How Neuroscience Can Vindicate Moral Intuition

Christopher Freiman
College of William & Mary
Department of Philosophy

Ethical Theory and Moral Practice 18 (2015): 1011-1025

Imagine that an anthropologist returns from her study of a group of people and reports the following:

- They refuse to kill one person even to avert the death of all involved—including that one person;
- They won't directly push someone to his death to save the lives of five others, but they *will* push a lever to kill him to save five others;
- They punish transgressors because it feels right, even when they expect the punishment to cause far more harm than good—and even when the harm done by the punishment exceeds the harm done by the transgression being punished.

The anthropologist's report might lead us to conclude that these people are at least confused, and perhaps even dangerous.

Here's some bad news. Those people are *us*. Or so suggests recent research in experimental psychology and the neurosciences. This research indicates that our moral intuitions have a vaguely deontological character and they prompt us to make any number of judgments that appear arbitrary or otherwise unjustified, such as those recommending the behaviors above. Our intuitions are allegedly the product of morally-insensitive evolutionary processes—they are emotionally-driven heuristics adapted to help our ancestors procreate, not to help us grasp the nuances of moral decision making. Our moral reasoning, in contrast, is characteristically utilitarian and appears sensitive to a rich complex of moral considerations. When our moral judgment is cognitive rather than affective, we gather information, balance competing concerns, weigh costs and benefits, and so on.

These findings have led moral theorists such as Peter Singer (2005) and Joshua Greene (2008) to doubt the credentials of our so-called "deontological" intuitions.[1] (I'll simply follow precedent by labeling the relevant intuitions "deontological" while acknowledging that they're unlikely to closely track the contours of a deontological moral theory.) If one accepts the preceding evolutionary account of moral judgment, the above conclusion seems compelling: our moral intuitions are attuned to matters of adaptive—not moral—significance. We therefore have good reason to deny that our intuitive moral judgment tracks the correct criterion of moral rightness.

Much of the philosophical work on the new research on moral judgment addresses this issue—viz. whether the research supports the claim that utilitarian reasoning is more likely to produce correct moral judgments than types of moral decision making that involve deontological intuitions.[2] I wish to address a different question. Suppose we simply grant, for the sake of argument, that we should prefer utilitarian reasoning in our search for the correct moral standard. What does this mean for our moral *practice*?

In brief, recent research in experimental psychology and the neurosciences provides information that helps vindicate moral intuition as a utilitarian decision procedure for personal conduct. Greene's most recent work advises that people generally trust their intuitions in local, everyday contexts but shift toward critical moral reasoning when faced with unfamiliar or controversial cases—particularly those concerning large-scale institutional problems (2013, 347ff). Our intuitions help us solve a variety of "micro-level" strategic problems but can produce conflict when relied upon to settle contested policy questions.

[1] See also Sinnott-Armstrong (2006).
[2] For arguments in this spirit, see Berker (2009) and Kamm (2009).

Although I believe this kind of view has much to recommend it, I contend that Greene partly misunderstands the practical implications of his own principles. If our ordinary moral judgments are to do the strategic work Greene wants them to do, he needs to endorse full-fledged Sidgwickian self-effacement for at least some areas of micro-level decision making. By the lights of some of Greene's own arguments, people must accept the correctness—and not simply the *usefulness*—of the relevant intuitions in their personal conduct to satisfy utilitarian standards. I argue that removing utilitarian reasoning from micro-level decision making is consistent with Greene's preferred strategy of using utilitarianism as a "common currency" for resolving moral conflict at the institutional level. To clarify upfront: the aim of this paper is to offer an internal criticism rather than to defend (e.g.) the relevant empirical research, utilitarianism, self-effacement, and so on. I'm making an argument about where Greene's own psychological and philosophical commitments should take him.

§1

Put briefly, Joshua Greene's *dual process* model of moral decision making indicates that characteristically deontological intuitive judgments are generated by automatic, emotionally-laden processes, whereas characteristically utilitarian judgments are generated by deliberate cognitive processes (Greene, 2008).

Before proceeding, let me explain my terms and my aims for this section. First, as noted earlier, my use of "deontological" simply follows the precedent in the relevant literature on moral psychology and is not intended to suggest that most deontological moral theories will endorse the intuitive judgments in question. Second, my purpose in this section is to briefly summarize the relevant research and explain how it features in

arguments purporting to debunk the normative authority of our deontological moral intuitions.

Perhaps the most celebrated finding concerns the "trolley problem" series of thought experiments.[3] Consider a case we can call TROLLEY. Imagine a trolley is speeding down a track on its way to killing five people. You can switch the trolley to a different track, but there is one person on that track. Most say that you should switch the track, killing one and saving five. In the FOOTBRIDGE variation, you are on a bridge above the track, standing next to a very large man. You can stop the trolley by pushing him onto the track. Most say that you should not push the man off the bridge.

There may be a principled moral difference between these two cases. In FOOTBRIDGE, but not TROLLEY, you use someone as a mere means. However, imagine a third variation: throwing the switch diverts the trolley around a loop before it will kill the five people.[4] Lying on the loop is a very large man whose body would stop the trolley. Throwing the switch uses this man as a mere means to saving the lives of others, but most judge it right to do so.

Greene and others used functional Magnetic Resonance Imaging (fMRI) to study why people make these conflicting judgments. Greene found that asking people for moral judgments about "personal" violations (e.g., directly pushing someone from a bridge) showed increased activity in the areas of the brain associated with emotion compared to moral judgments about "impersonal" violations (e.g., throwing a switch that would lead to someone's death) (2008, 43). He also found that even those who judged "personal" violations to be right would still experience emotions countervailing their cost-benefit

[3] For discussions of the trolley problem, see Foot (1967) and Thomson (1976).
[4] See Thomson (1985, 1402).

reasoning. People seem to have made utilitarian judgments in spite of their affective responses and thus took longer to arrive at their answer than those who made the deontological judgment (Greene 2008, 43). Further experimental results appear to corroborate the model.[5]

Emotion appears to drive deontological approaches to punishment as well. Retributivism holds that punishment should give wrong-doers what they deserve. Utilitarianism holds that punishment should produce beneficial effects. In practice, people tend to be retributivist. People generally punish in proportion to the extent to which transgressions anger them, rather than according to the punishments' future effects.

