The nature of moral judgments and the extent of the moral domain

A key question for research on the evolutionary origins of morality concerns just what the target of an evolutionary explanation of morality should be. Some researchers focus on behaviors, others on systems of norms, yet others on moral emotions. Richard Joyce (2006) offers an evolutionary explanation for the trait of making moral judgments. Here, I defend Joyce’s account of moral judgment against two objections from Stephen Stich (2008). Stich’s first objection concerns the supposed universality of moral judgments as Joyce conceives of them. I respond by undermining the empirical evidence upon which this objection is based. Stich’s second objection concerns the extent of the moral domain, which he takes to include far more than the considerations of harm and fairness central to Joyce’s account. In response, I outline several strategies for reconciling Stich’s observations with Joyce’s account.

Keywords: moral judgment; moral psychology; moral/conventional distinction; meta-ethics; evolution of morality; Joyce, R.; Stich, S.

Introduction

Much interesting work has been done under the heading ‘evolutionary explanation of morality’. Many hurdles – both empirical and conceptual – confront work of this kind, one being the question of just what the target of such an explanation should be. Some researchers focus on certain kinds of behaviour (Wilson [1979]), while others consider group-level phenomena like systems of norms (Boyd & Richerson [2005]), and yet others devote their attention to features of human psychology, such as so-called “moral sentiments” (Frank [1988]) or an innate “moral grammar” (Hauser [2006]). Richard Joyce seeks to give an evolutionary explanation for “the trait of making moral judgements” (2006: 17). Joyce brings together evolutionary biology, moral psychology, and metaethics to argue for an ‘error theory’ of morality.

In this paper, I defend Joyce’s account of the nature of moral judgments against two objections ably pressed by Stephen Stich (2008). The first objection concerns the supposed universality of moral judgments as Joyce conceives of them. I respond by undermining the empirical evidence upon which the objection is based. The second objection concerns the extent of the moral domain, which Stich believes encompasses more than the considerations of harm and fairness central to Joyce’s account of morality. In response, I outline several strategies for reconciling Stich’s observations with Joyce’s approach to explaining the evolution of morality.
Joyce’s evolutionary genealogy of morality

Joyce makes two observations about the nature of moral judgments. One concerns the normative form of moral judgements, the other the content of moral judgements.

Regarding the normative form of moral judgements, Joyce claims that moral judgements possess, putatively at least, “inescapable authority” (2006: 191). Take inescapability first. Consider the moral judgements ‘keeping promises is obligatory’ and ‘stealing is forbidden.’ When we make such judgements, we mean them to apply to everyone regardless of individual wants or needs. The former applies even if one really doesn’t want to keep one’s promise and the latter applies even if one really needs to steal. In this respect, moral judgements are unlike judgements about what is prudent, which apply only in virtue of individuals having particular wants and needs. In this respect, moral judgements resemble judgements like ‘doing X is polite’ and ‘doing Y is pious’; such judgements are all, putatively at least, inescapable. What marks moral judgements off from judgements about politeness and piety is that moral judgements purport to give everyone a reason to comply with them: this is authoritativeness. Were one to guzzle communion wine at an Alcoholics Anonymous meeting, that act would be impious and rude to boot. If one were to cite one’s atheism and desire to be an iconoclast, that would not prompt others to retract their judgements about the impiety and rudeness of the act, but they would allow that such judgments provide no reason not to swill away. By contrast, we do not take there to be any way of legitimately ignoring moral judgements.

It is important to keep questions about the nature of moral judgments vis a vis their supposed inescapability and authority distinct from questions about whether the inescapability and authority of moral judgments can be substantiated or justified. My focus at this point is on the first kind of question: it is the character of moral thinking that is relevant to Joyce’s claim about the evolution of the capacity for such thinking. Joyce dubs the conjunction of inescapability and authority “practical clout” (2006: 58) and considers it an important feature of moral judgements for which an evolutionary explanation of morality must account.
Regarding the content of moral judgements, Joyce claims that “human morality takes interpersonal relations as its central subject matter” (2006: 66). Joyce insists that, insofar as one is interested in explaining the evolution of human morality, it is interpersonal relations, especially considerations of harm and fairness, which take centre stage. Moreover, he claims, “the actions that morality prescribes with categorical force [i.e. inescapable authority] are those that constitute or promote, roughly speaking, cooperation” (2000: 714).

Joyce’s explanation for why the trait of making moral judgements would have been fitness-enhancing for our distant ancestors is, in essence, that throughout human evolutionary history there have been various evolutionary pressures favouring cooperative, helpful behaviour, and that moral judgements, in virtue of their “practical clout,” help motivate such behaviour. This is not an original idea. Michael Ruse, for instance, has argued that “because it is biologically advantageous for us to help and co-operate, morality…has evolved to guide and stiffen our will” (1998: 222). Nor is the idea that helpful behaviour cannot be motivated by other-regarding emotions such as sympathy or even by cold calculation of self-interest. Joyce’s claim is just that moral judgements bolster motivations to be helpful, thus giving agents with the tendency to make them an edge over agents equipped only with a prudential concern for their own interests and with emotions such as sympathy. Joyce writes:

\[\text{An individual sincerely judging some action in a morally positive light increases the probability that the individual will perform that action (likewise, mutatis mutandis, judging an action in a morally negative light)}\] (2006: 109).

