and its attendant *freedom*) consists of, and their answers will typically crumble.) It is entirely conceivable that this quality of “oomph” really is just a weird and quasi-mystical notion (like *freedom*, perhaps?), for which no adequate description, satisfactory to an analytic philosopher, can be provided. Nevertheless, in *TEOM* I made an attempt to give the notion some distinct content: I suggested that we might try to understand this oomph as a combination of *inescapability* and *authority*—both of which I described in some detail (tying the latter to a theory of practical reasons), and the conjunction of which I dubbed “practical clout.” A careful reading will reveal that I tempered this suggestion with qualifications, indicating that I took *clout* to be at best a promising way of articulating oomph. (I wrote: “That morality has practical oomph is a simple observation; whether that oomph should be cashed out as *clout* is a philosophical problem” (62).) I distanced myself from the claim that ordinary speakers would naturally express themselves in terms of “inescapability” and “authority”; I am attempting, rather, to “precisify or explicate the folk notion” (192) in terms that might be unfamiliar to a competent speaker and may even be coherently denied by her. In light of these reminders, let me turn to C&J’s objections.

They claim that my arguments against any moral naturalism that makes the practical nature of morality merely a *contingent* matter fail to take into consideration the important distinction between what would be affirmed from a “first-order perspective of someone who possesses a normally-functioning moral sense” and what might be asserted from the perspective of a theorist *of* moral sense. I do not wish to deny the distinction, but rather observe some limits. In order to focus our thinking, consider a parody of their argument. Ordinary speakers, competent at identifying and discussing *shapes*, will affirm that all squares are four-sided. It seems ludicrous to claim that this affirmation reflects only the perspective of the first person, and that the four-sidedness of squares is something that might be reasonably denied by the theorist *of* the shape sense.⁶ So how does C&J’s employment of the first-order/theorist perspective distinction in the case of the relation between morality and practicality differ from the distinction in the case of the relation between squares and four-sidedness?

It is here that my distinction between the intentionally vague term “oomph” and the more carefully defined term of art “clout” becomes relevant. (And I realize that some readers will be sniggering!) I maintain that *some* kind of special practical oomph is a necessary feature of moral judgment—though whether this quality can be given any clear articulation is an open question. I think that a theorist who claims that this practical oomph is merely a feature of the first-person perspective, and that from the theorist’s perspective morality is just another set of

⁶ One can imagine a philosopher denying that squares exist—perhaps by affirming some kind of radical Berkeleyan idealism—but that would be very different from the claim that squares have a number of sides totaling something other than four.

norms with no special binding qualities that need explaining, no distinctive “must-be-doneness” to it—is comparable to the theorist who asserts that squares are not really four-sided. The distinction between the first-person perspective and the theorist’s perspective can still be upheld: I have no problem with the theorist saying “When it is judged that someone is under a moral obligation, that judgment is imbued with a distinctive kind of practical oomph, but we theorists can make no sense of this quality, therefore nobody is ever really under a moral obligation.” Indeed, I have developed arguments of this structure myself (see most especially Joyce 2001). What I object to is C&J’s very different claim: “When it is judged that someone is under a moral obligation, that judgment is imbued with a distinctive kind of practical oomph, but we theorists can make no sense of this quality, therefore this distinctive kind of practical oomph *is not a feature of moral obligation at all.*” The issue is whether this quality of “practical oomph” is an expendable aspect of morality—whether a normative system stripped of this quality would warrant the description “moral.” I claim that it is not and it would not (respectively). Of course I realize how difficult it seems to assess the claim when it uses this purposely blurred term “oomph.” Yet it may be that this is the best that we can do, and it would be a mistake to try to do better. Nadeem Hussain notes that “part of what might attract one to an error theory about the moral in the first place is the thought that there is something deeply mysterious about moral concepts and the moral properties they supposedly pick out. Morality, one thinks, is an ideology, and mystification is the life-blood of ideologies. Surely it would be no surprise, then, if some fundamental unclarity is essential to morality’s ideological role. Given this essential unclarity, no surprise, then, if moral concepts seem to systematically escape analysis” (2004: 155-156).⁷ It seems to me not implausible that the sense of “practical requirement” with which natural selection may have endowed us—which emerges in the course of childhood development as the individual comes to internalize norms—is a primitive sort of feeling/thought which resists analysis, decomposition, explication, or naturalistic demystification.