Consider a neuroimaging study of the ultimatum game (Sanfey et al., 2003). In an ultimatum game, the first player proposes how to split a sum of money with the second player. If the second player accepts the offer, the money is divided as proposed. If the second player rejects the offer, neither receives any money. Responders tend to reject offers that are unfair (i.e., notably less than a fifty-fifty split)—even when the game is played only once. Yet the decision to take no money over some money is puzzling. In one-off games, punishment will not have a beneficial impact on players' future behavior. In effect, responders pay to punish unfair offers. Individuals appear motivated to punish for its own sake, a tendency driven by emotion: experimenters found that unfair offers produced increased activity in the brain region associated with anger and disgust (Greene

[5] For example, similar results were reported for variations on the "crying baby" thought experiment. See Greene (2008, 44).

2008, 54). Additional experimental evidence confirms individuals' retributivist preferences.[6]

What explains the idiosyncrasies of our moral psychology? Theorists such as Singer (2005) and Greene (2008, 59ff) speculate that the moral emotions benefited our ancestors in the *environment of evolutionary adaptedness* (EEA). Our adaptive moral behavior is driven by emotion rather than reasoning because emotions can be more reliable, expedient, and efficient than reasoning in recurring contexts (Greene 2008, 60). Given that cooperation requires a systematic constraint on violent behavior, it is more efficient to experience a feeling of absolute prohibition against inflicting harms than to puzzle out the right action on a case-by-case basis. This evolutionary explanation also purports to account for our negative emotional responses to personal, but not impersonal, forms of harm: face-to-face violence dates to the EEA, but impersonal violence does not. Moreover, a disposition to punish even when punishing is costly can be an effective deterrent against aggression given iterated interactions.[7]

Moral philosophers including Singer and Greene suggest that this new understanding of moral psychology debunks the normative authority of our intuitions.[8] Singer writes that advances in our understanding of ethics due to empirical work in moral psychology "do not themselves directly imply any normative conclusions, but they undermine some conceptions of doing ethics which themselves have normative conclusions. Those conceptions of ethics tend to be too respectful of our intuitions" (2005, 349). The evolutionary account purportedly explains why moral intuition but not explicit moral

[6] See, e.g., Baron and Ritov (1993).
[7] On this idea, see Mackie (1985, 215ff).
[8] See Singer (2005) and Greene (2008). See also Sinnott-Armstrong (2006, 351-2). Sinnott-Armstrong stresses that we ought to reserve final judgment pending replication, and further interpretation, of the experimental results.

reasoning (which readily adapts to new circumstances) seems inappropriately responsive to considerations that are arbitrary from a moral point of view (Singer 2005, 348). Given that they "reflect the influence of morally irrelevant factors," our intuitions are ill-equipped to track matters of moral significance (Greene 2008, 70). A 'cosmic coincidence' looms: it would be an incredibly (literally) lucky accident if the outputs of evolutionary processes were attuned to an independent order of moral facts.[9] If successful, the argument indicates that we have a reason to doubt that our native deontological intuitions reliably track the moral truth, a reason that allegedly does not apply to utilitarian reasoning.

Legitimate doubts may persist about the preceding attempt to "debunk" our deontological intuitions.[10] However, these doubts will remain unexplored here because I will simply grant that our deontological intuitions fail to track the moral truth. My interest centers on the role these intuitions should play in our practical judgment if we accept, with Greene and Singer, a utilitarian criterion of rightness.

At first blush, the new model of moral psychology appears to spell trouble for utilitarianism as a decision procedure. Our brains have specific adaptations tuned for life in an environment that differs strikingly from our own. For instance, interactions in the EEA tended to be iterated and confined to a small local circle of familiar participants. Now our interactions are generally not iterated, our trade partners are unfamiliar, and our reach is (potentially) global. In brief, the heuristics of moral intuition, while once useful, seem obsolete.

[9] See Greene (2008, 69-70). See also Street (2006) and Joyce (2006, chapter 6).
[10] For some reasons to doubt the success of this kind of debunking argument see, e.g., Berker (2009).

But Greene does not want to dispense with intuitive moral judgments entirely. He thinks they are quick and effective means of solving cooperation problems like the tragedy of the commons (2013, 15). These same intuitions, however, are ill-equipped to solve the problem he calls the "tragedy of commonsense morality"—namely, deep social conflicts due to differences in intuitions about controversial large-scale moral problems (Greene 2013, 4). So he wants to reserve a (limited) role for intuitions in moral decision making. In the next section, I'll explain why Greene thinks that reliance on our intuitions can foster cooperation and where I believe his analysis errs.

§2

Greene argues that some of our "moral instincts"—including empathy, reciprocity, and vengeful punishment—help us solve public goods problems. It's worth noting that utilitarians like Greene probably wouldn't want us to retain *all* of commonsense morality, even at the micro-level, given that it likely has elements that frustrate cooperation or produce other bad outcomes.[11] Due to space constraints, I'll only focus on some of the specific intuitions discussed by Greene and set aside the issue of those chunks of commonsense morality unlikely to pass utilitarian muster.

Let's start with punishment. As the discussion in the previous section indicated, our disposition to punish is often not motivated by calculated, direct reciprocity. That is, we don't simply punish an uncooperative player with the aim of deterring her from being uncooperative with us in the future. Rather, people often punish defection without receiving a benefit from that punishment in the future (Greene 2013, 56). Moreover, this punishment need not be motivated by explicit consideration of its social benefits but rather because people like the feeling of punishing wrongdoers (Greene 2013, 56).