In essence, Joyce claims that selection for certain kinds of helpful behaviour could have also favoured the tendency to make moral judgements, since making such judgements makes acting in those helpful ways more likely. Although a great deal of empirical detail has been left out of this brief summary, that detail is not crucial for current purposes. My focus here is on Joyce’s account of the nature of moral judgement. Specifically, I will consider two criticisms of Joyce’s account, both from Stephen Stich.

Two criticisms
Stephen Stich (2008) raises two criticisms of Joyce’s approach to explaining the evolution of morality. Stich claims, firstly, that moral judgements as understood by Joyce are not human universals and, secondly, that Joyce’s focus on helpfulness in explaining the evolution of morality is overly narrow. Neither criticism can be easily dismissed. There are replies to be made, but developing them will take some effort.

**Criticism 1: the universality of ‘Joyce-style’ moral judgements**

It is central to Joyce’s account of the evolution of morality that moral judgements have clout, that the demands of morality are considered inescapable and authoritative. It is in virtue of their clout that moral judgements are motivationally efficacious in a way that could drive selection for the tendency to make such judgements. In response to Joyce, Stich has questioned whether the making of moral judgements as Joyce describes them is really a human universal. Stich voices his suspicion that:

…the practice of making moral judgements 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 (2008: 234).

Stich’s criticism of Joyce depends largely on a critique of psychological research in the moral/conventional task tradition (Smetana [1981]; Turiel [1983]). Research in that tradition is one of Joyce’s main lines of evidence for the claim that making moral judgements (as he conceives of them) really is a human universal. A brief overview of the moral/conventional task paradigm is thus in order.

Elliot Turiel (1983) proposed that moral rules and conventional rules are distinguished – in practice, by ordinary folk – along several dimensions: scope, dependence on authority, seriousness of violation, and grounds or justification. According to the Turiel tradition, moral rules are thought to have universal scope and to hold independently of the pronouncements of authorities. That is, moral rules are seen as holding in all places and times rather than just locally, and breaking moral rules is still seen as wrong even if doing so is explicitly permitted by an authority. Furthermore, violating moral rules is considered relatively serious, compared to
violating conventional rules (which will be detailed in a moment). Finally, Turiel claimed, justifications for moral rules typically appeal to the harms or injustice suffered by victims when those rules are violated. Turiel proposed that conventional rules, by contrast, are treated as having limited scope (holding only locally) and as authority-dependent (such that breaking the rule is not wrong if permitted by an authority). Violations of conventional rules are seen as less serious than moral violations and, lastly, justifications for conventional rules tend to cite ‘minimising disruption’ or ‘maintaining social order’ as relevant considerations rather than harm or injustice.

The moral/conventional task has been administered to a wide range of subjects. Subjects have varied in age, cultural background and religious affiliation, as well as mental capacity, including, for example, psychopaths and sufferers of autism.¹ Turiel’s proposals about the authority-independence and universal scope of moral rules have apparently received striking support (Smetana [1981]). Subjects presented with paradigmatic cases of moral wrongdoing – such as one child gratuitously pulling another’s hair in the classroom – judge those acts to be wrong and maintain that judgement even when asked to imagine that the act occurred in a different country or that an authority figure explicitly permitted the act. By contrast, when asked to judge paradigmatic cases in which conventions are violated – such as a child wearing pyjamas instead of a uniform to school – subjects allow that such acts might not be wrong everywhere and that such acts would not be wrong if they were permitted by an authority.

Turiel’s proposals regarding differences in the perceived serious of moral versus conventional violations and in the typical justifications offered for the two kinds of rules have also been supported by this large body of empirical work. The weight of evidence seems to indicate that transgressions involving harm or injustice evoke a “signature moral response pattern,” in which “rules are judged to be authority independent and general in scope; violations are more serious, and rules are justified by appeal to harm, justice or rights” (Kelly et al. [2007: 119-120]).

It is worth taking a moment to assess the relevance of work in the moral/conventional task tradition to Joyce’s claims about the normative form of moral
judgments. Joyce notes that work in that tradition, especially data showing the authority-independence and universal scope of moral rules, “meshes well” with his claims about the practical clout of moral judgements (2006: 136). However, Joyce’s notions of inescapability and authority cannot be straightforwardly equated with universality of scope and authority-independence. Even if the results of moral/conventional task research are broadly friendly to Joyce’s claim about the practical clout of moral judgments, those results do not establish that claim, and nor would undermining such results suffice to reject Joyce’s claim. Whether studies that specifically investigated inescapability and authority would produce results consonant with Joyce’s view is an interesting question that will hopefully be addressed by future research.