However, in *TEOM* I was not content to conduct the argument in these mysterious terms. (Perhaps I ought to have been.) Rather, I endeavored to give some concrete articulation to *oomph*—cashing it out as a combination of *inescapability* and *authority*; i.e., “clout.” So understood, the battle lines get drawn at the dispute over agents’ *reasons*—and this, understandably, is where C&J make their stand. My mounting doubt over whether the practical nature of morality is really optimally captured by reference to an agent’s reasons makes me somewhat ambivalent about pressing my side of the argument. This reservation is partly due to a growing appreciation of the fact that the notion of *a reason* is so contested and confused in the field of philosophy (and elsewhere) that its introduction does not in fact represent an advance in clarity or specificity over the continued use of the intentionally hazy term “oomph.” This doubt was not very apparent in *TEOM*, however, so let me put it aside for the moment and tackle C&J’s argument on its own terms. Under these terms, then, my claim is that any form of moral naturalism that construes the relation between moral prescriptions and an agent’s reasons as merely contingent is inadequate, since it is an ineliminable platitude of moral discourse that those who are under moral obligations *have reason* to comply. C&J’s response is that this commits the error of conflating perspectives. They concede that from the first-person perspective one’s negative moral appraisal of a murderer (say) will include the

⁷ Recall also Wittgenstein’s observation that moral discourse consists largely of similes, yet “a simile must be a simile for something ... [but] as soon as we try to drop the simile and simply state the facts which stand behind it, we find there are no such facts” (1965: 10).

contention that he *had reason* to refrain from killing, but they think that it is open to the theorist to recognize that the murderer—being a psychopath, let’s suppose—as a matter of fact “possessed no goals that provided him with reason to desist.” But this is not something I need deny. The crux of the debate is not whether the murderer had or did not have reasons; it is whether the theorist’s denial that he had reasons is consistent with the theorist’s continued conformity with the first-person judgment that the actions of the murderer were morally wrong. If I were entirely confident in my explication of *oomph* in terms of reasons, I would object as follows: “We theorists agree that this psychopathic agent had no reason to refrain (for we are convinced that the only cogent theory of practical reasons is some form of instrumentalism), but on what grounds can one continue to affirm that his actions were nevertheless morally wrong? If the naturalistic property instantiated by the psychopath’s actions can be so easily divorced from the reasons he has for acting and refraining from acting, then in what sense is it an essentially *practical* property at all? But if it is not essentially practical, then what business do we have identifying it with *moral wrongness*?” (And I might go on to say something about squares and four-sidedness.) Nothing C&J say addresses this complaint.

Now, although I’ve just expressed much of the foregoing in the subjunctive mood, it is more or less the form that my argument in *TEOM* takes. And I continue to think that it is a perfectly defensible line of argument. Nevertheless, when push comes to shove I may retreat and regroup: I may admit that struggling to understand the practical element of morality solely in terms of *reasons* has not proved illuminating, and so I may even acquiesce to the theorist’s claim that someone may be under a moral obligation without having any reason to comply (though this is not an admission I will make quickly). But I will insist that the hopeful moral naturalist answers a challenge: If you don’t understand the practical element of morality in terms of reasons, then how *do* you propose to understand it? I am prepared to accept that it may turn out that this *oomph* can never be adequately analyzed, that it is a kind of magical and indescribable quality. So much the worse for moral naturalism, if that is so. But what I emphatically will not accept is any naturalist attempting to sidestep the challenge by claiming that there is nothing especially unusual about the practicality of morality that requires any special explanation. Nor will I accept that this elusive practical element is just one moral platitude among many, and that extirpating this problematic component would leave us with a kind of normativity still warranting the name “moral.”

C&J evidently would take one of these latter objectionable avenues, though it is not clear to me which. In any case, they offer a particular form of moral naturalism that they think (A) I have neglected to consider, (B) fits well with my case for moral nativism, (C) allows us to construe the evolved moral sense as a “truth tracking” faculty, and therefore (D) indicates how moral nativism leads to moral realism.⁸ Their favored theory is a kind of constructivism, according to which “moral facts just *are* facts about ... the sorts of attitudes others could hypothetically take toward certain courses of action.” They leave the details of the crucial counterfactual intentionally vague, making it difficult for me to respond with specific criticisms, but they say enough for me to highlight the shortcomings of their contender. Space, obviously, does not allow me to embark on a comprehensive critique, but I will briefly

⁸ That there is a question mark hanging over the claim that this kind of “hypothetical agreement” naturalism constitutes a kind of *realism* is something that C&J are alive to (in their note 3). I have discussed whether it deserves to be classified as “moral realism” in Joyce forthcoming *b*.

observe that their theory faces many of the same challenges as does Prinz's (though I will present the problems in a different order).