[11] Indeed, Greene (2013, 63ff) makes comments along these lines.

Let's look at a real-world example of commonsense punishment norms. The so-called *fundamental problem of exchange* is a matter of critical significance in modern commercial economies (Greif 2000). Trade is typically anonymous and one-off, as participants do not know each other, nor is there a promise of repeated interactions. Sanctions such as ostracism or future non-cooperation are therefore unlikely to deter cheating. A commitment problem arises because exchange is sequential—time lapses between the provision of a good or service and the payment of compensation.

Exchange resembles an assurance game: the player who moves first must decide whether to trust the second player before presenting her with the opportunity to cooperate or defect. In small communities, direct reciprocity may be enough to sustain cooperation. Individuals can know the reputation of potential cooperators, and punishments like ostracism or future non-cooperation can effectively deter defection. Yet in large-scale economies, the dynamics of a two-person assurance game are less relevant—we ought to fix our attention on the two-hundred-million-person assurance game. Commerce brings gains from trade and cooperators must be able to rely on the cooperation of their partner. Because exchange is typically impersonal, however, the previously mentioned punishments for cheating are ineffective.

The fundamental problem of exchange is thus precisely the sort of problem that seems likely to create a mismatch between our intuitions and our circumstances. It arises in response to the impersonal and non-iterated interactions that distinguish our modern economic lives. We should therefore expect our atavistic moral intuitions to be poorly suited to manage these conditions. Yet examination reveals our intuitions to have especial utility here.

Sustained social cooperation under the described conditions appears to require a willingness to sacrifice resources, on net, in order to reward fair behavior and punish unfair behavior.[12] To see why, consider a representative instance of the fundamental problem of exchange: the prosaic bilateral, one-shot exchange between a taxi driver and her passenger. Once the drive is complete, the passenger might easily exit without paying unless he expects the driver to punish him (by chasing him, or prosecuting him, and so on). But punishing the free-riding passenger will often come at a net cost.

The most efficient strategy for individual taxi drivers is to eschew punishment themselves and simply rely on the expectation effects established by other drivers' punishments—that is, to free ride on the punishments delivered by others. In brief, punishment presents a public goods problem.[13] Generally speaking, these kinds of problems can be solved with the cooperation of only some threshold number of individuals. We don't need *every* taxi driver to punish in order to produce the general expectation of punishment; enough of them will do. Thus, unless an individual is at the threshold of sufficient cooperation, defection is the dominant strategy—the best strategy regardless of how others act. If enough others refrain from punishing, one might as well refrain oneself given that the deterrent effect will not be produced regardless of what one does. If enough others punish, one still might as well refrain from punishing given that the deterrent effect will be produced regardless of what one does.

It is crucial to emphasize that the individual choice to defect—that is, failing to punish—generally maximizes *social* utility and not simply personal utility. Here's why.

[12] For a discussion of "strong reciprocity" and its role in sustaining cooperation, see Gintis et al. (2006). Greene (2013, 61) offers a discussion of related ideas as well.

[13] For more on how collective action problems pose these sorts of difficulties for utilitarian agents, see Harsanyi (1977).

The public good in question is the deterrent effect produced by the social expectation that free riders will be punished. This good is preserved when an individual driver defects because one driver's failure to punish will not affect the social expectation that free riders will be punished, especially given that each interaction is anonymous and non-iterated, as is characteristic of interaction in large-scale economies.

The upshot of the negligibility of an individual act of punishment is that the expected social utility of contributing to the punishment norm will almost always be lower than that of available alternative acts. Mancur Olson explains why rational altruists will not contribute to public goods:

> Even if the member of a large group were to neglect his own interests entirely, he still would not rationally contribute toward the provision of any collective or public good, since his own contribution would not be perceptible […] He would know that this sacrifice would not bring a noticeable benefit to anyone. Such a rational [actor], however unselfish, would not make such a futile and pointless sacrifice, but he would allocate his philanthropy in order to have a perceptible effect on someone (1971, 64).

Instead of spending time and resources performing a "futile and pointless" act of punishment, a utilitarian-minded driver can do more good by working overtime and donating the extra income to charity, giving a ride to someone in need, and so on. The trouble, however, is that if all drivers think about the punishment problem in these terms, they all will defect and the social benefit will not be provided (i.e., no one will punish, the costs of cheating will plummet, and free riding will increase). The result is that all are worse off. Thus, we have a case of the well-known problem that directly aiming at utility maximization can frustrate the end of utility maximization.

Greene notes that reliance on our commonsense moral intuitions "can stabilize cooperation without a Leviathan" (2013, 55). However, it's worth emphasizing that even if a formal infrastructure is in place to enforce the relevant norms, individuals within the

infrastructure must be disposed to punish when doing so is not cost-effective in order for the infrastructure to function properly. Availing oneself of police or legal services will often cost more than the expected gain of rectifying a loss. Taxi drivers who spend hours at a police station to rectify a lost fare of twenty dollars will probably lose more than twenty dollars in opportunity costs.

So, how might utilitarian-minded people go about achieving cooperation in these sorts of cases? One option is to endorse a form of indirect utilitarianism. Utility will be maximized if agents select actions on the basis of their conformity to certain moral rules (specifically, those rules that maximize utility when followed by most actors[14]) rather than directly on the basis of their expected utility.[15] For example, people who follow a rule mandating contributions to public goods will achieve the benefits due to a general social expectation of punishment. In brief: an act utilitarian criterion of moral rightness justifies a rule utilitarian decision procedure rather than an act utilitarian one.

