Cognition xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/COGNIT

Original Articles

The rat-a-gorical imperative: Moral intuition and the limits of affective learning

Joshua D. Greene [*]

Department of Psychology, Center for Brain Science, Harvard University, United States

ARTICLE INFO

Article history:
Received 10 April 2016
Revised 6 December 2016
Accepted 6 March 2017
Available online xxxx

Keywords:
Deontology
Utilitarianism
Consequentialism
Reinforcement learning
Model-free learning
Machine learning
Ethics
Normative ethics
Moral judgment

ABSTRACT

Decades of psychological research have demonstrated that intuitive judgments are often unreliable, thanks to their inflexible reliance on limited information (Kahneman, 2003, 2011). Research on the computational underpinnings of learning, however, indicates that intuitions may be acquired by sophisticated learning mechanisms that are highly sensitive and integrative. With this in mind, Railton (2014) urges a more optimistic view of moral intuition. Is such optimism warranted? Elsewhere (Greene, 2013) I've argued that moral intuitions offer reasonably good advice concerning the give-and-take of everyday social life, addressing the basic problem of cooperation within a "tribe" ("Me vs. Us"), but that moral intuitions offer unreliable advice concerning disagreements between tribes with competing interests and values ("Us vs. Them"). Here I argue that a computational perspective on moral learning underscores these conclusions. The acquisition of good moral intuitions requires both good (representative) data and good (value-aligned) training. In the case of inter-tribal disagreement (public moral controversy), the problem of bad training looms large, as training processes may simply reinforce tribal differences. With respect to moral philosophy and the paradoxical problems it addresses, the problem of bad data looms large, as theorists seek principles that minimize counter-intuitive implications, not only in typical real-world cases, but in unusual, often hypothetical, cases such as some trolley dilemmas. In such cases the prevailing real-world relationships between actions and consequences are severed or reversed, yielding intuitions that give the right answers to the wrong questions. Such intuitions—which we may experience as the voice of duty or virtue—may simply reflect the computational limitations inherent in affective learning. I conclude, in optimistic agreement with Railton, that progress in moral philosophy depends on our having a better understanding of the mechanisms behind our moral intuitions.

1. Introduction

How reliable are our moral intuitions? Under what circumstances should we accept or reject their advice? And what, exactly, is the alternative to intuitive moral judgment? Are not all judgments ultimately grounded in intuition? These questions are central to scientifically informed discussions of normative ethics. In an insightful and illuminating recent paper, Peter Railton (2014) argues that some researchers, myself among them, have painted a philosophical portrait of moral intuition that is too unflattering. Railton argues that moral intuition need not be "fast" and "automatic", and therefore need not be correspondingly myopic or biased. He draws on psychological and neuroscientific research showing that affective intuitions are the products of sophisticated learning systems that are both flexible and integrative (Behrens, Woolrich, Walton, & Rushworth, 2007; Grabenhorst & Rolls, 2011; Quartz, 2009; Schultz, Dayan, & Montague, 1997; Singer, Critchley, & Preuschoff, 2009; Tobler, O'Doherty, Dolan, & Schultz, 2007). These learning systems, he argues, attune us to the subtle contours of the decision landscape, and the intuitions generated by these systems embody their hard-won wisdom.

Here I offer a friendly counterpoint to Railton's optimistic assessment of moral intuition. He and I have, I think, no fundamental disagreement concerning the strengths and limitations of affective learning and the intuitive judgments that such learning supports. Instead, our disagreement is one of emphasis, but nonetheless significant for that. In what follows I briefly review

[*] Address: Department of Psychology, 33 Kirkland St., Cambridge, MA 02138, United States.
E-mail address: jgreene@wjh.harvard.edu

http://dx.doi.org/10.1016/j.cognition.2017.03.004
0010-0277/© 2017 Elsevier B.V. All rights reserved.

Railton's case for optimism. I then present a framework for assessing the general strengths and weaknesses of intuitive judgment, focusing on the distinction between model-based and model-free strategies for learning and deciding (Crockett, 2013; Cushman, 2013; Sutton & Barto, 1998). Drawing on this framework, I explain why even very sophisticated learning processes can produce intuitive judgments that are systematically misguided. I then return to the key normative question and argue that one's assessment of moral intuition will depend on one's goal as a moral thinker: Is the goal to organize and justify our most central and widely shared moral practices? Or is it to help us solve moral problems, to answer the moral questions that divide us?

If one believes, as I do, that the primary aim of moral philosophy should be to solve moral problems, then it makes sense to emphasize the limitations of our moral intuitions, including intuitions produced by sophisticated learning processes. This is because moral philosophy, so conceived, must focus on cases of moral disagreement, both across people (moral controversies) and within people (moral paradoxes). In such cases, we should expect our moral intuitions—including intuitions generated by sophisticated learning processes—to fail us often. Finally, I close with some optimistic remarks concerning a conclusion on which Railton and I agree: Understanding the mechanics of moral intuition is not only a worthy scientific endeavor, but also essential for progress in moral philosophy.

2. Attunement and the optimistic view of moral intuition

In keeping with a long philosophical tradition (Aristotle, 1941), Railton argues that intuition can be sophisticated, flexible, and generally smart, reflecting a lifetime of hard-won experience. (See also Haidt, 2003; Pizarro & Bloom, 2003). This view is presented in contrast to a seemingly more pessimistic view of moral intuition (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene, 2013; Haidt, 2001, 2012; Singer, 2005), and intuition more generally (Kahneman, 2003, 2011), according to which "fast", "reflexive", "point-and-shoot" intuitions often bias our judgments.

To illustrate his argument for optimism, Railton describes the case of a defense attorney, in the midst of a murder trial, whose highly attuned intuitions enable her to win an important legal and moral victory. Despite the overwhelming strength of the evidence she has set before the jury, she senses that she is failing to reach them. An inner voice, which grows increasingly persistent, tells her that she must cast aside her trademark detached, meticulous style and instead speak from the heart. And so she does, drawing up powerful words from a previously untapped reservoir of conviction. She meets each juror's eyes and one by one conveys to them the simple truth she feels in heart. And thus she wins the case.

A key feature of this example is that the protagonist, while relying heavily on her burgeoning jurist's intuitions, was not merely acting in a "fast", "automatic", "point-and-shoot" way. Indeed, she cast aside her habitual detached style, which the jury perceived as cold and condescending. Nor did she arrive at her winning strategy simply by reasoning from the observable facts. Instead, her winning performance was the product of an extended dialogue between her conscious reasoning and her, at times inexplicable, gut feelings about how (not) to win the case. Critically, these feelings were not generic reflexes and certainly not innate responses. Instead, these feelings reflected the lessons of a broad range of experiences, the significance of which she could only dimly appreciate at the outset. In short, she succeeded by relying, in a thoughtful way, on her sophisticated, well-attuned intuitions.

This example is fictional, but Railton also provides ample empirical support for the psychological lessons he draws from this case. A great deal of evidence indicates that humans, like other mammals, have a core set of systems for affective learning that are flexible, highly attuned to the available evidence, and therefore likely to produce behavior that we would naturally regard as rational (Behrens et al., 2007; Grabenhorst & Rolls, 2011; Quartz, 2009; Schultz et al., 1997; Singer et al., 2009; Tobler, O'Doherty, Dolan, & Schultz, 2007). Railton focuses on recent advances in cognitive and computational neuroscience, but classic studies of expert judgment (Chase & Simon, 1973; deGroot, 1946/1978) make the same point: After years of learning from experience, chess experts, for example, can intuitively "see" certain moves as good and fail to even consider the bad moves favored by lesser players.

With this view of intuitive judgment in the background, Railton reviews some classic hypothetical scenarios from the moral psychology and philosophy literatures. He considers Haidt's case of Mark and Julie, the adult brother and sister who decide to have sex, just once, using multiple forms of birth control, in hopes that they will enjoy it and become closer (Haidt, 2001; Haidt, Bjorklund, & Murphy, 2000). People typically respond to this case with disgust and vigorously condemn Mark and Julie's behavior. What's more, people typically stand by their condemnation, even as they struggle to articulate a coherent justification for it—a phenomenon that Haidt calls "moral dumbfounding". From this, one might conclude that people's stubborn adherence to their affective intuitions is "dumb", but Railton disagrees. In Haidt's telling, things work out well for these siblings, but as Railton observes, their behavior was nonetheless reckless and foolish. They were, as he puts it, playing Russian roulette with their relationship. People's insistent condemnation of this behavior may not be dumb at all, even for people who struggle to articulate the reasons behind it.

Railton's more general conclusion after considering the available scientific research, some classic cases form the ethics literature, and his own extended example is that our moral intuitions are smarter than many have thought, implicitly reflecting the hard-won benefits of experience.