First, they face the *content* problem. What guarantee do we have that a group of humans will not agree to the most callous of policies? (see Shafer-Landau 2003: chapter 2; Velleman 1988; Sobel 1994). What sort of policy, for example, could be justified to Genghis Khan and his henchmen? The constructivists' response is to find some way of idealizing the agents or their circumstances of response. Moral facts may be identified with what a group of *fully-informed* agents would agree to, or a group of *rational* agents, or a group of *rational and impartial agents*, etc. But the content problem is just as obstinate for C&J as we saw it was for Prinz. How do we know that even *being rational* (say) will exclude a preference for ethnic cleansing? It is only by assuming a substantive (and contentious) theory of rationality that one could be confident that this kind of idealization will secure the right kind of normative output. Moreover, it is not obvious that the property of *what rational agents would agree to* has any determinate content at all (i.e., C&J also face the incompleteness problem). By comparison, there is, presumably, no fact about what rational agents would choose as their favorite color (see Joyce 2001: 84ff.). The conspicuous concern here is that the only way to guarantee (A) convergence among idealized agents (i.e., a solution to the incompleteness problem), and (B) convergence towards the desired normative output (i.e., a solution to the content problem), is to idealize the agents *in a moral* sense. Moral facts may be identified, for example, with what would be agreed upon by rational *and virtuous* agents. Such agents presumably won't sanction ethnic cleansing. But clearly this route becomes viciously circular if our original task was to provide a naturalization of moral properties.⁹

C&J seem to recognize this objection, though they do not express it in its most damaging form. They seem anxious that constructivism might turn out to place *odd* or *non-moral* things in the moral category (e.g., eating duiker meat when the moon is full); they fail to mention that things that are intuitively *immoral* might be classed as acceptable. In any case, their response is curious. They point out that my strategy of "genealogical debunking" at best represents a challenge to the epistemological status of moral judgments—a challenge that may then be met by the moral constructivist (and the hypothetical contractualist in particular). But it is difficult to see how this is supposed to alleviate the content problem. My claim is that discoveries about the genealogy of a set of beliefs might remove the epistemic warrant that we might have assumed these beliefs enjoy (based on some principle of conservatism, say), thus rendering them in need of epistemic justification. C&J are correct that I have said nothing to exclude the possibility of some theorist then coming forward to provide that lost justification. (The debunking should be read as a challenge, not a knockout blow.) Suppose that the moral constructivist then volunteers to restore warrant to our moral beliefs. It is *now* that the content problem arises: as revealing a glaring inadequacy in the constructivist's case.

These are not the only troubles for C&J's brand of moral constructivism. In light of my previous comments, it should come as no surprise that I am also doubtful that this naturalistic theory will satisfy the desideratum of accounting for the special practical oomph of morality—which amounts to observing that they also face the practical relevance problem. To see this, start by acknowledging the infinitude of hypothetical attitudes: A given action—stealing a newspaper, say—might be such that drunken Vikings would heartily approve of it,

⁹ That this problem is reminiscent of the Euthyphro dilemma is worth mentioning, if only because the dilemma is one that Carruthers himself confidently employs in dispatching theistic ethics in one curt paragraph (1992: 13-14).

be such that zealous medieval samurai would think it dishonorable-but-not-forbidden, be such that Soviet communists seeking to promote the Workers' Revolution would regard it as obligatory, be such that rational agents who share the aim of reaching free and unforced agreement would judge it unacceptable. Let us say that the given action actually *has* all these dispositional properties simultaneously, and many more besides. The question, then, is why should a person, contemplating the action, give a damn about any of them? Why should she be allowed (indeed, expected) to utterly ignore the opinions of hypothetical drunken Vikings (not denying them, note, but ignoring them), but not equally free to ignore the pronouncements of hypothetical rational agents aiming to reach an unforced agreement? What's so practically relevant about the latter that it deserves to be identified with the realm of *moral* facts?

One answer to this last question is that the latter dispositional facts (concerning rational agents trying to reach agreement) promise to match up with the content of our pretheoretical moral opinions in a way that the other dispositional facts so dramatically do not. But this is something that I have already questioned. Even putting the content problem aside, however, we are still left wondering why one particular sort of hypothetical agreement supplies a special kind of normative force to actual agents that the infinitude of other hypothetical agreements do not.¹⁰ The obvious way of putting this is in terms of *reasons*. C&J do not hint at how they would respond to this query, but it is a question that Carruthers has tackled elsewhere. Concerning hypothetical agents behind a veil of ignorance, he asks "Why should I have any reason to accept the rules that *they* would accept?" and "Why would this [the hypothetical agreement of hypothetical agents] be something worth dying for?" (1992: 44). Carruthers's answer is that we just *do* care about justifying ourselves to our fellows, and he speculates that this may be an innate human tendency. (This point is made in C&J's penultimate paragraph, too, but there it is not presented as a response to the problem I am raising.) "Since we can no longer appeal to theological authority to resolve moral disputes, and since no body of traditional belief can now hope to secure universal assent, the only way in which we can have a chance of achieving moral consensus is through reasoned agreement" (Carruthers 1992: 44).