However, this familiar reply generates a familiar objection, namely that rule utilitarianism is incoherent.[16] Let's consider the problem by way of an example: a taxi driver sees her passenger exit without paying and knows that pursuing punishment will be costly and fruitless. Still, her moral code tells her to follow a rule mandating punishment for the free rider. This result is problematic because the reason for breaking the rule (viz. utility maximization) is the very same reason that recommends the rule in

[14] I say "most" actors rather than "all" because, as Brad Hooker argues, a consequentialist theory ought to provide guidance for dealing with people who lack moral motivation or follow the wrong moral rules. See Hooker (2002, 83).
[15] On this idea see, e.g., Brandt (1979) and Harsanyi (1982).
[16] For discussion of this sort of objection see, e.g., Smart (1956), Lyons (1965).

the first place. It seems irrational or inconsistent to follow a rule when rule following frustrates the very end it is supposed to further.[17]

I see three main strategies for resolving this problem, each of which is either unattractive or seemingly unavailable to Greene. The first is to deny that rule utilitarianism is justified by its tendency to maximize utility but rather by its concordance with our moral intuitions in reflective equilibrium; the second is to assert that it can be rational to follow rules even in cases where rule following frustrates the very end it intends to further; the third is to advise agents to efface utilitarian considerations from their moral decision-making processes. I'll consider each possibility in turn.

One possibility concedes that it is irrational for an agent who follows a rule in the interest of maximizing expected utility to follow that rule when rule-following *fails* to maximize expected utility but denies that rule consequentialism is justified in virtue of maximizing expected utility. Brad Hooker, for example, rejects the idea that rule consequentialism should be grounded in an "overarching commitment to maximize the good" (2002, 101). On this view, there is nothing inconsistent about consequentialism recommending actions that fail to maximize expected utility. The theory is defended by an appeal to its concordance with our considered moral judgments in reflective equilibrium (Hooker 2002, 9-16).

There are two major costs to accepting this reply. First, the proposed revision deprives consequentialism of one of its chief philosophical attractions, what Philippa

[17] The incoherence objection is typically lodged against rule utilitarianism as a criterion of rightness whereas here I am framing it as an objection to the use of rule utilitarianism as a decision procedure intended to satisfy an act utilitarian criterion of rightness. Thanks are due to [omitted] for emphasizing this point. The incoherence objection still has force in this context because it is, at a minimum, a *prima facie* problem if a moral theory only avoids self-defeat in the case that its adherents act irrationally.

Foot calls its "spellbinding force"—namely its intimate connection with the compelling idea that rationality entails preferring more value to less (1985, 198). Those who choose this alternative cannot regard morality as simply the application of a maximizing conception of practical rationality to the problem of distributing social costs and benefits.[18] Indeed, Greene endorses a maximizing conception of consequentialism (2002, 320 n.28; 2013, 152ff).

Second, although it is a virtue of Hooker's rule consequentialism that it can consistently underwrite a stable scheme of social cooperation, we cannot avail ourselves of his conception for present purposes. Conceiving of rule consequentialism as the systematic reconstruction of commonsense morality will not engage the concerns of those who are drawn to consequentialism partly because of their skepticism about commonsense moral intuitions.

As noted, it seems irrational or inconsistent for an agent who follows a given rule in order to maximize expected social utility to follow that rule when doing so *fails* to maximize expected social utility. However, some might reply to this objection by denying that the decision to follow the rule would be irrational. Although a complete treatment of this reply is beyond the scope of this paper, I will briefly explore reasons to reject it.

The worry about the consistency of a commitment to rules is a long-standing issue in both game theory and moral theory. According to classical game theory, players ought to be modular rational, i.e., they ought to maximize expected utility at each particular choice point.[19] Yet we've seen that being modular rational may produce suboptimal results. For

[18] On this idea, see Rawls (1999, 21).
[19] For a more complete discussion of modular rationality, see Skyrms (1998, chapter 2).

example, allowing free-riding makes drivers worse off overall even though allowing free-riding maximizes expected utility at each choice point. Thus, theorists like David Gauthier (1986) and Edward McClennen (1990) suggest building the notion of committed behavior into our conception of rational choice. The idea is that rationality obliges us to adopt the *globally* optimal strategy even if the strategy entails choices that fail to maximize expected utility at particular constituent choice points. By divorcing itself from the theory of modular rationality, we could generate the desired output—that drivers should follow the punishing rule even when doing so is suboptimal. Indeed, if rationality implies the conjunction of the joint choice of an optimal joint strategy and commitment to that optimal joint strategy, rule utilitarianism simply falls out of our revisionary theory of rationality.[20]

The basic objection to this view is that rational plans must be "subgame perfect"—at each node of her decision tree, a rational player makes the choice that maximizes expected utility from that point on. Commitment can be a rational strategy for its expected deterrent effects but it is not irrational to break that commitment if it fails to produce the intended effects. To borrow an example from Brian Skyrms, it could be rational to build a doomsday device that will automatically retaliate to an attack in the expectation that the device will deter such an attack in the first place (1998, 22ff). However, suppose the device malfunctions after the other side launches an attack. It did not retaliate automatically, so you are now confronted with the choice of whether to destroy the other side. Skyrms suggests that there is nothing irrational or inconsistent about choosing *not* to carry on with the retaliation. The reason is because the

[20] See Skyrms (1998, 39) for a discussion of this point.

commitment to retaliation no longer achieves its intended purpose—i.e., deterring an attack on one's own side.

Or consider the taxi driver again. Signaling that one is committed to delivering costly punishment may deter free-riding (e.g., by posting signs in one's taxi stating that passengers who do not pay their fares will be prosecuted). However, there is nothing irrational about a choice to abandon one's commitment if the strategy does not deter—e.g., choosing not to punish once free-riding has already occurred. The strategy is no longer an effective means to one's end.