3. Intuitions as learned, flexible, and integrative: some clarifications

Before moving on to a more detailed consideration of the strengths and limitations of learned intuitions, I'd like to make three clarifications concerning my previously stated views, which Railton contrasts with his own. The first clarification concerns the respective roles of domain-general processes for learning and deciding versus domain-specific decision processes that are highly genetically constrained. The second and third clarifications concern the ways in which intuitive judgments, in general, are and are not flexible and integrative.

While I have at times emphasized the likely role of genetic influences on intuitive moral judgment (Greene, 2003, 2013; Greene & Haidt, 2002, chap. 1–2), I've long maintained that moral intuitions depend critically on learning (Greene, 2002, 2013, chap. 3). With respect to this question of "nature vs. nurture", trolley dilemmas (Foot, 1967; Greene et al., 2001; Thomson, 1985) in particular present an interesting case. This is because they elicit responses that are, in some respects, surprisingly consistent across cultures (Hauser, Cushman, Young, Kang-Xing Jin, & Mikhail, 2007). More specifically, people from a wide range of cultures typically judge that it's worse to save five lives by pushing the man off the footbridge than by hitting a switch that turns the trolley onto one person. What's most interesting is that this consistency appears in the apparent absence of explicit teaching or accessible knowledge of the principles that govern such patterns (Cushman, Young, & Hauser, 2006; Hauser et al., 2007). Consistent with this, very few people can provide a coherent justification for distinguishing between these two cases (for example, by appealing to the distinction between means and side-effect). The apparent universality of the footbridge-switch effect, combined with people's widespread inability to explain why these two cases ought to be judged differently, has prompted some to posit the existence of a dedicated and innate "universal moral grammar" (Hauser et al., 2007; Mikhail, 2000, 2011).

For over a decade, the empirical evidence has consistently shown that there is no single cognitive system, let alone an innately configured single system, responsible for producing this widely observed pattern of judgment. Instead the evidence, which I will not review here, favors a dual-process view (Cushman, 2013; Greene, 2013, 2014a, 2014b; Greene et al., 2001; Koenigs et al., 2007). While I have from the outset rejected single-system theories, I have at times (Greene, 2007; Greene & Haidt, 2002) suggested that the negative reaction to causing "personal" harm, as in the footbridge case (Greene et al., 2001, 2009), reflects a domain-specific, innately supported affective response. My view on this question has since shifted. Following Cushman (2013) and Crockett (2013), I'm increasingly convinced that learning[1] plays a dominant role in generating these patterns of judgment and that whatever genetic influences are at work—which may be very important—are likely operating on domain-general cognitive systems (Greene, 2014b; Shenhav & Greene, 2010), rather than on a domain-specific "module" related to morality, or to causing personal harm more specifically.[2] However, as I will explain below, if it turns out that these patterns of judgment depend more on learning and less on genes, or depend entirely on domain-general processes rather than one or more domain-specific ones, that may not do much to vindicate them.

Next, what about the characterization of intuition as "fast" and inflexible, rather than "slow" and integrative? Here, despite appearances, I have no substantive disagreement with Railton. We both agree that "fast", automatic responses serve as *inputs* into judgment, and neither of us believes that such responses by themselves *determine* our judgments. For example, in responding to the footbridge case, the judgment is not automatically determined by whether the judge has an automatic negative response to pushing the man to save the five. Instead, the judgment will depend on whether and to what extent one *relies* on that response as it competes with the utilitarian thought that it's better to save more lives. (See also Kahneman (2003), who distinguishes the "System 1" responses that inform and often dominate judgment from the explicit judgments themselves, which always involve "System 2".) The distinction between having intuitions and relying on intuitions is illustrated by studies employing the Cognitive Reflection Test (Frederick, 2005) to measure and manipulate (Pinillos, Smith, Nair, Marchetto, & Mun, 2011) people's style of moral judgment (more intuitive vs. reflective) (Paxton, Bruni, & Greene, 2014; Paxton, Ungar, & Greene, 2012). These results indicate that people's moral judgments are influenced by whether they are in general, or at the moment of decision, disposed to rely on their intuitive responses. Consistent with this, a recent fMRI study (Shenhav & Greene, 2014) dissociates a more reflexive emotional response that depends on the amygdala ("How bad do I feel about pushing?") from a more integrative and reflective (but still affective) weighing process that depends on the ventromedial prefrontal cortex ("Does saving four extra lives justify doing this terrible thing?"). And, of course, as decisions get more complicated (without the simplifying assumptions of hypothetical dilemmas) there will be more intuitive inputs, more complex assessments of likely consequences, and more affective integration over the above.

Thus, it seems that difficult decisions, moral and otherwise, typically involve a deliberative dialogue between "fast" automatic processes and "slow" controlled processes. (See also Cushman, 2013.) Such deliberative processes are flexible and integrative in the sense that they involve a conscious and controlled consideration of multiple inputs. However—and this is the point I have long emphasized (Greene, 2003, 2007, 2013, 2014a)—some of these inputs are "fast", inflexible, and often decisive. As one deliberates about trolley dilemmas, for example, a typical person has no choice but to find the action in the footbridge case more emotionally disturbing than the action in the switch case. And while it is possible to ignore or override that emotional signal, most people do not. More generally, and despite our capacity to do otherwise, our inflexible "fast" responses often dominate our moral judgments (Haidt, 2001, 2012).

Finally, I would like to distinguish between flexibility in the *formation* of an intuition and the flexibility of an intuition at the time of its *deployment* in judgment. As Railton observes, the learning mechanisms that produce intuitive responses may be highly flexible in that they integrate over a broad and temporally extended range of experiences *during the learning process*. But in the moment of decision, the intuitive response itself may be "fast" and correspondingly inflexible. Once again, you can override your horror at the thought of pushing the man off the footbridge, but you can't make the horror disappear. It is this inflexibility, present at the time of decision, that matters most when we are considering the limitations of intuitive judgment.

4. Learning and the limits of affective intuition

We'll agree, then, that intuitive judgments can be smart, reflecting a rationally defensible integration of knowledge gained from past experience. Here one might be tempted to reserve this praise for intuitions acquired through one's own learning, but that would be a mistake. Individual learning, social/cultural influence, and genetic influence all reflect trial-and-error learning. It is only the time scales and transmission mechanisms that differ. For example, an animal with an innate fear of its natural predators has benefitted from the trial-and-error experience of its ancestors (or would-be ancestors), and the lessons embodied in its instincts may be superior to whatever lessons it might draw from its own limited experiences. In the sense that's relevant here, all intuitions are learned, and all intuitions can be smart thanks to the learning they embody.

If our learned intuitive judgments are so smart, why, then, do we ever need anything else? Because there are some things that intuitive judgment simply can't do. To acquire a good intuitive response through learning requires (at least) two things. First, it

[1] Here by "learning" I mean learning in the ordinary sense of learning over an individual lifespan. However, even innately specified stimulus-response mappings ("Pavlovian" responses, which may or may not involve Pavlovian conditioning) involve learning on an evolutionary time scale.

[2] In Greene (2013, pp. 224–240) I refer to this hypothesized module as the "myopic module". Consistent with the evidence provided previously, I continue to believe that the mechanism responsible for attaching negative affective values to prototypically violent actions (such as pushing people off of footbridges) is myopic, with its blindness to side effects, etc. However, I now think it's a bit misleading to call this mechanism a "module", as the "module" seems to be our domain-general system for habitual control of behavior (Crockett, 2013; Cushman, 2013). With that said, this system is modular in certain key respects, and what matters for normative purposes is the myopia, not the domain-specificity of the mechanism. If all of this is correct, it implies that the blindness to side effects observed in the switch/footbridge contrast is in fact a general feature of the model-free learning system.

Please cite this article in press as: Greene, J. D. The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.004

requires *good data*, by which I mean a sufficiently representative set of cases from which to learn. Second, it requires a *good trainer*, by which I mean a mechanism (which need not be a human teacher) that provides evaluative feedback that is aligned with the values that we—the ultimate evaluators, whoever we happen to be—hold. If a learning process lacks one of these features, the resulting intuitions will be unreliable.

In considering what counts as "good data" and "good training", two points deserve special attention. First, whether a set of training cases is "sufficiently representative" depends not only on the training cases themselves, but also on the kinds of problems one hopes to solve after training. Learning to fish in the Caribbean may provide adequate training for fishing in the Mediterranean, but may not provide adequate training for ice fishing in Manitoba. Second, whether the feedback delivered during training constitutes good training is relative to one's ultimate goals. To learn how to fish, it's not enough to encounter a broad range of fish, similar to those available in the target environment. The fishing instructor must know which fish are worth catching and inform her trainees accordingly. And what counts as worth catching may vary from one fishing operation to the next, depending, for example, on whether the goal is to catch fish that are tasty, marketable, sustainable, or challenging to reel in.