But this seems inadequate as an answer. I'll grant that all humans wish to justify their actions to others, and I'll also grant for the sake of argument that this may be an innate tendency. The problem concerns *whose* agreement we're interested in securing. The folks whose attitudes matter to a person, whose approval she would hope to obtain (if only tacitly) and whose disapproval would make her uncomfortable—to whom, in short, she would seek to justify herself—may be a very parochial bunch. If a Yanomamö tribesman slaughters an innocent stranger whom he chances upon in the jungle, he may be acutely concerned with how this decision will be received by his fellows back at the village, but he doesn't give a fig for whether he could justify it to his poor victim, or, for that matter, whether the action would win the approval of a group of rational agents who share the aim of reaching free and unforced agreement. He may very well explicitly deliberate about group stability and the avoidance of conflict (cf. Boehm 1999), but it will be the stability of *his* particular group that concerns him. If he became aware that his action will be highly acclaimed by his fellow

¹⁰ The complaint is not a million miles from that famously leveled by Ronald Dworkin (1975) against John Rawls's constructivism. "[Rawls's] contract is hypothetical, and hypothetical contracts do not supply an independent argument for the fairness of enforcing their terms" (pp.17-18). See also Ackerman 1980: 336-42; Brudney 1991.

tribesmen, but would be strongly disapproved of by a hypothetical group of rational agents sharing the aim of reaching agreement, it seems perverse to think that the latter justificatory framework provides a greater normative oomph lacking in the former (that it should be “something worth dying for”), or that abiding by the latter will enhance his fitness better than attending to the former—or, indeed, that the latter represents any kind of practical consideration for him at all.

To some extent, it seems, C&J will be unbothered by this, since they are willing to countenance the possibility of someone being under a moral obligation with which he has no reason to comply. But the extent of the problem may be far greater than they imagine. It is not merely the occasional psychopath who may have no reason to care about morality (by C&J’s lights), it may be anyone who cares more about the opinions of her friends, family, and colleagues—imperfect and fallible though she acknowledges them to be—than about the opinions of a bunch of non-existent rational agents struggling to construct a social contract. And that, it would seem, is the typical case. If it is her friends, family, and colleagues to whom she is concerned to justify her actions, then chances are she is in fact not concerned with the proclamations of hypothetical ideal agents *at all*, which is to say (by the moral constructivist’s lights) that she is in fact unconcerned with the moral realm *per se*. Even if the opinions of her real friends and family have the same extension as the hypothetical opinions of some ideal agents (specified in some determinate manner), in caring about the former she won’t be caring about the *moral wrongness* and *moral rightness* of actions (as defined by C&J). By analogy, an atheist who happens to think it a bad idea to covet his neighbor’s wife does not thereby care about God’s commands.

In sum: The kind of “hypothetical agreement” contractualism favored by C&J does not seem promising as a kind of moral realism. It may not yield a determinate output at all (if there is no fact about what “rational agents aiming at unforced agreement” would decide upon), or it may yield an entirely disagreeable output. It may demarcate a realm of facts that is not merely *contingently* connected to people’s reasons (which is a bullet C&J are willing to bite), but is such that people *typically* have no reason to care about it. When a property displays such an ill fit with our entrenched desiderata of what moral properties should be like, we have reason to reject the moral naturalism that champions it.

I must confess that I grow weary of attacking moral naturalism. Speaking as both a moral skeptic and an atheist, I find myself classifying defenders of moral realism along with apologists for theism (and I have never bothered to argue against theists). Both, to my mind, have about them an air of slightly desperate conservatism: an anxious determination to ensure that popular belief systems turn out as *true*. I do not accept as a general rule the orthodox methodological principles underlying such an approach; I do not think it the job of the philosopher to leave ordinary beliefs and attitudes as unruffled as he or she can. How much more invigorating philosophy might be if it *ruffled* us; how much more intriguing life might be if we opened our minds to the possibility that we’ve all been dramatically mistaken about the nature of the world.