For current purposes, it is sufficient to note two things. Firstly, it is important to Joyce’s case – as Joyce himself presents it – that the moral/conventional distinction hold up. Secondly, whether that distinction holds up is also of more general interest. Stich’s denial that the tendency to make ‘Joyce-style’ moral judgements is a human universal relies heavily on empirical work by him and his collaborators (Kelly et al. [2007]) challenging the moral/conventional distinction. Outside of the disagreement between Stich and Joyce, the attack by Kelly et al. (2007) on the moral/conventional distinction has been taken seriously by several prominent philosophers and psychologists, including Haidt & Joseph (2007: 372), Prinz (2007: 282), and Doris & Nichols (forthcoming).

Kelly et al. (2007) deny that all transgressions involving harm or fairness evoke the signature moral response pattern described above. Kelly et al. open their case against the moral/conventional distinction with the observation that researchers working in the moral/conventional task tradition have presented subjects with a very limited range of stimuli. Moral/conventional task studies have typically asked subjects to judge schoolyard variety harms like hair pulling. Kelly et al. created more complex harm scenarios in order to test whether subjects’ responses to a wider range of stimuli would accord with previous findings.

Kelly et al. constructed contrasting pairs of harm scenarios that varied with respect to either where the harm described occurred, when the harm described
occurred, or whether the harm described was sanctioned by an authority. One of the scenario pairs used by Kelly et al. was “Whipping/Time” (2007: 123-4). This scenario pair was meant to test whether subjects’ moral judgements really were temporally universal in scope. Another scenario pair used by Kelly et al. was “Military/Authority” (2007: 125). This scenario pair was meant to test whether subjects’ moral judgements really were authority independent. Subjects were asked to indicate whether causing the described harm was OK or not, and also to rate the badness of the harmful behaviour.

Kelly et al. report a significant difference in the number of subjects who judged whipping to be OK in the past and the number of subjects who judged whipping to be OK now: “many subjects think whipping was OK 300 years ago but they do not think it is OK now” (2007: 126). Also, while more than half of subjects responded that physical abuse in military training was OK when not prohibited, very few judged it to be OK when prohibited, showing, Kelly et al. claim, that “many subjects do not judge the harmful transgression to be authority independent” (2007: 127). Kelly et al. thus claim a significant number of their subjects made moral judgements that were temporally local and authority dependent. Kelly et al. report that structurally similar scenario pairs produced similar results for geographical differences, with subjects making judgments that were not generalized to cover harms occurring in other places.

Kelly et al. claim their study casts doubt on the idea that there is a signature moral response pattern evoked by rule violations involving harm. In particular, moral judgements about harms of other than the simple schoolyard variety need not exhibit authority independence or universality of scope. Notice, while the harm scenarios Kelly et al. described were more complex than schoolyard variety ones, they were not entirely outlandish or even particularly far-fetched. If ordinary folks’ moral thinking about such cases does not exhibit authority independence and universality of scope, then Joyce’s claim that common thought and talk about morality supposes moral judgements to have ‘clout’ may well be imperilled.

Kelly et al. end their paper by asking:
Is there something special about these simple harm transgressions [i.e. the
‘schoolyard’ harms described in previous studies] that is not shared by the
more ‘grown-up’ transgressions that we also used in our study? (2007: 129)

There is a relevant difference between ‘schoolyard’ harm scenarios and those
that Kelly et al. used. Below, I will describe the difference and show how it allows for
a response to the claim that the Kelly et al. study undermines the moral/conventional
distinction. First, though, I must acknowledge a recent critical response to Kelly et al.,
and situate my discussion with respect to that response.

Paulo Sousa (2009) focuses on methodological issues concerning data analysis
in the Kelly et al. study. Sousa argues that the way in which Kelly et al. analysed their
data made their findings seem more striking than they really were. Specifically, Kelly
et al. compared the total number of ‘yes’ answers to the ‘Is it OK…?’ question for
each scenario with the total number of ‘yes’ answers to the ‘Is it OK…?’ question for
the contrasting scenario. Sousa points out that, if the question is whether subjects are
making authority-dependent or authority-independent moral judgments, then what
matters is how many subjects said that physical abuse in military training was not OK
when prohibited by authority and said that physical abuse in military training was OK
if it was permitted by authority. In Sousa’s re-analysis of the data, such subjects are a
minority for many of the scenario pairs. However, for some scenario pairs
administered by Kelly et al., subjects like that were common. That result requires an
explanation even if Sousa’s methodological criticism is granted. My discussion below
goes some way toward providing that explanation. If my explanation is correct, then
we need not suppose that the subjects in the Kelly et al. study in some cases made
moral judgments that were authority dependent or of limited scope.

My first point in responding to Kelly et al. is that moral/conventional task
studies typically describe harms that are very clearly transgressions, wrongful harms,
harms that ought not be inflicted on the victim. In the very simplicity of ‘schoolyard’
harms lies their strength as experimental probes. Early research on the
moral/conventional distinction used examples like pulling hair and pushing people off
swings precisely because those are paradigm cases of harmful moral violations.
Subsequent moral/conventional task studies continued to employ these simple cases
for that reason, as well as to allow comparison of results across different subject populations, such as preschoolers and psychopaths.