Critically, a decision-maker who lacks intuitions that are informed by adequate learning is not necessarily out of luck. It may be possible to make one's choice, not based on one's feelings about the available actions, but based on an explicit goal and an explicit understanding of which actions are most likely to achieve that goal. For example, a fishing enthusiast who has never before gone sport fishing might deploy methods that are very different from any she has previously used. This first-time sport fisher might rely on some old fishing habits while suppressing others, but the overall strategy may be based on an explicit goal and an explicit plan for achieving it, not a general reliance on intuition.

These lessons are illustrated and made more precise by the distinction, originally drawn by computer scientists, between model-based and model-free reinforcement learning (Sutton, 1988; Sutton & Barto, 1998). In what I regard as a very important recent development, two researchers (Crockett, 2013; Cushman, 2013) have independently proposed that dual-process moral psychology, and dual-process psychology more generally, are best explained in terms of the more basic computational distinction between model-based and model-free algorithms for leaning and deciding. I will now briefly summarize the key ideas behind this distinction before returning to moral psychology.

Model-based learning involves accumulating information about the decision environment and using that information to build a causal model of that environment. For example, a rat in a maze might learn to obtain a reward by exploring the maze and building an internal map of the maze, which includes the location of the reward. Critically, a map is an integrated *causal model*, encoding information concerning the expected effects of moving in various directions from various starting points. When a rat with an internal map revisits the maze, it can use its map to guide it toward rewards and away from hazards. Spatial maps are causal models, but a causal model need not be a spatial map. For example, one can learn to operate a machine by constructing an explicit understanding of what its various levers and buttons do and which actions are necessary to achieve a given result. Model-based learning and decision-making corresponds to what we would naturally identify as reasoning and planning: using an understanding of how the world works to identify a sequence of actions that will get one to one's goal.

Model-free learning and decision-making work in a fundamentally different way. Instead of building an explicit model of the world, model-free learners attach positive or negative values directly to actions (or action-context pairs) based on whether and to what extent those actions have been rewarded in the past. For example, if a rat stumbles upon the rewarding cheese after making a right turn out of a red room, the next time it finds itself in the red room (or a similar room) it will feel an urge to turn right. It will feel this way, not because it knows (i.e. explicitly represents) that the next room contains a reward, but simply because the delivery of the previous reward attached a good feeling to the action that produced the reward. Critically, these learned habitual responses can be chained together to produce adaptive sequences of actions, using what is known as temporal difference reinforcement leaning (TDRL). The rat can represent the arrival in the red room (which puts it one move away form a reward) as a reward in itself. Then, if the rat subsequently enters the red room by turning left from a blue room, it can acquire a positive feeling about turning left in the blue room and encode a subsequent arrival in the blue room as rewarding, thus setting up the next step in the learning process. By chaining together a set of context-dependent habitual responses, the rat can learn to wind its way through a maze to a reward. In so doing, it's guided not by an understanding of where it's going and how it's getting there, but simply by feelings that tell it what to do at each stage of the process.[3]

Neither strategy, model-based or model-free, is inherently superior to the other. Instead, these strategies have their respective strengths and weaknesses. Model-based learning and decision-making is computationally expensive. A model-based reasoner must store a lot of information about the environment (the causal model) and, when deciding what to do, must search through the entire decision tree to identify a sequence of actions that will realize the goal. (A complete search is necessary if the agent desires an optimal solution and has no algorithms available for simplifying the search process.) The model-free strategy, by contrast, is computationally cheap. The model-free agent need not store a map of the whole territory. Instead, it can simply associate positive or negative feelings with action-context pairs. And, more importantly, the model-free agent need not search through a complex tree of possible action sequences. It can simply choose the action that feels best at any given moment.

An important corollary to the increased computational costs of the model-based strategy is the increased probability of error. For every additional step involved in acquiring, maintaining, searching through, and optimizing over a causal model of an environment, there is an additional opportunity to make a mistake. In a world in which model-free learning will suffice, the model-based agent runs the risk of "over thinking".

The model-based strategy, while disadvantaged by its computational expenses, has the advantage of flexibility. Returning to our hypothetical rat, if the location of the reward changes, or if the starting location changes, or if the learned pathway is suddenly blocked, the model-free rat may have to start over, and under the burden of counterproductive habits. By contrast, a model-based rat that has mapped out the territory can adjust to these changes simply by updating its map to account for a new goal location, new starting point, or a newly blocked (or opened) passage. And there is, as one might expect, evidence that rats can, in a limited way, engage in model-based learning and decision-making (Moser, Kropff, & Moser, 2008; Tolman, 1948). Such instances of model-based learning and decision-making are limited, however, in that they are domain-specific and, correspondingly, do not

[3] See also classic work on "somatic markers" in decision-making (Bechara, Damasio, & Damasio, 2000; Bechara, Damasio, Tranel, & Damasio, 1997; Damasio, 1994).

involve the kind of "slow", highly flexible and multimodal executive processing supported by the prefrontal control network in humans.

A related disadvantage of the model-free approach is that a model-free agent can't act in an explicitly goal-directed way and therefore cannot prospectively consider, or reconsider, the value of a goal or sub-goal. Cushman (2013, pg. 279) explains:

> The contrast between these algorithms is elegantly captured by the devaluation procedure, a well-studied behavioral paradigm (e.g., Dickinson, Balleine, Watt, Gonzalez, & Boakes, 1995). A rat is trained to press a lever to obtain a food reward. During training, the rat is kept on a restricted diet to motivate performance. But then a critical test is performed: The rat is taken out of the apparatus, fed until it shows no more interest in food, and then immediately returned to the apparatus. Under some conditions, it is observed to resume pushing the lever, even though it now has no desire for food. In other words, although the food reward has been "devalued" through satiation, the habitual behavior remains intact. This apparently irrational action is easily explained by a model-free mechanism. The rat has a positive value representation associated with the action of pressing the lever in the "state" of being in the apparatus. This value representation is tied directly to the performance of the action, without any model linking it to a particular outcome. The rat does not press the lever expecting food; rather, it simply rates lever pressing as the behavioral choice with the highest value. A model-based algorithm, in contrast, has the capacity to recognize that the specific outcome associated with pressing the lever is food. Thus, because the rat does not desire food, it would place little value on the action of pressing the lever.

Here, it's not the decision environment that's suddenly changed, but the rat and its values.

Some rat decisions involve competing values. Suppose that a rat is first trained repeatedly to obtain and consume a food reward. Then, a substance that induces nausea a few hours post-consumption is added to the food. Because the rat is already in the habit of eating the food, guided by model-free learning, it may take several bouts of nausea for the rat to learn that the food is no longer worth eating. By contrast, a rat with more minimal prior training, and therefore more model-based in its decision-making, will more quickly adjust to this new reality and stop eating the tainted food (Adams & Dickinson, 1981).

Thus, a creature may decide poorly if it's slow to adjust to a changing world. However, even in a perfectly stable environment, a creature can go terribly wrong if it fails to pick up on the right cues. Suppose that a rat's favorite food is tainted with a potent poison, undetectable to the rat, which builds up slowly in the rat's system over long periods of time and ultimately causes death. Here, the problem is not that the rat's world has changed (as in a relocated reward) or that what's good for the rat has changed (as in "devaluation" through satiation). Instead there is a more fundamental misalignment between subjective reward and value. In this world, the rat's brain provides bad training, mistaking something very bad for something very good. (I'm here putting aside legitimate questions concerning the nature of the "the good" for rodents.) This erroneous reward signal would be a problem for any rat, whether it's relying on a model-based strategy or model-free strategy. Nevertheless, there is a sense in which a model-based rat is closer to a solution. If only the model-based rat could understand English, you could explain to it that, in this case, indulging his tastes will lead to consequences that the rat would regard as very, very bad. A rat with a model has the cognitive infrastructure necessary to represent consequences, the values of consequences, and the causal relations between actions and consequences. To avoid this horrible death, a model-based rat would "simply" need to expand its model. By contrast, if you were to attempt to inform and reason with a model-free rat about this problem, it would respond to your exhortations about poisonous cheese like this: "Hey! Nice lever!". The only way to persuade a model-free rat would be to add some nausea-inducing stuff to its favorite food until its learned its life-saving lesson. Of course, it's a bit silly to talk about rats that speak English, but not as silly as one might think. This is because humans speak English (among other languages) and we, like our distant rodent cousins, must ultimately rely on some combination of model-based and model-free strategies for learning and deciding. More on this shortly.