These sorts of worries underlie my concerns about Greene's approach, despite his apparent sympathy with the notion that moral emotion can function as a commitment device. Greene imagines that Art and Bud face a Prisoner's Dilemma. Art wants to threaten Bud to keep him quiet, but there's a problem. If Bud confesses, they'll both go to prison and it will be pointless for Art to follow up on his threat after they've served their time (Greene 2013, 40). But Greene notes that Art can solve the problem by programming a robot to carry out his threat against Bud—a robot that, critically, cannot be stopped by anyone once it has been programmed. If Bud knows of this robot, he has an incentive not to confess (Greene 2013, 40). Punishing from anger (rather than calculation) can serve a similar function. As Greene puts it, if Bud knows that Art is a vengeful punisher, he won't confess because he'll know that Art will make good on his threats even when doing so is costly and can no longer deter (Greene 2013, 40).

The trouble is that Greene seems to advise that we regard commonsense intuitions—like those regarding punishment norms—as mere (but admittedly weighty) heuristics. Greene does not directly address the issue of self-effacement; however, while he says that

we should be "wary" of explicit utilitarian reasoning in small-scale decision making, he does not go so far as to say we should efface it outright.[21] Thus, he appears to stop short of endorsing full-fledged self-effacement. But I'll argue that this is a move Greene needs to make if our commonsense intuitions are to solve the strategic problems he thinks they can solve.

Let me explain. Viewing the relevant suite of moral emotions as a heuristic is like putting an "emergency override" button on Art's robotic punisher. Suppose Art discovers that Bud has confessed despite his threat—a confession made more likely by Bud's knowledge of the robot's emergency override button. At this point, Art has no reason *not* to initiate an emergency override and spare Bud because there's no way the robot can accomplish what it was designed to accomplish. It cannot deter Bud's confession because Bud has already confessed. (Think back to Skyrms's example: you shouldn't restart the malfunctioning doomsday device once the other side has launched its missiles.) So adding an emergency override to the robot would defeat the very purpose of the robot.

Similarly, suppose a utilitarian-minded taxi driver regards "righteous indignation" as, in Greene's words, a "useful illusion" for ordinary moral decision making.[22] When someone free rides, she experiences the feeling of righteous indignation and a desire to punish—but she also knows that going through with a punishment would do more harm than good. She's in the position of Skyrms's doomsday device operator or Art with his unstoppable robot. Her deterrent device has failed to deter and so it looks irrational for

[21] See Greene (2013, 350).
[22] More specifically, Greene (2013, 274) writes that "our taste for justice is a useful illusion." On the term "righteous indignation," see Greene (2013, 59).

her to go ahead and use it anyway.[23] Now we're back to the original problem: if sufficiently many taxi drivers think this way and decide not to punish, then free riding will increase.

The move toward effacing utilitarian reasoning from the relevant moral decision-making process altogether—which is the final strategy I'll consider—can solve this problem.[24] Some utilitarians suggest that agents ought to internalize, rather than simply comply with, the relevant set of rules.[25] That is, agents should not follow the rules on the basis of an expected utility calculation but rather because they are emotionally committed to performing the actions implied by the rules for their own sake. It is clear how this move solves our problem. The taxi driver would be irrational to punish in order to maximize expected social utility when she expects the punishment to fail to maximize expected social utility. But if she punishes for its own sake, she could not be convicted of irrationality for performing particular acts of punishment that incur a net social cost.

Indeed, some evolutionary game theorists and experimental economists hypothesize that one reason passengers rarely free ride is that many drivers would be sufficiently angry at cheaters to incur the costs of punishing them.[26] The drivers' punishing disposition is beneficial precisely because it is insensitive to costs and benefits. The potential cheater must anticipate that the driver will punish him regardless of the cost or future effects. The "intuitive" punishment rule is effective because individuals elect to

[23] This argument is reminiscent of objections lodged by Bernard Williams (1988) against R.M. Hare's two-level utilitarianism.
[24] On the "self-effacing" question, see e.g., Sidgwick (1981, 489-90), Parfit (1984, chapter 1).
[25] For an argument in this spirit, see Hooker (2002, 76).
[26] See, e.g., (Fehr et al. 2002).

enforce it for its own sake, even at a net cost, and are thus not tempted to defect from the rule.[27]

The effect of apparently irrational punishing behaviors can therefore be to sustain socially beneficial practices.[28] A desire to satisfy retributive emotion—to punish for its own sake—structures individuals' subjective payoffs to punishment in socially useful ways. Our native moral emotions can facilitate cooperation by functioning as de facto commitment devices.[29] They motivate actors to play one-off games *as if* they were iterated. This may not be merely a fortunate coincidence. The evolutionary account of these intuitions explored earlier suggests that they are heuristics adapted to ancestral conditions in which most interactions were iterated; therefore it is not surprising that utilizing these heuristics leads us to treat interactions *as if* they were iterated.[30] It is precisely because interactions tend *not* to be iterated in the modern world that these emotions are so useful.

The example of punishment furnishes one illustration of the strategic difficulties facing utilitarian reasoning, but the difficulties arise for virtually any public goods scenario. The trolley problem provides another example. As Greene notes, utilitarians have good reason to dissuade people from pushing others in front of trolleys in real-world conditions (2013, 211). Following the deontological norm in the trolley problem scenario,

---

[27] Greene (2013, 15) writes that our moral emotions are "gut-level instincts that enable cooperation within personal relationships and small groups." But this statement omits mention of their ability to enable cooperation within impersonal relationships and large groups. After all, retributive punishment (or the threat thereof) appears to foster cooperation in the taxi case, where the actors involved are strangers and the interaction takes place within an enormous community. Indeed, the actors can cooperate in this case while also endorsing different answers to the controversial large-scale moral questions that underlie the "Tragedy of Commonsense Morality."

[28] For further discussion of the prudence of "vengeance-seeking," see Frank (1988, 83), Joyce (2006, 119).

[29] For a different perspective on this point, see Frank (2007).