Of course, not all harms are considered to be transgressions. Some harms are seen as morally justified, such as those inflicted by doctors during treatment. Some harms may even be judged morally obligatory, such as those involved in meting out deserving punishment to wrongdoers. In general, whether a harm counts as a transgression will depend on the context in which the harm is inflicted. It thus seems sensible to ask whether, across the contrasting scenario pairs Kelly et al. constructed, it is the temporal differences (or the differences in authoritarian approval) per se that generate the differences in subjects’ responses.

Recall, a significant proportion of subjects in the Kelly et al. study judged that it was OK to whip derelict sailors 300 years ago but not OK to whip derelict sailors now and that physical abuse in military training is OK when not prohibited but not OK when it is prohibited. Recall also, Kelly et al. claim that what is going on here is that subjects are making moral judgements about harm that are (temporally) limited in scope and authority-dependent. However, the very complexity of the harm scenarios used may be part of the explanation for the pattern of results Kelly et al. found.

Suppose subjects make their particular moral judgements, such as ‘whipping Mr Williams in 2004 is not OK’ and ‘whipping Mr Williams 300 years ago was OK,’ by applying to the particular cases described a more general moral judgement to the effect that (roughly) ‘inflicting harm is permissible when necessary as means to achieving a sufficiently valuable end.’ That such a general moral judgement might be in play is quite plausible. After all, something very like it no doubt accounts for the observations above about common views on when dentists and executioners can permissibly inflict harm on others.

The general moral judgement about when harm is OK that I just sketched takes circumstances into account. Perhaps subjects reading about Mr Williams being whipped 300 years ago think that in those wild times on the high seas, whipping may have been the only punishment fierce enough to deter sailors from drinking on duty and thus the only way to safeguard the lives that could depend on a watchful lookout’s
warning. In today’s setting, by contrast, whipping is probably not seen as necessary to maintaining shipboard discipline. If this speculation is on track, then the general moral principle sketched above would generate different particular judgements when applied to the contrasting scenarios described in the study by Kelly et al.

I am tempted by a similar speculative account of what is going on when subjects judge that physical abuse in military training is OK when not prohibited but not OK when it is prohibited. Suppose again that subjects make their particular moral judgements by applying something like the general principle sketched above. Suppose in addition that subjects take the pronouncements of relevant authorities to be a good guide to the necessity or otherwise of certain harms in bringing about valuable ends. This further supposition is by no means unreasonable. Deference to (supposed) experts and epistemic division of labour is a feature of our everyday practices, in moral judgement no less so than in other arenas. If these two suppositions are correct, it would account for the responses of subjects who judged that physical abuse in military training is OK when permitted by military authorities but not when it is prohibited.

Notice that on my speculative account, it is not necessary to suppose that subjects are making limited-scope, authority-dependent moral judgements. It is not the differences in temporal location or authoritarian approval that make the difference to subjects’ responses, on my account. Those things are relevant, to be sure, but relevant to the application of a general moral judgement that we have no reason to doubt is considered universal in scope and independent of authority.

**Criticism 2: the extent of the moral domain**

As well as questioning whether moral judgements have the particular normative form Joyce supposes, Stich questions Joyce’s idea that “the primary sort of behaviour moral judgement was selected to motivate is cooperative or prosocial behaviour” (2008: 234). He notes the prevalence across different cultures and time periods of norms that seemingly have little to do with helping, harming or fairness, such as norms about what clothing can be worn, what foods can be eaten, how to show deference to high ranking people, and how to dispose of the dead. He then says:
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 (2008: 235).

This claim strikes me as deeply puzzling. According reciprocity pride of place in an explanation of morality despite the existence of many other kinds of norms would strike one as implausible (surely?) only if one took those other norms to also be moral norms. But Stich is very well aware that the issue of which norms are moral norms is a vexed one. Indeed, his discussions of this issue are nuanced and useful (Nado et al. [2009: 619-21]; Sripada & Stich [2007]). Stich does not assume that norms about clothing, food, corpses and so on are moral norms; he is officially agnostic on just how the category of moral norms is to be defined (and even whether such a category can be defined at all). So, since Joyce’s claim is not that reciprocity has pride of place in an explanation of norms simpliciter, it is not clear to what Stich is objecting here.

What reason is there to think that morality is about more than interpersonal relations, including especially issues of harm and fairness? Perhaps a good place to begin this discussion is with an examination of the empirical work meant to support the claim. A key figure here is Jonathan Haidt. Haidt identifies four moral “domains,” by which he seems to mean ‘broad classes of actions that are treated as morally evaluable.’ These are the domains of suffering (concerned with harm), reciprocity (concerned with fairness), hierarchy (concerned with respect and loyalty), and purity (concerned with pollution). Only the first three of these four domains are concerned with interpersonal relations, which, as Haidt notes, may strike one as strange:

There is an odd corner of moral life, odd at least for modern Westerners, who tend to think of morality as strictly concerned with how we treat other people. That corner is the profound moralisation of the body and bodily activities, such as menstruation, eating, bathing, sex, and the handling of corpses (Haidt & Joseph [2004: 60]).
Haidt claims that there are cultures in which morality is not focussed centrally on issues of harm or fairness, being based instead around the ideas of purity and pollution. Haidt et al. (1993) studied subjects of different cultures and socioeconomic standing (SES) in order to determine whether those factors influenced the “domain of morality”. In particular, the study investigated subjects’ responses to harmless but disgusting or disrespectful actions. To remove harmfulness as a confound, subjects were asked whether the act they were being asked to judge was harmful, and only data from subjects who judged the acts in question to be harmless was considered.7

It is crucial to be clear about how Haidt et al. operationalised what it is to include an act in the moral domain. Haidt et al. contrasted two “stances” that might be adopted toward an act: a “moralising” stance and a “permissive” stance. To adopt a moralising stance toward an act is to “endorse interference” and to “universalise” one’s judgement that the act is wrong. That is, to moralise an act is to judge that people should be prevented from performing the act and punished if they do so and, furthermore, to judge that performing the act is wrong everywhere, not just locally (1993: 613). To adopt a permissive stance is to fail to make one or both of those judgements about the act.

On Haidt’s account, one would count as taking a permissive stance toward an act if one judged that it is wrong and should be prevented or punished but only locally. Suppose a headhunter believes that his fellow tribesmen must take heads after battle and should be punished (perhaps by beheading) for failing to do so. Suppose he also believes that the foothunters from across the valley are under no such obligation and do not deserve punishment for failing to take heads after battle. This fellow would count as permissive with respect to taking heads. This would probably come as a shock to the folk he has beheaded for failing to take heads, were those folk in a position to be shocked by anything. Calling this headhunter ‘permissive’ is terminologically jarring at least. There may also be a deeper problem, though. Haidt may be ruling out the very possibility of moral relativism being true. If one thinks the wrongness of acts is always a local and culturally-specific phenomenon, one cannot count as moralising by Haidt’s definition. Haidt calls the moral relativist-style response pattern (i.e. endorsing interference without universalising) an “enforceable-conventional” stance, which contrasts with “fully moralised” (endorsing interference
and universalising) and “personal morality” (universalising without endorsing interference) stances, suggesting he really might not see it as any kind of moralising at all (Haidt et al. [1993: 622]). Moral relativism may be false, but it ought not, by definitional fiat, fail to count as a moral view at all.

Having given their account of moralising, Haidt et al. aimed to determine what kind of stance subjects would adopt toward the paradigmatic moral and conventional transgressions used in the moral/conventional task, as well as in response to scenarios describing disrespectful or disgusting actions. They found that culture and SES did have an effect on the domain of morality. Specifically, Brazilian more so than North American subjects and low-SES more so than high-SES subjects moralised harmless but disgusting or disrespectful actions. Haidt takes this experimental work, along with the observations of cultural anthropologist Richard Shweder (e.g. Shweder [1991]; Shweder et al. [1997]), to support the conclusion that there are:

…cultural differences in the domain of morality. There does not appear to be a single list of content areas—even defined abstractly as harm, rights, and justice—that can capture the moral world of all peoples (Haidt et al. [1993: 625]).

It is time to step back and assess the bearing of Stich’s claims (against the background of experimental work like that of Haidt) on Joyce’s account of the evolution of morality. Joyce does focus on the fitness benefits of certain kinds of helpfulness in giving his evolutionary explanation of morality. While he is too cautious to rule out the possibility of moralities with other contents, he does say that concerns about interpersonal relations, especially involving harm and fairness, make up the bulk of “actual human morality” (2006: 66). If it turns out that helpfulness is but one small corner of morality, explanations for the evolution of morality that focuses mostly on helpfulness might be thought importantly incomplete.

In what follows, I will discuss several strategies for defending Joyce’s approach to explaining the evolution of morality. These are the foot-stamping strategy, the redescription strategy, and the extend-the-strategy strategy. The upshot
of my discussion is that such an approach is ultimately defensible, though much depends on empirical questions that are currently open.

*The foot-stamping strategy*

Stich suggests that interpersonal relations might be only part (and a small part at that) of the content of morality. Haidt’s work can convey the same impression. One strategy in response is to insist that judgements (or norms, systems of rules, etc.) simply must have a certain content in order to count as moral judgements. Neil Levy can be read as adopting something like this strategy. Levy claims that moral systems:

…must be devoted, largely if not wholly, to concern for the welfare of other people [and] must systematise norms of justice and fairness… If it does not have the right kind of content, we should be reluctant to call any system of prescriptions a morality (2004: 44).

Judgements about purity, pollution, sacredness and profanity are, according to this strategy, just of the wrong kind to count as moral judgements. This is not to deny that what is eaten, what is worn, how the dead are treated and so on matter to many people, or to say caring deeply about such things is silly or primitive. It is, however, to exclude such concerns from the domain of morality proper. This might be termed the foot-stamping strategy.