All of this points, now in a more precise way, to the conclusions outlined above and suggested by an analogy I've presented earlier (Greene, 2013, 2014a): Our intuitions are like the automatic settings on a digital camera ("portrait mode" "landscape mode"). They are efficient, but inflexible. By contrast, our capacity for deliberation is like a camera's manual mode, in which all of the relevant settings can be adjusted by hand. The manual mode is very flexible, allowing one to set a goal and devise and implement a plan for achieving it—that is, to use an explicit model. But the manual mode is not very efficient, and it introduces additional opportunity for error. As the camera analogy suggests, it's not that intuition ("automatic settings") is inherently bad or that deliberation ("manual mode") is inherently good, or vice versa. It may be perfectly sensible (and unavoidable) to rely on one's intuitive "automatic settings" most of the time. These two approaches to judgment and decision-making are just different, with complementary strengths and weaknesses.[4]

Our computational perspective on learning allows us to flesh out this analogy a bit more: First, intuitive decision-making is likely to fare poorly in a changing world. More specifically, intuitions do poorly when the causal relationships between actions (in context) and consequences differ between the world in which the intuitions were acquired and the world in which they are subsequently deployed. This is illustrated by the model-free rat whose string of habitual responses cannot easily adjust to a relocated cheese. Second, even in a stable world, affective learning does poorly when there is a mismatch between the values implicitly embodied in the training/learning process and the values of the agent. This point is illustrated by the rat who continually eats poisoned food because it has received no signal urging it to do otherwise. (Once again, this kind of bad training can be a problem for model-based rats as well if the model is incomplete. See footnote 5).

Critically, these conclusions reflect *fundamental* limitations on intuitive, affective decision-making. Model-free learning strategies can get far more sophisticated than those described here (e.g., Mnih et al., 2015), but no matter how sophisticated a learning process is, if it informs decision-making by attaching values directly to actions based on prior reward history, then it will be subject to these limitations. No intuitive decision-maker can overcome the

[4] I wish to highlight a potentially misleading feature of the camera analogy. The camera's automatic settings do not change once it's left the factory, but people's "automatic settings" are constantly evolving through learning. With that in mind, a better analogy would be a "smart" camera that adjusts its automatic settings based on user feedback, such as which photos are kept vs. deleted. (Thanks to an anonymous reviewer for this suggestion.) The key point, however, is that at the time of decision one is stuck with the automatic settings that one has, regardless of how circumstances might have changed. See above [pg. XX] regarding flexibility in the deployment vs. acquisition of intuitive responses.

Please cite this article in press as: Greene, J. D. The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.004

problems that result from unrepresentative training data or from a training algorithm with misaligned values.[5]

What's more, this is not just a problem for decisions that are immediately determined by "fast" intuitive responses. It applies, as suggested above, to slower, more deliberative decisions that are influenced by "fast" inputs. One might hope that a more extended deliberative process will weed out the bad intuitive inputs and boost the good ones, but there is no reason why this has to be (Greene, 2014a). In the absence of some explicit (model-based) understanding of how our intuitions are likely to go wrong, an integration over good and bad inputs may simply produce a judgment that is a compromise between good and bad inputs. And if the bad inputs enter with greater strength, they may simply dominate the decision, however extended and integrative that decision process may be. As computer scientists say, it's GIGO: garbage in, garbage out.

5. Two goals of moral philosophy

What, then, does this mean for moral philosophers and others who aspire to answer moral questions? Let's retrace the argument's main thread: Some researchers have raised doubts concerning the reliability of intuition in general (Kahneman, 2003, 2011) and moral intuition more specifically (Greene, 2007, 2013, 2014a; Singer, 2005), characterizing intuitions as "fast" and correspondingly inflexible. I, among others (e.g., Singer, 2005; Wright, 1994), have also argued that our moral intuitions may be biased by genetic influences (Greene, 2003, 2013, chap. 1–3). This combination heightens concerns about the rigidity of our moral "instincts": What if our moral intuitions are stuck in the Pleistocene epoch? Even worse, what if the biological directive to spread our genes, both now and in the past, gives us moral instincts that are good from a biological perspective and bad from a moral perspective?[6] I think these considerations give us good reasons to worry about our moral intuitions. But Railton's portrait of moral intuition would seem to offer more hope: If our moral intuitions are acquired through individual experience, rather than inherited from our ancestors, and if they are acquired through a sophisticated learning process, rather than by other means, does this not give us legitimate cause for optimism concerning their normative authority?

The answer, as we now know, is, "It depends". Railton knows this, too. In a section entitled "Limitations" (pp. 845–846) he previews the conclusions reached above:

> Statistical learning and empathy are not magic—like perception, they can afford only prima facie, defeasible epistemic justification. And—again, like perception itself—they are subject to capacity limitations and can be only as informative as the native sensitivities, experiential history, and acquired categories or concepts they can bring to bear. If these are impoverished, unrepresentative, or biased, so will be our statistical and empathic responses.

Thus, as noted at the outset, we have no fundamental disagreement concerning the general strengths and weaknesses of learned affective intuition. And, as it happens, my failure to disagree with Railton on this point recapitulates an earlier episode from the borderlands between descriptive and normative psychology. Daniel Kahneman (representing the "heuristics and biases" tradition, focused on errors of judgment) and Gary Klein (representing the "naturalistic decision making" tradition, focused on skilled judgment) co-authored a paper entitled "Conditions for Intuitive Expertise: A Failure to Disagree" (Kahneman & Klein, 2009). They conclude that, "evaluating the likely quality of an intuitive judgment requires an assessment of the predictability of the environment in which the judgment is made and of the individual's opportunity to learn the regularities of that environment." (pg. 515).

These shared conclusions point to the next question: Are we, as normative moral thinkers, deploying intuitions produced by good training algorithms operating in a good learning environment? Or are our moral intuitions "impoverished", "unrepresentative", or "biased" by our "native sensitivities"? Or, to frame our question in the most practically useful way: Are we more likely to *overestimate* or *underestimate* the reliability of our moral intuitions?

With the question framed thus, my answer is clear: We give our moral intuitions way too much credit (Greene, 2013, 2014a). For Railton, the observation that our affective moral intuitions are subject to certain systematic limitations serves as a cautionary caveat, following a generally optimistic account of their authority. In my opinion, this story buries the lede. For moral philosophers, both lay and professional, the limitations of our intuitive moral thinking ought to loom large.

[5] Here and throughout, I follow Cushman (2013) and Crockett (2013) in aligning intuition with decision-making based on model-free learning while aligning explicit, conscious reasoning—in particular, reasoning based on the evaluation of consequences and the deployment of beliefs about causal relationships between actions and consequences—with model-based decision-making. I think this is defensible, for reasons given here, and more elaborately by Cushman and Crockett. However, there is a noteworthy wrinkle. Judgment driven by model-free learning is necessarily intuitive, and explicit reasoning about actions, values, and consequences is necessarily model-based. However, some judgment driven by model-based learning may also be intuitive. A prime example is the model-based navigational system employed by mammals such as rodents (Moser et al., 2008; Tolman, 1948) and humans (Doeller, Barry, & Burgess, 2010). Despite employing a cognitive map, rats are presumably not deploying the kind of "slow", domain-general reasoning abilities enabled by the human fronto-parietal control network. In other words, this kind of thinking is more "fast" than "slow", certainly in the case of rats and presumably much of the time in humans. Likewise, when humans effortlessly comprehend and produce novel sentences or apply their knowledge of "folk physics", they may be relying, unconsciously and intuitively, on cognitive causal models. This observation matters for the argument made here (see below) because it could be that some moral intuitions are in fact model-based, and therefore not subject to all of the limitations of model-free learning and decision-making. At the same time, there are good reasons to think that *intuitive* model-based reasoning will indeed be limited in ways that parallel the limitations of model-free learning and decision-making. The reason is that domain-specific, unconscious reasoning of any kind is bound to be rather inflexible, even if it exhibits some notable flexibility within its domain. Take, for example, the case of linguistic intuition. If the linguistic environment were to suddenly change so as to make an alternative grammatical structure desirable, it would nevertheless be very difficult, probably impossible, to change one's linguistic intuitions on the spot. The newly grammatical sentences would still sound "wrong", and making the correction will be much harder than, say, changing one's driving route upon learning that the location of one's event has changed. Critically, this inflexibility is likely to exist whether those linguistic intuitions were produced via model-free learning or some kind of domain-specific, unconscious model of one's language. In the same way, even if some moral thinking relies on unconscious moral models, it will be severely limited and inflexible insofar as those models are unconscious, since this implies that they cannot be directly accessed and adjusted by the kind of high-level, "slow" thinking that makes human thinking uniquely flexible. As suggested above, this point is illustrated by the case of the rat that unknowingly consumes poison (pg. XX): Even if the rat's behavior is guided by a model, the rat is doomed if it cannot integrate information about the poison into that model. At this point one might wonder why, for present purposes, we should even bother talking about model-free vs. model-based learning, if what really matters is the distinction between efficient "fast" thinking and flexible "slow" thinking. I am sympathetic to this point. Indeed, I think that the present consideration of model-based vs. model-free learning serves to bolster a more general argument based on the limitations of intuitive moral judgment (Greene, 2007, 2013, 2014a). Incorporating these newer ideas is very useful, however, because some moral judgments, in addition to being intuitive, seem to have the precise signatures of model-free learning, which enables us to make a more precise and compelling diagnosis of their likely limitations.