REFERENCES

- Ackerman, B. 1980. *Social Justice in the Liberal State*. Yale University Press.
- Blair, R.J.R. 1995. A cognitive developmental approach to morality: Investigating the psychopath. *Cognition* 57: 1-29.
- Blair, R.J.R., Colledge, E., Murray, L., & Mitchell, D.G. 2001a. A selective impairment in the processing of sad and fearful expressions in children with psychopathic tendencies. *Journal of Abnormal Child Psychology* 29: 491-98.
- Blair, R.J.R., Monson, J., & Frederickson, N. 2001b. Moral reasoning and conduct problems in children with emotional and behavioural difficulties. *Personality and Individual Differences* 31: 799-811.
- Blair, R.J.R., Mitchell, D.G., Richell, R.A., Kelly, S., Leonard, A., Newman, C., & Scott, S.K.. 2002. Turning a deaf ear to fear: Impaired recognition of vocal affect in psychopathic individuals. *Journal of Abnormal Psychology* 111: 682-86.
- Boehm, C. 1999. *Hierarchy in the Forest*. Harvard University Press.
- Brudney, D. 1991. Hypothetical consent and moral force. *Law and Philosophy* 10: 235-70.
- Carlsmith, J. & Gross, A. 1969. Some effects of guilt on compliance. *Journal of Personality and Social Psychology* 11: 232-239.
- Carruthers, P. 1992. *The Animals Issue*. Cambridge University Press.
- Cleckley, H.M. 1941. *The Mask of Sanity: An Attempt to Reinterpret the So-called Psychopathic Personality*. St Louis, Mo.: C.V. Mosby Company.
- Dworkin, R. 1975. The original position. In N. Daniels (ed.), *Reading Rawls*. New York: Basic: 16-52.
- Freedman, J. 1970. Transgression, compliance, and guilt. In J. Macaulay & L. Berkowitz (eds.), *Altruism and Helping Behavior*. Academic Press: 155-161.
- George, A. (trans.). 1999. *The Epic of Gilgamesh*. Penguin.
- Hussain, N.J.Z. 2004. The return of moral fictionalism. *Philosophical Perspectives* 18: 149-187.
- Joyce, R. 2001. *The Myth of Morality*. Cambridge University Press.
- Joyce, R. 2006. Is human morality innate? In P. Carruthers, S. Laurence & S. Stich (eds.) *The Innate Mind: Culture and Cognition*. Oxford University Press.

Joyce, R. 2007. Morality, schmorality. In P. Bloomfield (ed.), *Morality and Self-Interest*. Oxford University Press.

Joyce, R. Forthcoming a. Is moral projectivism empirically tractable? *Ethical Theory and Moral Practice*.

Joyce, R. Forthcoming b. Moral anti-realism. *Stanford Encyclopedia of Philosophy*.

Ketelaar, T. & Au, W.T. 2003. The effects of feelings of guilt on the behavior of uncooperative individuals in repeated social bargaining games: An affect-as-information interpretation of the role of emotion in social interaction. *Cognition and Emotion* 17: 429-453.

Regan, J. 1971. Guilt, perceived injustice, and altruistic behavior. *Journal of Personality and Social Psychology* 18:124-132.

Shafer-Landau, R. 2003. *Moral Realism*. Oxford: Clarendon Press.

Sobel, D. 1994. Full-information accounts of well-being. *Ethics* 104: 784-810.

Tangney, J.P. 1994. The mixed legacy of the superego: Adaptive shame and guilt. In Bornstein (ed.), *Empirical Perspectives on Object Relations Theory*. Washington, DC: American Psychological Association: 1-28.

Tangney, J.P., Mashek, D., Stuewig, J., Magaletta, P., et al. 2007. Working at the social-clinical-community-criminology interface: The GMU inmate study. *Journal of Social and Clinical Psychology* 26: 1-28.

Tangney, J.P. & Dearing R. 2002. *Shame and Guilt*. New York: Guilford.

Tibbetts, S.G. 2003. Self-conscious emotions and criminal offending. *Psychological Reports* 93: 101-126.

Velleman, D. 1988. Brandt's definition of "good." *Philosophical Review* 97: 353-71.

Wittgenstein, L. 1965. Lecture on ethics. *Philosophical Review* 74: 3-12.

Yamauchi, E.M. (ed.). 2001. *Africa and Africans in Antiquity*. East Lansing, MI.: Michigan State University Press.

Zhong, C. & Liljenquist, K. 2006. Washing away your sins: Threatened morality and physical cleansing. *Science* 313: 1451-1452.