[30] Greene (2013, 58) suggests something like this possibility.

like following the deontological norm in the punishment scenario, produces the public good of beneficial social expectations. It helps us avoid creating an atmosphere of distrust, fear, hostility, and uncertainty.[31]

The problems involved in providing this public good parallel the problems involved in the punishment case. One person's contribution to the social expectation that we can use trolleys and so on in peace—that is, her adherence to the deontological norm—generally does not affect whether this public good is provided, especially given that most modern social interaction is anonymous and non-iterated. Thus, each person has reason to defect in particular cases when defection maximizes net utility. However, if everyone thinks in these terms, the public good will not be provided because the social expectations will unravel.

Here, as in the punishment case, if people are to commit to the relevant rule without a breach of modular rationality, they need only follow their deontological intuitions. Someone could be judged irrational for abstaining from pushing a bystander in front of the trolley in the interest of maximizing social benefits when her abstention incurs a net social *cost.* Thankfully, however, the model of moral judgment under consideration indicates that people abstain from pushing others into harm's way because they experience an emotional aversion to the prospect of doing so, not because of a cost-benefit calculation. Here again, our deontological preferences are effective mechanisms for the enforcement of social rules like collective action norms.

§3

[31] On the importance of expectation effects for utilitarianism, see, e.g., Harsanyi (1977).

Greene devotes considerable attention to the ways in which our moral intuitions are useful for solving public goods problems. In this section, I'll argue that they are similarly useful for facilitating coordination.

The strategic problems posed by impersonal, non-iterated interactions also suggest the importance of a *public* moral decision procedure: for cooperation to flourish, individuals must accept the rules governing social practices and know that others accept those same rules.[32] The previous section explains why deontological intuitions are critical to securing the first goal—that is, ensuring individuals' commitment to the rules. This section will explain why deontological intuitions are also critical to securing the second goal—that is, ensuring that individuals arrive at a set of rules that is commonly accepted and publicly known.

Utilitarians, including Greene, generally agree that a moral decision procedure must be public, as public rules enable individuals to coordinate their plans and expectations with one another.[33] An effective social morality will supply what John Rawls calls "a common basis for determining mutual expectations" (1999, 49).

Consider again the fundamental problem of exchange. In order to effectively deter potential cheaters and thus avoid the costs of mutually disadvantageous conflict, it is not enough for taxi drivers to reliably punish cheaters. Potential cheaters must *anticipate* that they will be punished. Indeed, potential free riders must expect that most drivers will react to being cheated by punishing even at a high cost. But establishing reliable expectations is especially difficult in this situation, as agents are not playing an iterated

[32] For a related discussion of publicity, see Rawls (2005, 66-71).
[33] On Greene's endorsement of Rawls's publicity condition, see Greene (2002, 326 n.37). See also Rawls (1999, 115). On the publicity condition generally, see Sidgwick (1981, 489-90), Parfit (1984, 40-3), Railton (1984).

game in which reputation effects have a chance to take root. This is precisely the sort of interaction that characterizes most of our everyday economic lives. So how can cheaters know to expect punishment in the absence of specific knowledge about each person they encounter?

Here is one possibility: they know how *they* would feel were they in the driver's situation and thus predict what the driver's mental state will be. James Woodward and John Allman note that social emotions and moral intuition play a central role in simulating others' mental states:

> A large body of evidence suggests that we often detect and represent the mental states of others (including their beliefs, preferences, intentions, and emotions) by simulating these via our own emotional processing—that is, in representing the mental states of others, we activate the emotional areas and processing in ourselves that are involved in the those mental states when experienced by others. By further simulating how we would behave in the presence of this mental state in ourselves, we may also be able to predict successfully how others will behave, given that they have this mental state (2007, 194).

This mechanism of simulation and prediction is surpassingly efficient: it enables agents to avoid mutually disadvantageous conflict without requiring costly cooperative infrastructure like third party arbitration and enforcement.

For this mechanism to function, however, there must be a homogeneous affective response across cooperators (e.g., across passengers and drivers). And this is precisely what our evolved moral psychology has given us. Evidence suggests that (almost) everyone experiences the relevant moral emotions, even those who ultimately resist them. Nature facilitates consensus on these norms *gratis*. Thus, our deontological intuitions come "ready-made" with an efficient solution to the problem of coordinating our plans and expectations with one another.

Settling on a new coordination point would be costly, maybe prohibitively so. Even if a collective conversion to the use of utilitarian reasoning as a decision procedure would bring about a gain in utility relative to our current coordination point (although my arguments in the previous section provide reason to doubt this claim), we still need an account of how to get there from here. The need for a coordinating procedure remains even for a view that eschews cost-benefit reasoning in favor of internalizing an emotional commitment to moral rules. If the view recommends that people internalize a *new* set of rules rather than the one provided readymade by evolution, it must ensure that people internalize the same rules. How, then, do people converge on new norms in the absence of centralized direction? Ironically, the answer is often evolution of some sort.

To illustrate, let's set aside questions of adjudication between deontological intuitions and utilitarian reasoning to consider an analogous rivalry between evolved norms and their seemingly superior counterparts: English versus Esperanto. Proponents of Esperanto argue that the constructed language enjoys decisive advantages over natural languages like English, whose arbitrary and suboptimal features are products of its accidental evolutionary history. In particular, Esperanto is culturally neutral and easier to learn. Supplanting English with Esperanto seems like a net gain. Modeling language as a simple pure coordination game, Esperantists might tout the following payoff matrix (assume cardinal utilities where 10 = best and 0 = no coordination):

	English	Esperanto
English	8,8	0,0
Esperanto	0,0	10,10

Figure 1: Language as a Pure Coordination Game

{English, English} and {Esperanto, Esperanto} are both Nash equilibria, but {Esperanto, Esperanto} Pareto dominates {English, English}. So why isn't this paper written in Esperanto?