[6] The most straightforward concern here is that moral intuitions may be either directly self-serving or indirectly self-serving by favoring in-group members (tribalism). Add to this the possibility that moral judgment maybe distorted by motivations to engage in social signaling, advertising one's reliability as a cooperation partner through words and deeds that may destroy more value than they create (Everett, Pizarro, & Crockett, 2016; Jordan, Hoffman, Bloom, & Rand, 2016), as in the case of mass internet shaming (Ronson, 2016).

Please cite this article in press as: Greene, J. D. The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.004

As noted at the outset, this conclusion depends on a specific conception of moral philosophy's primary purpose. Once again, I believe that moral philosophy's principal aim should be to *solve moral problems*, to help us resolve practical moral disagreements concerning live, controversial issues.[7] I will not defend this philosophical orientation here. It is simply my starting point, and it is by no means shared by all moral philosophers. Aristotle (1941), for example, is not principally concerned with resolving moral controversies. Instead, his philosophy is primarily descriptive, characterizing what he believes are the psychological features and behavioral tendencies of those who are generally, and generally regarded as, virtuous. Likewise, Kant (1785/2002) aims to organize and justify what he regards as good morals and is not focused on changing people's minds about controversial moral issues. He wants to explain why the wrongs of everyday life are wrong, deriving from first principles the inherent immorality of lying, stealing, and killing (Kant, 1785/2002). (And also masturbating; Greene, 2007; Kant, 1930). His goal is to put his preferred version of commonsense morality on a solid deductive foundation, analogous to the foundational principles of mathematics.

As explained below, I think it's no coincidence that Aristotle and Kant, with their relatively optimistic views of everyday moral conviction, are not primarily focused on resolving moral disagreements. The contrasting, problem-solving approach to moral philosophy finds its clearest historical expression in the founding utilitarians: Bentham (1781/1996), Mill (1863/1987), and Sidgwick (1907). Like Kant, they sought to organize moral thinking, but they were also social reformers, challenging some of the most firmly held moral convictions of their day. They argued against slavery, supported free speech and free markets, and defended what we now call women's rights, worker's rights, animal rights, and even gay rights (Bentham, 1978; Driver, 2009). As I will explain, it's likewise no coincidence that these social reformers questioned the authority of moral intuition. In the next two sections we'll consider two ways by which biased moral intuitions can be acquired.

6. Applied moral philosophy and the "bad training problem

When focused on real-world moral disagreements, it's hard to be anything but skeptical about people's moral intuitions. You might think that your own moral intuitions are splendid, but it's hard to feel that way about moral intuitions in general. This is because, as Haidt (2001, 2012) has compellingly argued, most real-world moral disagreements are fueled by conflicting moral intuitions. When people disagree about abortion, capital punishment, gay marriage, or the distribution of wealth within and across nations, they are not (merely) disagreeing about relevant facts and principles. They have very different gut feelings about what's right or wrong in these cases. When moral intuitions conflict, moral intuitions must be wrong at least 50% of the time.

Nevertheless, this pessimistic account of moral intuitions in conflict is consistent with the optimistic thought that most people's moral sensibilities are pretty good for most everyday purposes. As I've argued elsewhere (Greene, 2013), it's useful to divide moral problems into two categories: "Me vs. Us" and "Us vs. Them". The moral problems that we're forced to address in our everyday lives are generally of the "Me vs. Us" variety. They are about cooperation in the broadest sense, the "Tragedy of the Commons" (Hardin, 1968), the perennial tension between self-interest and the interests of others. The most familiar and basic moral norms, the kind we first learn as children, address the "Me vs. Us" problem, placing restrictions on what any given Me may or may not do to a member of Us: No hitting, stealing, lying, promise-breaking, etc. These relatively concrete norms are variations on the most celebrated (and variously interpretable) norm of all, the Golden Rule, which instructs people to treat the other members of Us the way one would like to see Me treated.

Returning to Railton's case for optimism about moral intuition, it's worth noting that his examples of successful affective intuition are drawn from life within the tribe. Consider, first, the lawyer. She wins her case through adept social cognition, reading the thoughts and feelings of others and figuring out what she needs to say and do in order to win their agreement. (And this need not be cynical or merely strategic. She may very much believe in the argument she is making.) In light of our discussion of learning mechanisms, her success should come as no surprise. After a lifetime of negotiating relationships with other humans, and after years of trial-and-error in the courtroom, we would expect her social "instincts" to be well attuned. She's had a good, representative training set and a good, value-aligned training algorithm.

Now consider, once again, the case of Mark and Julie, the incestuous siblings. Here, too, our sense that their behavior is risky and foolish is grounded in relevant experience, but in this case the experience is more likely to be evolutionary and cultural rather than personal. The gut feeling that says, "I don't even want to think about it" is not a particularly controversial one, signaling a split between moral tribes. Instead it comes, at least in part, from our species' learning over evolutionary time that matings between close relatives are likely to produce diseased offspring (Lieberman, Tooby, & Cosmides, 2007). In addition to whatever genetic dispositions against incest we may have, human tribes may have also learned over historical time that sexual relationships and sibling relationships don't mix well. Here, too, we have what is plausibly a good data set and a good training algorithm. With respect to the biological basis for the incest taboo, the data may not be representative because Mark and Julie are (by stipulation) using adequate birth control. But insofar as Mark and Julie really are playing Russian roulette with their relationship, that implies that such behavior carries a high risk of bad consequences. And that implies that the intuitive judgment that Railton here defends is backed up by good data.

Cases like that of Mark and Julie contrast with cases of public moral controversy. These are conflicts, not between isolated individuals, but between groups with competing moral values and interests, cases of "Us vs. Them". Such problems are beset by what I have called the "Tragedy of Commonsense Morality" (Greene, 2013, chap. 3). This is the problem of cooperation one level up, at the level of groups rather than individuals, fueled by incompatible moral intuitions that are common sense within groups, but not between groups. Here the problem arises not from simple selfishness, but from tribalism, selfishness at the level of groups. It also arises from disagreements about how social life should be organized, disagreements concerning the terms of cooperation within a tribe: Collectivist vs. individualist (Henrich et al., 2001), punitive vs. peace-making (Nisbett & Cohen, 1996), egalitarian vs. hierarchical (Kahan et al., 2011, 2012), and so on. In addition to favoring different terms of cooperation, tribes differ in their attitudes about sex and death, the gas pedals and breaks of tribal growth. Finally, they differ in their "proper nouns", the particular individuals, texts, traditions, etc. that are invested with moral authority, giving rise to many moral differences that are effectively arbitrary.

My general conclusion is that our moral intuitions do a pretty good job of dealing with the perennial problems of everyday moral life ("Me vs. Us") and do a much worse job of handling disagreements across tribes ("Us vs. Them") (Greene, 2013). This conclu-

[7] One might suppose that moral philosophy, in addition to resolving disagreement, should also work to disrupt unjustified agreement. I agree (with good justification!) that philosophy may play such a role, but in practice at least one person must disagree, or at least harbor a doubt, in order to begin such a process. I would be very surprised if there are examples of societal shifts in moral attitudes that did not begin with advocates who rejected the prevailing consensus.

sion, which fits well with the available psychological research (chap. 2–3), also makes perfect evolutionary sense (chap. 1). To the extent that we evolved biologically for social life, it's because living together as social beings provides a competitive advantage. We are able to put ''Us'' ahead of ''Me'' precisely because our capacity for cooperation within a tribe, our capacity for teamwork, enables Us to outcompete Them. Evolution does not select for creatures that are universally cooperative, as there is no competitive advantage in this, and natural selection depends on competition. Nor is it possible for our genes to encode specific programs that tell us how to be successful selective cooperators across the full range of possible human tribes. Instead, we have evolved to *learn* how to be successful members of the tribes into which we are born (Haidt, 2001, 2012; Henrich, 2015). And this requires the integrated operation of ''native sensibilities'' and sophisticated learning abilities.