Adopting the foot-stamping strategy seems to me unwise at this stage in our understanding. Joyce is far from alone among philosophers in suggesting that the distinctive normative form of moral judgements lies in their putative inescapability and authority. Admittedly, studies in psychology have not investigated judgements with exactly that normative form. Still, what psychologists have found makes it seem not unlikely that future work might turn up folks who judge that performing certain acts is self-polluting and wrong regardless of what one happens to believe, need or desire, and who take that supposed fact to give everyone reason to avoid such acts. It seems a live possibility that some folks might make judgements that have the (supposedly) distinctive normative form of moral judgements but that concern subject matter far removed from harm, fairness or any kind of interpersonal relations. I think
that if one were to now stamp one’s foot about constraints on the content of morality, one would be in the same motion sticking one’s neck out for the empirical axe, which may not have far to fall.

The redescription strategy

An alternative strategy to simply denying that concerns about purity and pollution count as moral is to redescribe such concerns in terms of more familiar, interpersonal moral concerns, especially the concern with harming others. This seems to be Joyce’s favoured strategy for dealing with Stich’s challenge regarding the content of morality. Joyce warns against “assuming that what we in the West might count as a purely self-regarding action should be categorised as such in another culture” (2008: 253).

I call this the redescription strategy. It opens with the observation that, although some experimental work has controlled for the effect of harm judgments on responses to disgusting transgressions, examples of supposed purity-based moralities are often drawn from broadly religious contexts. Haidt, for instance, supports his claim about cross-cultural variation in the moral domain by noting that “a great deal of the moral law of Judaism, Hinduism, Islam, and many traditional societies is explicitly concerned with regulating purity and pollution” (Haidt & Joseph [2004: 60]).

While gods (or ancestral spirits, as the case may be) are not typically seen as subject to brute physical harm, it is typically thought that such beings can be pleased or displeased. Violating a purity-based prohibition on eating certain foods might seem like a purely self-regarding violation from the standpoint of one who does not accept the religious framework within which the prohibition is embedded. For a believer, though, who thinks such a transgression offends the gods (ancestors, etc.) by flouting their wishes, such a violation may not be seen as purely self-regarding at all. Purity-based rules like this may well be experienced as interpersonal, relating an agent not to other human persons but to supernatural persons (who, moreover, occupy particularly important positions in the agent’s worldview).
If the redescription strategy as executed above seems like too much of a stretch, consider the following variation. Notice that polluting transgressions are typically taken to have bad consequences (sometimes quite severe ones) for the transgressor’s community, should the proper purification rites not be performed. Viewed in this light, seemingly strange injunctions to avoid certain foods (for example) can be recognised as more familiar injunctions to avoid bringing harm to others. Purity-based rules might even be redescribed as straightforward prudential advice, since the transgressor himself is typically believed to suffer harm of some kind as a result of being polluted. Consider the following from anthropologist Mary Douglas, who studied the Yurok Indians of Northern California:

The Yurok so much believed that contact with women would destroy their powers of acquiring wealth that they held that women and money should never be brought into contact. Above all, it was felt to be fatal to future prosperity for a man to have sexual intercourse in the house where he kept his strings of shell money. In winter, when it was too cold to be out of doors, they seem to have abstained altogether… [It is] significant for understanding their idea of female pollution that for the Yurok men there was a real sense in which pursuit of wealth and of women were contradictory (2002/1966: 189-90).

One might worry that the redescription strategy does not take seriously enough the idea that purity and pollution as such are central to some cultures’ worldviews. Haidt would likely grant that actions that defile places, people, and objects are often believed to have harmful consequences but insist that, for participants in purity-based norm systems, the pollution caused by such acts figures as a deliberative consideration in its own right. The redescription strategy, on this kind of view, just reflects educated, Western, liberal biases (Haidt [2007: 1001]).

Empirical work teasing apart considerations of purity and pollution from harms to self and others (human and otherwise) would be invaluable at this point. To my knowledge, though, such work has not yet been conducted, and the methodological challenges involved in carrying it out would be daunting. Consider: asking an Oriya man in India whether touching a menstruating woman would be
wrong if doing so carried no risk of harm to self or others nor did it offend anyone (gods included) may be tantamount to asking him to entertain a conditional with an impossible antecedent: ‘If that were so,’ he may think, ‘who knows what would be right or wrong?!’ This problem worsens if we suppose impurity is seen not just as instrumentally bad but also as intrinsically harmful. Being an educated, liberal Westerner, I find it hard to get into the head of a participant in a purity-based norm system. For all I know, someone with that mindset may think that to be made impure (by the touch of a lower-caste individual, for instance) is to suffer a kind of harm. If this were so, it would make experimentally teasing concerns about purity apart from concerns about harm even harder.

The worries above aside, some concerns about purity and pollution should probably be expected to resist redescription in terms of harm or other interpersonal factors. After all, one of the striking findings of Haidt *et al.* (1993) was that actions judged to be harmless but disgusting triggered the same sort response as was triggered by the prototypical harmful transgressions used in the moral/conventional task. The next strategy I will discuss is not an alternative to the redescription strategy but rather meant to complement it.