The reason is that English was already "locked in" when Esperanto arrived. Speakers find themselves in the English equilibrium and require a public conversion to Esperanto to make their own conversion worthwhile. As the number of players in this coordination game increases, not only does coordinating a public conversion become less feasible, but also the payoffs to conformity to the established norm increase. New speakers (or the parents of new speakers) gain more by choosing English over Esperanto. They can coordinate their speech with more speakers by choosing English. If one's interest is communication, it is not rational to speak Esperanto in a community of English speakers. This is true regardless of whether the global adoption of Esperanto would be more efficient than the global adoption of English. To reap the benefits of speaking a language—to communicate—speakers must converse in the same language. The norms of English may be arbitrary to some degree, but adherence to the norms of English is not. The norms' arbitrariness does not prevent them from facilitating mutual advantageous cooperation. Ensuring that agents coordinate on a preferred point can be less important than simply ensuring that agents coordinate.

As the preceding indicates, English's dominance is the product of increasing returns, not calculated design. The payoffs associated with conformity to its norms increase as the number of people conforming to its norms increases. This is why, in Hume's terms, languages are "gradually establish'd by human conventions without any promise" (2003, Bk. III, Pt. II, Sec. II). We do not contract to speak a language, nor must we submit to

coercive institutions that enforce linguistic unanimity. Linguistic convention "arises gradually, and acquires force by a slow progression, and by our repeated experience of the inconveniences of transgressing it" (2003, Bk. III, Pt. II, Sec. II). On Hume's model, social norms are established through uncoordinated feedback mechanisms rather than explicit agreement or centralized direction: conforming to linguistic convention is convenient; transgressing linguistic convention is inconvenient. People converge over time because of increasing returns dynamics—adopting the language of one's community pays.

Similarly, increasing returns dynamics favor our allegedly native deontological intuitions. To reap the benefits of abiding by a social morality—to secure reliable and publicly known conditions of cooperation—agents need to use the same basic rules. (As noted earlier, "righteous indignation" serves as an effective deterrent in the taxi case partly because the participants know what to expect from each other.) As our moral psychology stands, we are in Nash equilibrium: no gain comes about from the unilateral adoption of new moral rules. Thus, *why* deontological intuitions are locked in may be a reflection of "the influence of morally irrelevant factors"; *that* deontological intuitions are locked in is not a morally irrelevant factor if individuals aspire to field the mutually best responses and achieve social coordination. Path-dependence may therefore be an ineliminable feature of real-life social norms.

Certainly individuals can and do find themselves locked into suboptimal equilibria.[34] And our native deontological intuitions are admittedly imperfect. Thus, I do not want to

[34] See Bicchieri (2006, chapter 5). Our deontological intuitions can bring us to a Nash equilibrium even if utilitarian reasoning would be a better coordination point (although, again, my arguments in the previous section cast doubt on this claim). Consider the case of language again: even if we'd all be better off speaking Esperanto, I don't have a reason to switch unless sufficiently many

deny that there are costs to relying on them. Yet "real-life consequentialism," as Greene says, "aims to take nearly everything into account" (2008, 64). So any real-life consequentialist theory must take coordination problems into account. The relevant question, then, is not: counterfactually, in a world in which savvy utility calculators design moral norms from scratch, would there be reason to adopt these newly engineered norms?, but rather: here and now, is there reason to *convert*?[35]

Let's pause for an objection.[36] If we are "locked in" to at least some of our imperfect-but-functional intuitions, are we doomed to what Greene calls the "Tragedy of Commonsense Morality"? Even if people generally share a taste for (e.g.) retributive punishment that motivates them to incur the (net) costs of calling the police to catch free riders, what about moral issues like abortion or euthanasia? Here our intuitions tear us apart rather than bind us together.

Greene's solution is to use the "common currency" of utilitarian reasoning to resolve controversial moral issues, particularly those that concern public policy (2013, 190ff). We generally share the capacity for utilitarian reasoning and can use it to assess those macro-level moral problems that our commonsense intuitions are ill-equipped to handle. These are the kinds of problems that people don't encounter on a day-to-day basis and for which their intuitions favor conflicting solutions.

---

others switch as well. Similarly, if a useful moral code is one that will help us coordinate our plans and expectations—that is, it will function as a *shared* decision procedure—then individual moral agents should switch to a new code only if others switch too. Thanks are due to an anonymous referee for helpful questions on this point.

[35] An anonymous referee notes that utilitarians are able to coordinate with non-utilitarians in the real world. However, I would predict that this coordination is due to utilitarians' reliance on intuitive moral judgment for micro-level decisions.

[36] Thanks are due to an anonymous referee for raising this objection.

I think that Greene can preserve this view of macro-level moral decision making even if he revises his views about micro-level moral decision making to account for self-effacement. The idea here is that the best outcome would result from acceptance of the correctness of (something like) deontology at the level of personal conduct and the correctness of utilitarianism at the level of institutional design. The difference between this proposal and Greene's is that it would make a clean break between micro-level and macro-level principles instead of allowing utilitarian reasoning to lurk in the background of all moral decision making.

Is this kind of moral split realistic? I see at least two reasons for optimism. For one, the dual process model seems to fit nicely with such a division, at least at first blush. That is, we're naturally "of two minds" when it comes to morality—we can appreciate the attractions of the two different moral standards. Moreover, people are disposed to view violations of the intuitive deontological standards as *wrongs* rather than mere departures from useful rules of thumb.[37]

There is also philosophical precedent for splitting apart micro-level and macro-level standards. For instance, Rawlsian public reason liberalism rests partly on the claim that, in Norman Daniels's words, people "wear two morally distinct hats" (1996, 152). As Rawls puts the point, "The principles of justice for institutions must not be confused with the principles which apply to individuals and their actions in particular circumstances. These two kinds of principles apply to different subjects and must be discussed separately" (1999, 47). One motivation for this divide is to enable citizens who subscribe to different micro-level moral standards to find common ground that can help them

[37] For instance, Greene (2013, 114) reports that subjects say that it is wrong to push someone off of a footbridge in the trolley problem.

resolve the sorts of macro-level conflicts that concern Greene. Although I lack the space to examine the arguments for this view, it's worth noting that many moral and political philosophers find it plausible.