What this suggests, most immediately, is that our moral intuitions concerning members of other tribes are likely to be unreliable. That is, we may be prone to racism, nationalism, xenophobia, etc. And, indeed, Railton identifies racism as a critical failing of affective intuition (pp. 845–846). But the problem of tribalism is, as suggested above, much broader than that. Morally contentious topics such as abortion, capital punishment, climate change, and national health insurance are not explicitly about ''Us vs. Them'', but our views on these matters are likely to be heavily influenced, and in some cases completely dominated (Cohen, 2003), by our tribal identities, allegiances, and interests. As Kahan and colleagues have argued (Kahan et al., 2011, 2012), persistent political disagreement over what would appear to be purely factual issues, such as the reality and causes of rising global temperatures, make a lot more sense if we view people as attempting to solve a local social problem rather than a global empirical problem. What a sophisticated affective learner may glean is that, around here, believing in climate change (or doubting its reality) is not very rewarding. And likewise for our attitudes about abortion, gay marriage, government-supported healthcare, terrorism, and so on.

Thus, when it comes to the moral problems that divide us, our intuitive judgments are dogged by the second problem identified above, the bad training problem. It's possible that, sometimes, our sophisticated affective learning mechanisms enable us to learn our way out of our tribal biases. To take an example of Railton's (pp. 850–851), many Americans have recently shifted their views about same-sex marriage, and gay rights more generally. Railton suggest, very plausibly, that this shift was largely intuitive, promoted by feelings of connection with, and empathy for, relatives, friends, and co-workers who, until recently, would have remained closeted.[8] More generally, there are good reasons to think that we are slowly developing post-tribal, moral sensibilities (Pinker, 2011). But given that our tribal disagreements are generally fueled by conflicting moral intuitions, and assuming that these culturally variable intuitions are not differentiated genetically, we can be confident that our affective learning mechanisms plenty often get us *into* our tribal biases.

With this in mind, we can revisit Railton's lawyer case, which includes an intertribal element that I've not yet mentioned. The defendant, who's been accused of murder, is a member of a disadvantaged minority. In Railton's story, the lawyer succeeds by making a successful emotional appeal to the jury's higher ideals of justice. At one point she urges them, with a subtle pause, to prevent racial bias from clouding their judgment: ''My client just happened to be the right height, weight, . . . and color . . .'' I assume that Railton's successful lawyer is on the side of the angels in this case, although we're never told that explicitly. Assuming she is, let's consider a version of this story in which she's less successful thanks to a more formidable opponent. Suppose that the opposing prosecutor is even more affectively attuned than she is. He plays his ''dog whistle'' perfectly, subtly activating the jury's fears and prejudices, all the while obscuring the evidence just enough to allow those darker feelings to take control. And thus an innocent man is convicted of murder.

As noted above, Railton is aware that affective attunement can be destructive in precisely this way. My point in twiddling the nobs on this example is to highlight the persuasive role played by factors that are, from a purely psychological perspective, incidental in the original example. The original case of the attuned lawyer paints an optimistic picture of moral intuition, not only because it illustrates the subtlety and complexity of intuitively informed judgment, but also because it aligns, in an entirely optional way, the lawyer's being socially and emotionally attuned with the lawyer's being morally right. Psychologically speaking, the winner in Railton's original case is not so much *morally good* attunement, but *socially effective* attunement, which may be used for good or ill. And, as I've emphasized, ''for ill'' is not at all uncommon when moral tribes clash.

7. Normative ethics and the ''bad data problem: a tale of three rats

Most moral philosophy is not about specific, real-world controversial topics such as abortion and raising taxes. Instead, moral philosophers address more general, theoretical, and abstract problems. This is for good reason. The deeper one digs into specific moral controversies, the more one encounters recurring abstract questions: Which entities deserve full moral consideration and why? What makes someone a responsible moral agent? What role should personal loyalties play in one's decisions? How should we weigh individual rights against the good of others? If we are to find philosophically satisfying answers to specific, practical moral questions, it seems, we're going to have to find some satisfying general moral principles. And this is what many philosophers, myself included, have set out to do.

How do you know if you've found a good principle? One strategy for finding good principles is to logically derive them from self-evident premises. While there have been some famous attempts to do just this (Kant, 1785/2002), there are no substantive moral principles (ones that can be used to resolve real moral controversies) that are generally agreed to have been so derived. In lieu of moral theorem-proving, a more widespread approach to testing moral principles is to consider their implications under various circumstances and to ask whether those implications seem morally acceptable. In other words, we can test moral principles in something like the way we test scientific theories. But rather than testing moral theories against empirical data, we test them against our intuitions or ''considered judgments'' about specific cases, or types of cases. Principles that seem to get things mostly right, but not completely right, may then, like scientific theories, be modified in hopes capturing more of the ''data''. This method, which is very old (Plato, 1987), and perhaps inevitable, has been dubbed the method of ''reflective equilibrium'' (Rawls, 1971).

With this ubiquitous approach to moral theorizing in the background, I'll make two points in this section: First, the method of

[8] While this is certainly a case in which views have shifted on a matter of intertribal disagreement, it may also be an unusual case in which tribal commitments have facilitated the march of moral progress. Many have marveled at how quickly attitudes on same-sex marriage (etc.) have shifted, especially in light of what seems to be slower progress on matters of race. This difference in the speed of change may be due to the fact that many gay people have an insider tribal advantage. Gay people are routinely born into families and cultures that are hostile to gay people, but a White couple with strong racial biases, for example, will not unexpectedly find themselves the parents of a Black child. The rapid advance in gay rights, then, may be because the moral tension is as much *intra*tribal as it is *inter*tribal. If gay people were only born into gay-friendly families and cultures, I suspect that the cause of gay equality would not have progressed so quickly.

Please cite this article in press as: Greene, J. D. The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.004

reflective equilibrium is unlikely to succeed so long as we allow our moral intuitions to escape scientifically informed reevaluation. And this is true even when our intuitions are acquired through sophisticated learning processes. To make progress, we must aim for what I have elsewhere called "double-wide" reflective equilibrium (Greene, 2014a). This means factoring into our moral theorizing, not only candidate principles and particular judgments, but also an understanding of the strengths and limitations of the cognitive processes that produce those principles and judgments. I have argued, more specifically, that when we factor in a better scientific understanding of moral psychology, consequentialism[9] will become more attractive (Greene, 2007, 2013, 2014a). My second point in this section—as a case in point—is that a computational perspective on moral learning makes consequentialism more attractive. More specifically, the foregoing account of moral learning explains why there is a natural alignment between consequentialism and the more flexible type of learning described above, i.e. model-based learning and deciding. (See also Cushman, 2013 and Greene, 2013, chap. 8) Combine this with the premise that difficult moral problems require model-based thinking, and we have a case for consequentialism.

Consider the rat in the devaluation paradigm. After being extensively trained to press the lever for food, the rat is removed from the apparatus, allowed to feed *ad libitum*, and then returned to the apparatus. Upon return, he presses the lever over and over, leaving a pile of unwanted food pellets on the floor. Were this an English-speaking rat, you might ask him what he's doing: "You don't seem to want the food. Why do you keep pressing?" If he's a simple sort of rat, he might shrug and say, "I don't know. I just feel like pressing." Or perhaps a lame excuse: "I really need the exercise". A more philosophically minded rat, however, might insist that he presses the lever for a more noble reason. He does this, he explains, not to receive some crass reward, but out of a sense of duty. And if such a rat were inspired by the success of mathematics, he might attempt to derive this rat-a-gorical imperative from principles of pure rodent reason.

A different rat philosopher might conceive of this question—to press or not to press?—as a fundamentally a matter of character. She might opine that any rat who can sit, unmoved, in an apparatus like this, with a lever like that, must be poorly attuned to the subtle contours of rodent virtue.

Finally, a third philosopher rat might observe herself pushing the lever, recognize that it's not doing anyone any good, and stop pressing. She might feel a bit uncomfortable about this. Like the other rats in this experiment, she has the strong sense that the lever simply *must be pressed*. And, what's more, she knows from experience that such feelings are generally worth heeding. But she also knows that, in this unusual situation, foisted upon her by fiendish experimenters, her feelings are very likely misguided. And so, with some hesitation, she decides to take a rest and save her strength for some more productive enterprise.