*The ‘extend-the-strategy’ strategy*

The ‘extend the strategy’ strategy concedes for argument’s sake that morality encompasses issues of purity and pollution that cannot be redescribed in terms of harm or other interpersonal concerns. This strategy seeks to explain moral thinking about these matters in the same kind of way that Joyce explains thinking in moral terms about helpfulness. That is, the strategy seeks to establish that there are fitness benefits to behaving in the ways effectively motivated by moral judgements concerning purity and pollution. In essence, this strategy depends on treating Joyce’s focus on helpfulness as dispensable to an evolutionary explanation of the trait of making moral judgements. The evolution of that trait is to be explained by reference to its efficacy in motivating (by and large) fitness-enhancing behaviours, but if helpful behaviours turn out to be only one among several kinds of fitness-enhancing behaviours motivated by moral thinking, that does not invalidate the general approach.
Some care is needed in employing this strategy. It will not do to cite modern examples like the Jewish *kashrut* – dietary rules – prohibiting pork consumption together with the supposed health risks of eating pigs (see Simoons [1994]). It is not the current utility (in evolutionary terms) of specific norms about purity and pollution that is relevant, but the utility to our distant ancestors of having the capacity for moralising about such matters. Also, the evidence needed to substantiate claims about how making purity-based moral judgements provided fitness benefits to our distant ancestors is likely to be scattered throughout the literatures of numerous disciplines, if it is available at all. I will not develop this strategy in any great detail here. I will only indicate some of the issues that would need to be confronted in doing so.

One may worry that even if some supposed pollution-avoidance is susceptible to evolutionary explanation in the way just sketched, much of it seems to carry no fitness benefits or to be actively detrimental to one’s reproductive prospects. For example, recall the Yurok belief that bringing women and money into contact or even proximity can destroy one’s ability to accumulate wealth. Judging by the distribution of tribesmembers’ birthdays throughout the year, this belief is apparently sufficient to make many Yurok men abstain from sex during the winter months when they, their wives, and their money are all trapped inside together (Douglas [2002/1966: 189]). A purity-based moral judgement that motivates one to go months without sex seems unlikely to advance one’s reproductive fitness. I will not multiply examples here, but will instead rest content with the assertion that a trawl through the purity-based prohibitions and requirements of various cultures would net many examples that support this worry.

One may also worry that where supposed pollution-avoidance *does* plausibly secure some fitness benefit, it is hard to see how one’s motivation to shun supposedly polluting thing(s) would need bolstering (or at least, harder than it is to see how one’s motivations to help others might need it). For example, some of the most promising cases for the extension strategy are certain food taboos, rules about ritual washing, and regulations governing disposal of the dead. Preserving personal and local purity in such cases may incidentally secure significant health benefits. However, one’s motivations to clean up any corpses lying about, for instance, or to avoid the sacred
mushrooms that make the shaman spasm and froth, are probably fairly robust. Strong emotional responses such as disgust and fear are effective motivators that require little in the way of bolstering from clouty moral judgements.

The two worries just laid out present the extension strategy with something of a dilemma. Some remarks can be made to blunt the horns, though.

First, the success of the strategy does not depend on every purity-based norm having a plausible adaptive rationale, just as the success of Joyce’s initial case does not depend on every morally motivated instance of helpfulness being fitness-enhancing. A key question as far as this strategy is concerned is just what adaptive benefit might accrue to having a faculty for moralising that is sensitive to considerations of purity and pollution. That is a currently open empirical question.

Also on this point, where one looks for a fitness benefit of purity-based moralising is crucial. I have so far been considering the possibility that making purity-based moral judgements might benefit individuals. The fitness benefits of purity-based moralising may also (or mostly) accrue to cultural groups. Shared purity norms are important markers of group identity that may strengthen group cohesion (Bowie [2000: 73]; Meyer-Rochow [2009]). While considerable empirical detail would need to be filled in to make the extension strategy work, the prospects for explaining purity-based moralising in terms of its adaptive benefit is not as unpromising as the first worry above may suggest.

Second, while the motivational boost provided by purity-based moralising may well be superfluous when supposedly polluting things trigger strong emotional responses, such responses need not always be triggered by such things. Sticking with the example of food taboos, Begossi et al. (2004) report that among certain Amazonian tribes, piscivorous (i.e. fish-eating) fish are a taboo food for pregnant women. Such fish are not in and of themselves particularly repulsive. They may even be tasty. As it turns out, however, these fish contain relatively high concentrations of toxins and contaminants, in virtue of being near the top of their food chain (Begossi et al. [2004: 1341]). In cases like this, the motivation to avoid forbidden foods may well sometimes need boosting. If judging in moral terms achieves this, then doing so
would help secure the benefits (in this case health benefits) of refraining from eating forbidden foods.

Also on this point, it seems worth mentioning that disgust responses and moral judgements are not entirely separate; the latter can help create the former. Suppose a fundamentalist Christian happened upon two men holding hands outside a hospital. Imagine she learned that the two men are friends, one of whom is recently bereaved. Now imagine instead that she learned the two men are lovers awaiting the decanting of their first cloned child. I am willing to bet that in the second case, a strong disgust reaction would follow hard on the heels of the Christian’s realisation that she had happened upon homosexuals playing God. Insofar as moralising actions can create disgust responses (or prompt other strong emotions), the second worry raised above may be less damaging to the extension strategy than it seems.