In closing, I'd like to reiterate that my aim has not been to defend any particular view of moral psychology, utilitarianism, or self-effacement. It could be the case that the dual process model is incorrect, or that it fails to provide any vindication for utilitarianism as a moral standard, or that self-effacement renders utilitarianism unacceptable. I take no stand on these issues. Rather, I made an argument about the form our moral decision making should take *if* we accept Greene's account of moral psychology and moral philosophy.

References

Baron J, Ritov I (1993) Intuitions about penalties and compensation in the context of tort law. Journal of Risk and Uncertainty 7(1): 17-33

Berker S (2009) The normative insignificance of neuroscience. Philosophy and Public Affairs 37(4): 293-329

Bicchieri C (2006) The grammar of society: the nature and dynamics of social norms. Cambridge University Press, New York

Brandt R (1979) A theory of the good and the right. Oxford University Press, Oxford

Daniels N (1996) Reflective equilibrium and justice as political. In: Justice and justification. Cambridge University Press, Cambridge, pp. 144-178

Fehr E, Fischbacher U, Gächter S (2002) Strong reciprocity, human cooperation and the enforcement of social norms. Human Nature 13(1): 1-25

Foot P (1967) Abortion and the doctrine of double effect. Oxford Review 5: 28-41

Foot P (1985) Utilitarianism and the virtues. Mind 94(374): 196-209

Frank R (1988) Passions within reason: the strategic role of the emotions. W.W. Norton and Company, New York

Frank R (2007) The status of moral emotions in consequentialist moral reasoning. In: Zak P (ed) Moral markets: the critical role of values in the economy. Princeton University Press, Princeton, pp 42-62

Gauthier D (1986) Morals by agreement. Oxford University Press, Oxford

Gintis H, Bowles S, Boyd R, and Fehr E (eds) (2006) Moral sentiments and material interests: the foundations of cooperation in economic life. The MIT Press, Cambridge, MA

Greene J (2013) Moral tribes. Penguin, New York

Greene J (2008) The secret joke of Kant's soul. In: Sinnott-Armstrong W (ed) Moral psychology, volume 3: The neuroscience of morality: emotion, brain disorders, and development. MIT Press, Cambridge, MA, pp 35-80

Greene J (2002) The terrible, horrible, no good, very bad truth about morality, and what to do about it. Ph.D. dissertation, Department of Philosophy, Princeton University

Greif A (2000) The fundamental problem of exchange: a research agenda in historical institutional analysis. European Review of Economic History 4(3): 251–284

Harsanyi J (1977) Rule utilitarianism and decision theory. Erkenntnis 11(1): 25-53

Harsanyi J (1982) Morality and the theory of rational behavior. In: Sen A and Williams B (eds) Utilitarianism and beyond. Cambridge University Press, Cambridge, pp 39–62

Hooker B (2002) Ideal code, real world. Oxford University Press, Oxford

Hooker B (1996) Ross-style pluralism versus rule-consequentialism. Mind 105(420): 531-52

Hume D (2003) A treatise of human nature. Oxford University Press, Oxford

Joyce R (2006) The evolution of morality. The MIT Press, Cambridge, MA

Kamm FM (2009) Neuroscience and moral reasoning: a note on recent research. Philosophy and Public Affairs 37(4): 330-345

Lyons D (1965) Forms and limits of utilitarianism. Clarendon Press, Oxford

Mackie JL (1985) Morality and the retributive emotions. In: Mackie J and Mackie P (eds) Persons and values, volume 2. Oxford University Press, Oxford, pp 206-219

McClennen E (1990) Rationality and dynamic choice: foundational explorations. Cambridge University Press, New York

Olson M (1971) The logic of collective action. Harvard University Press, Cambridge MA

Parfit D (1984) Reasons and persons. Oxford University Press, Oxford

Railton P (1984) Alienation, consequentialism, and the demands of morality. Philosophy and Public Affairs 13(2): 134-71

Rawls J (2005) Political liberalism. Columbia University Press, New York

Rawls J (1999) A theory of justice. The Belknap Press, Cambridge

Sanfey A, Rilling J, Aronson J, Nystrom L, and Cohen J (2003) The neural basis of economic decision-making in the ultimatum game. Science 300(5626): 1755-1758

Sidgwick H (1981) The methods of ethics. Hackett, Indianapolis

Singer P (2005) Ethics and intuitions. The Journal of Ethics 9(3-4): 331-52

Skyrms B (1998) Evolution of the social contract. Cambridge University Press, Cambridge

Smart JJC (1956) Extreme and restricted utilitarianism. The Philosophical Quarterly 6(25): 344-354

Street S (2006) A Darwinian dilemma for realist theories of value. Philosophical Studies 127(1): 109-166

Sinnott-Armstrong W (2006) Moral intuitionism meets empirical psychology. In: Horgan T and Timmons M (eds) Metaethics after Moore. Oxford University Press, Oxford, pp 339-366

Thomson JJ (1976) Killing, letting die, and the trolley problem. The Monist 59(2): 204-17

Thomson JJ (1985) The trolley problem. The Yale Law Journal 94(6) 1395-1415

Williams B (1988) The structure of Hare's theory. In: Seanor D and Foton N (eds) Hare and critics. Clarendon Press, Oxford, pp 185-196

Woodward J and Allman J (2007) Moral intuition: its neural substrates and normative significance. Journal of Physiology - Paris 101(4-6): 179-202