Humans, of course, live in more complicated worlds and have more nuanced philosophical ideas than these whiskered versions of Kant, Aristotle, and Mill. But with respect to the underlying neurobiology and corresponding cognitive mechanisms, these rodent-human comparisons may be startlingly apt (Blair, 2007; Crockett, 2013; Cushman, 2013). Consider, as ever, the footbridge case. Ample evidence, which I will not review here (Greene, 2013, 2014a, 2014b), shows that people judge it wrong to push the man off the footbridge because of an affective response to actions of this sort, independent of their net consequences (Miller, Hannikainen, & Cushman, 2014). What's more, the neural circuitry that enables this response (involving the amygdala and vmPFC, among other regions) appears to be the very same circuitry that, in a rat, would prevent it from pressing a lever that has previously delivered a shock (Blair, 2007; Crockett, Clark, Hauser, & Robbins, 2010; Glenn, Raine, & Schug, 2009; LeDoux, 2000; Phelps, Delgado, Nearing, & LeDoux, 2004; Shenhav & Greene, 2014).[10]

With respect to this action's triggering features, what seems to matter is that this harmful action is active, intentional (as opposed to merely foreseen, or accidental), and that it involves pushing, i.e. a direct application of "personal force" (Cushman et al., 2006; Greene et al., 2009). This is not the last word on the distinguishing features of the action in the footbridge case, and, as Railton observes, one can construct cases involving actions that have these features and that, nevertheless, don't seem quite so bad.[11] Nevertheless, these features do seem to have a reliable effect. Indeed, they seem to define our very concept of "violence": It's hard to envision an action by which one person harms another actively, intentionally, and through the direct application of personal force that is not naturally described as "violent".

How is it, then, that we are disposed to respond so negatively to the action in the footbridge case? Why does it make the hair on our amygdalas stand up (Glenn et al., 2009; Shenhav & Greene, 2014)? As Crockett (2013) and Cushman (2013) argue, following on related ideas from Blair (2007), it's probably because we've used our mammalian brains to learn from our own experiences, and indirectly from those of others (Olsson, Nearing, & Phelps, 2007; Olsson & Phelps, 2007), that violent actions tend to have bad consequences. Such actions are bad for the victims. This, combined with some empathy or sensitivity to distress cues (Blair, 1995; Blair, Jones, Clark, & Smith, 1997; Singer et al., 2004), may make such actions aversive. Moreover, in a typical social group, such actions are likely to elicit punishment, whether through social disapproval, material deprivation, or bodily harm. In any case, the available evidence suggests that we recoil at the actions that we call "violent" because we have, either directly or indirectly, been trained to find such actions aversive. And this is a good thing because, once again, such actions tend to produce bad *consequences*.

With all of this in mind, let's now consider the fiendish perversity of the footbridge case. What this case essentially does is take a kind of behavior that very reliably produces bad consequences in the real world and says, "Suppose that this is guaranteed to produce the best possible consequences. Now do you like it?" The

[9] Consequentialism is the view that consequences are the ultimate source of value and that whether actions are right or wrong is ultimately a function of their consequences alone. Utilitarianism is the more specific view that consequences should be assessed impartially and that consequences are to be measured in terms of their effects on experienced well-being—roughly, the reduction of suffering and the promotion of happiness.

[10] It's worth noting that the philosophical devaluation paradigm and the footbridge case differ in that the relevant affective responses are positive in the first case and negative in the second. The neural substrates of reward and punishment are highly overlapping, but not identical.

[11] In Railton's variation one saves oneself and a busload of others by pushing an innocent person into a terrorist with a bomb, who is about to board the bus. Both get actively and intentionally pushed, and die from the explosion as the bus pulls away. In his informal surveys, Railton finds that most people approve of this action. This is an interesting and complicated case, with many additional factors introduced. It may turn out that this action does trigger a negative emotional response similar to that of the footbridge case, but that it is, due to the context, easily overridden or defused. Speaking for myself, I would feel better about allowing the man to fall into the terrorist and out the door than I would feel about pushing him. And I would feel better about knocking them both off the bus as a foreseen side effect of pulling a lever. But I take Railton's point that the features identified above and in previous experiments (action, intention, personal force) need not be entirely determinative of the emotional response to the action. Our affective responses may be attuned to other contextual features that we've yet to explore. However, what matters for present purposes is something that Railton would not dispute, which is that our affective responses to actions such as these depend on their direct or indirect reinforcement histories and are therefore likely to give biased responses if trained with unrepresentative data.

Please cite this article in press as: Greene, J. D. The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.004

footbridge case takes the moral maze of everyday life and turns it upside down. It transforms behavioral poison into reward, reversing the most reliable of connections between social action and social consequence. Should we trust our intuitions about this case? Everything we have learned about affective learning says "No". If our task is to evaluate a case in which—truly—the best possible outcome is produced by an act of lethal violence, perpetrated against an innocent person, then we are operating with *very bad training data*.

And yet philosophers have generally drawn a very different conclusion from cases like this one. The general lesson that philosophers have taken from this and similar cases is that consequentialism, and utilitarianism more specifically, must be wrong: Any theory that endorses pushing the man off the footbridge must be deeply flawed. Of course, this conclusion was not reached solely because of the footbridge case. For decades, ethicists have reflected upon, and reacted strongly to, the hypothetical consequences of promoting the greater good. Rawls (1971), for example, asks us to evaluate a society in which happiness is maximized when a majority enslaves a minority. Nozick (1974) asks us to imagine a monster that gets more happiness out of eating a person than a person gets out of her entire life. Sandel (2009), in his introduction to moral and political philosophy, devotes most of the chapter on utilitarianism to a string cases in which doing horrible things is artificially stipulated to promote the greater good. Examples include Romans delighting in the spectacle of lions tearing Christians to pieces, a society whose well-being inexplicably depends on a single child's being locked away in miserable solitude, and, for good measure, the footbridge case. Of course, there have been many abstract theoretical arguments levied against consequentialism, but in my estimation (and the estimation of countless introductory textbook authors), it's our gut reactions to this theory's hypothetical implications that really puts it on the ropes.

Elsewhere I have offered a more systematic, scientifically fortified, defense of consequentialism/utilitarianism (Greene, 2013), and that is not my purpose here. Here, my more specific point is this: Given what we know about the mechanics of affective learning, we cannot possibly give consequentialism a fair hearing if we insist on evaluating it based on our feelings about its hypothetical implications. To do this is to insist that we rely on intuitions trained on *bad data*. The data are bad, not because they are bad for guiding judgment in everyday life ("Me vs. Us"), but because they are bad for guiding judgment in hypothetical worlds in which the *usual relationships between actions and consequences are reversed*. All of this suggests, more generally, that we can't trust our intuitions about strange hypothetical cases.

At this point, some readers may be surprised to hear me, a longtime fan of trolley dilemmas, say such a thing. To be clear, I am a fan of trolley dilemmas as *scientific tools*, not as *normative guides*. Trolley dilemmas are useful, not because they are representative, but because they are artificial high-contrast stimuli that enable us to dissociate cognitive processes that are otherwise hard to dissociate (Cushman & Greene, 2012). The understanding that we gain from studying trolley dilemmas (etc.) helps us understand why we *shouldn't* rely on them for normative guidance. Indeed, I got into the business of studying moral dilemmas scientifically because I was, and continue to be, skeptical of their apparent normative implications.

Being wary of weird hypothetical cases (and real cases that are comparably weird) sounds like simple common sense. But, as noted above, much of the most influential moral theory over the last century has ignored this advice. Part of the problem is that the hypotheticals don't need to seem all that strange in order to lead us astray. Consider, for example, Rawls' assumption that utilitarianism could, under some not-too-implausible circumstances, endorse slavery. The idea that slavery could maximize happiness is in fact extremely unrealistic (Greene, 2013, pp. 275–284). For slavery to maximize happiness it would have to be the case that the typical slave owner gets more happiness out of owning a slave than the typical slave loses by being enslaved. Put it this way: Would you choose to be a slave for half your life in order to have a slave (or equivalent economic benefits) for the other half? Is there any realistic world in which your answer would be "yes"? If our collective answer to these questions is a resounding "no", then a world in which slavery maximizes happiness is, in fact, a very, very unrealistic world, one in which "slavery" means nothing like what it means to us in this world.

Testing moral principles against our intuitions about strange hypothetical cases is a general problem for moral theorizing, but this problem is especially acute for consequentialism because this philosophy affords itself such little wiggle room. Because consequentialism is systematic (giving definite answers to all cases, given enough empirical information) and specific concerning what matters (aggregate happiness, in the case of utilitarianism), it gives critics an unlimited ability to construct intuitively damning hypotheticals against it. Virtue-theory is sufficiently vague that no Aristotelian will ever be forced to approve of a nasty-feeling hypothetical behavior. Likewise, if a Kantian with a straightforward principle ("Lying is wrong") is caught in a difficult situation ("But what about saving someone's life by lying to a would-be murderer?; Kant, 1983), the Kantian can always hold out for a more sophisticated interpretation of the principle in question. Such moral theorists give themselves ample wiggle room to avoid getting trapped by unpleasant hypotheticals. But a consequentialist/utilitarian, by being systematic and precise, has nowhere to hide (Greene, 2014a). For any action that feels terribly, horribly wrong because of its typical real-world bad consequences, one can always construct an unrealistic hypothetical world in which its consequences are artificially stipulated to be good. And if we are willing to trust such intuitions, trained up on unrepresentative data, then our moral theorizing will inevitably be distorted.