There is one more reply I think can be made in response to Stich’s comments about the extent of the moral domain and the importance of helpfulness in a Joyce-style explanation for the evolution of morality. This final reply is less of an argumentative strategy and more of a straightforward observation: call it the ‘origin vs. elaboration’ point. It is important to note that the claim Stich attributes to Joyce – that the selective pressures favouring the capacity to make moral judgments sprang mostly from the fitness benefits of certain kinds of helpfulness – is compatible with there being many norms having nothing to do with interpersonal relations, even granting that such norms are genuinely moral norms. Joyce’s claim concerns the initial driving forces in the evolution of morality. Making it does not commit one to the claim that the domain(s) of action it was fitness-enhancing for our distant ancestors to judge in moral terms are co-extensive with those over which the human capacity for moral judgement is now exercised. Of course, giving an account of how and why the moral domain expanded would not be trivial. The bare logical point stands, though: even indisputable evidence of cross-cultural differences in the domain of the moral would not sink a Joyce-style account of the evolution of morality.

Conclusion
Richard Joyce takes the appropriate explanatory target of an evolutionary account of morality to be the trait of making moral judgements. We want to know why we have evolved to judge in moral terms. Joyce’s answer is that doing so – in virtue of the normative form and typical content of moral judgments – got our ancestors to act in fitness-enhancing ways. I have defended Joyce’s approach to explaining the evolution of morality against two objections ably pressed by Stephen Stich. The first concerned the supposed universality of moral judgments as Joyce conceives of them. I have responded to it by undermining the empirical evidence upon which it is based. We can interpret Kelly et al.’s results in a way that leaves the moral/conventional distinction intact. The second objection concerned the extent of the moral domain. My response to this objection was to outline several strategies for reconciling Stich’s observations with a Joyce-style approach to explaining the evolution of morality. While much in the way of empirical detail remains to be filled in, the strategies outlined – especially the redescription and extension strategies – show how this approach can be plausibly maintained. [7414 words]

Notes

1. See Kelly & Stich (2008: 354) for a review and further references.
2. Subjects who received this scenario first saw a version that ran as follows:

   Three hundred years ago, whipping was a common practice in most navies and on cargo ships. There were no laws against it, and almost everyone thought that whipping was an appropriate way to discipline sailors who disobeyed orders or were drunk on duty. Mr. Williams was an officer on a cargo ship 300 years ago. One night, while at sea, he found a sailor drunk at a time when the sailor should have been on watch. After the sailor sobered up, Williams punished the sailor by giving him 5 lashes with a whip.

   Subjects were then asked to respond to a different version:

   Mr. Adams is an officer on a large modern American cargo ship in 2004. One night, while at sea, he finds a sailor drunk at a time when the sailor should have been monitoring the radar screen. After the sailor sobered up, Adams punished the sailor by giving him 5 lashes with a whip.

3. Subjects given this scenario were first shown the following material:

   For many years, the military training of elite American commandos included a simulated interrogation by enemy forces in which the trainees were threatened and physically abused. Most people in the military believe that these simulated interrogations were helpful in preparing trainees for situations they might face later in their military careers. Though no one was ever killed or permanently disabled by the physical abuse they received during these simulated interrogations, the trainees often ended up with bruises or injuries that lasted for a week or more. Recently, the Pentagon issued orders prohibiting physical abuse in military training. Sergeant Anderson is a soldier who trains elite American commandos. He knows about the orders prohibiting physical abuse and his immediate superiors have ordered him not to do it. Nonetheless, he regularly threatens and physically abuses trainees during the simulated interrogations that he conducts.

   Subjects were then shown this:

   Now suppose that the Pentagon had never issued orders prohibiting physical abuse in military training, and that Sergeant Anderson’s superiors had told him that the use of physical abuse was acceptable in simulated interrogations.
4. For example, in the ‘Military/Authority’ scenario pair, Kelly et al. report that only 9% of subjects said physical abuse in military training was OK when it was prohibited but the percentage of subjects saying physical abuse in military training was OK ‘jumped’ to 58% when physical abuse in military training was permitted by military authorities.

5. See e.g. Haidt et al. (1993); Haidt (2001); Haidt & Joseph (2004), (2007); Haidt (2007).

6. Haidt sometimes mentions five domains (e.g. Haidt [2007: 1001]): harm/care, fairness/justice, in-group/loyalty, authority/respect, and purity/sanctity.

7. In this respect, Haidt’s work is superior to that of Shaun Nichols (2002), who surveyed subjects’ responses to moral, conventional, and disgusting transgressions but did not include such a question in his study design, leaving open the possibility that harmfulness played an unrecognised role in generating the pattern of results he found.

8. Attridge (2004: 71-76) gives many examples of supposedly polluting transgressions, the dangers such pollution was believed to pose, and the corresponding purification rites.

9. Although, Ogbeide (1974) is an excellent example of the kind of empirical work needed to assess the fitness benefits or drawbacks of purity-based norms. Ogbeide details the negative nutritional impact of food taboos on children and pregnant women in traditional Nigerian societies.

References