Before closing, I'd like to address one lingering question about the role of intuition in moral theorizing. It's sometimes said that one cannot avoid relying on moral intuition because all moral judgment ultimately comes down to intuition. For example, in the footbridge case, isn't the intuition that it's wrong to push the man simply competing with the "intuition" that it's better to save more lives? No. And, here, the distinction between model-free and model-based strategies helps us understand why. It is true that the utilitarian judgment ultimately comes down to one or more brute claims of value, such as the claim that happiness is good or the claim that (*ceteris paribus*) more people's being happy is better than fewer people's being happy. But these core utilitarian premises are not tied to specific actions or action types.[12] Instead they are instances of what Sidgwick (1907) called "philosophical intuitions", which he contrasts with "perceptual" and "dogmatic" intuitions about specific actions and classes of actions, respectively. The difference between model-free and model-based decision-making is not that one involves some kind of rock-bottom, affective judgment of value, while the other does not. The difference, instead, is that model-free learning, unlike model-based learning, attaches values directly to *actions* (in context), independent of their consequences. In other words, model-free learning gives us intuitions that are "perceptual" or "dogmatic". By contrast, for a model-based agent,

[12] It's true that judgments favored by utilitarianism may, in addition to being supported by impartial cost-benefit reasoning, be supported by intuition. In fact, this is generally the case for uncontroversial moral judgments. For example: Is it okay to push people off of footbridges because you don't like the way they dress? Here impartial cost-benefit thinking and the dominant, action-based affective response agree. But the critical point above is that utilitarian judgment does not *depend* on the support of such affective responses, as illustrated by cases such as the footbridge case.

Please cite this article in press as: Greene, J. D. The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.004

the value is attached to the *goal*, a consequence, and actions acquire value based on an understanding (a model) of which actions are likely to lead to which consequences. Thus, all moral judgment involves some kind of brute evaluation (an "intuition" of some kind at some level), but some evaluations are of broad goals that apply across countless contexts, while others are automatic responses to specific actions and action types.

Finally, the distinction between model-free and model-based valuation helps us understand, in a deeper way, what consequentialism is really all about (Crockett, 2013; Cushman, 2013; See also Greene, 2013, chap. 8). Consequentialism isn't just another moral theory, supported by it's own set of intuitions. Consequentialism is what you get when you apply model-based thinking to the general problem of morality at the level of first principles. Consequentialism says that the only thing that *ultimately* matters is consequences and that actions derive their value from their relation to consequences. In other words, "the good" comes before "the right". Consequentialism does not say that it is always good to *think* in terms of costs and benefits. Doing this might result in very bad consequences, thanks to our inherent cognitive limitations and biases (Hare, 1981). Instead, consequentialism says that sometimes, probably most of the time, it's best to make moral decisions in a model-free way, relying on what we hope are good moral habits. But, says consequentialism, the *ultimate standard* by which to judge our moral thinking is its consequences.

Deontology and virtue ethics deny this. They say that sometimes the right thing to do is not the action that produces the best consequences, and they say that the right way to think in general may not be the kind of thinking that generally produces the best consequences. What this amounts to is a willingness to favor model-free thinking over model-based thinking at the level of first principles, to insist that our deepest moral insights come from—or just happen to coincide with—a process designed to give good-enough behavioral guidance in a computationally cheap way.

8. Conclusion

Can we trust our moral intuitions? It depends on what we are trying to do. If our goal is to navigate the give-and-take of everyday social life, then our moral intuitions are likely to serve us relatively well. But if our goal is to solve moral problems, to answer controversial or philosophically challenging moral questions, the limitations of our intuitive judgments loom large.

Does it help if our intuitions are acquired through sophisticated learning mechanisms (Railton, 2014)? My answer is a qualified "No". The science of affective learning teaches us, in a more precise way, just how misguided it is to trust our moral intuitions—at least in cases of moral disagreement, both external and internal. My claim is not that our moral intuitions are always wrong in such cases, but rather that they will be wrong too often for us to rely on them. When it comes to real-world moral controversies, our intuitions are too often given bad training by an affective learning system that helps us get along with our tribe-mates by reinforcing our tribal biases. And when it comes to moral theorizing, we fall prey to bad data, relying on moral intuitions that have been trained up in one maze and then tested in another. As moral theorists, we mistakenly expect our intuitive moral imperatives to apply in all possible worlds, no matter how different those worlds are from the world in which our moral intuitions were acquired. We fail to appreciate that our moral feelings are, by their very nature, approximations of something else, and that, sometimes, that "something else" is not worth approximating.

Amidst this pessimism, I want to close by highlighting and amplifying a more optimistic message. As Railton observes, our moral intuitions do not "bear upon their sleeves the seal of validity" (pg. 833). Consequently, he urges us to learn more about the mechanisms of moral learning, to better understand when our moral intuitions are likely to serve us well and when they are not. Thanks to our capacity for "slow," domain-general, model-based thinking, we need not be slavishly bound by our model-free habits. We don't have to press the lever simply because it feels like the right thing to do, and we don't have to build our moral theories around the idea that we ought to. Instead, we can model our own moral thinking and use that understanding to make better decisions. Science can't, by itself, tell us what's right or wrong. But if our goal is to solve difficult moral problems, a scientific understanding of moral thinking may be our best hope for progress.

References

Adams, C. D., & Dickinson, A. (1981). Instrumental responding following reinforcer devaluation. *The Quarterly Journal of Experimental Psychology, 33*(2), 109–121.

Aristotle (1941). Nichomachean ethics. In R. McKeon (Ed.), *The basic works of Aristotle* (pp. 927–1112).

Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral Cortex, 10*(3), 295–307.

Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1997). Deciding advantageously before knowing the advantageous strategy. *Science, 275*(5304), 1293–1295.

Behrens, T. E., Woolrich, M. W., Walton, M. E., & Rushworth, M. F. (2007). Learning the value of information in an uncertain world. *Nature Neuroscience, 10*(9), 1214–1221.

Bentham, J. (1781/1996). *An introduction to the principles of morals and legislation (collected works of Jeremy Bentham)*. Oxford, UK: Clarendon Press.

Bentham, J. (1978). Offences against one's self. *Journal of Homosexuality, 3*(4), 389–406.

Blair, R. J. R. (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition, 57*(1), 1–29.

Blair, R. J. R. (2007). The amygdala and ventromedial prefrontal cortex in morality and psychopathy. *Trends in Cognitive Sciences, 11*(9), 387–392.

Blair, R. J., Jones, L., Clark, F., & Smith, M. (1997). The psychopathic individual: A lack of responsiveness to distress cues? *Psychophysiology, 34*(2), 192–198.

Chase, W. G., & Simon, H. A. (1973). The mind's eye in chess. In W. G. Chase (Ed.), *Visual information processing* (pp. 215–281). New York: Academic Press.

Cohen, G. L. (2003). Party over policy: The dominating impact of group influence on political beliefs. *Journal of Personality and Social Psychology, 85*(5), 808.

Crockett, M. J. (2013). Models of morality. *Trends in Cognitive Sciences, 17*(8), 363–366.

Crockett, M. J., Clark, L., Hauser, M. D., & Robbins, T. W. (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proc. Natl. Acad. Sci., 107*(40), 17433–17438.

Cushman, F. (2013). Action, outcome, and value a dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292.

Cushman, F., & Greene, J. D. (2012). Finding faults: How moral dilemmas illuminate cognitive structure. *Social Neuroscience, 7*(3), 269–279.

Cushman, F., Young, L., & Hauser, M. (2006). The role of conscious reasoning and intuition in moral judgment testing three principles of harm. *Psychological Science, 17*(12), 1082–1089.

Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain*. New York: Grosset/Putnam.

deGroot, A. D. (1946/1978). *Thought and choice in chess*. The Hague: Mouton.

Dickinson, A., Balleine, B., Watt, A., Gonzalez, F., & Boakes, R. A. (1995). Motivational control after extended instrumental training. *Animal Learning & Behavior, 23*(2), 197–206.

Doeller, C. F., Barry, C., & Burgess, N. (2010). Evidence for grid cells in a human memory network. *Nature, 463*(7281), 657–661.

Driver, J. (2009). The history of Utilitarianism. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. <http://plato.stanford.edu/archives/sum2009/entries/utilitarianism-history/>.

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General, 145*(6), 772.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford Review, 5*, 5–15.

Frederick, S. (2005). Cognitive reflection and decision making. *J. Econ. Perspect., 19*(4), 25–42.

Glenn, A. L., Raine, A., & Schug, R. A. (2009). The neural correlates of moral decision-making in psychopathy. *Molecular Psychiatry, 14*, 5–6.

Grabenhorst, F., & Rolls, E. T. (2011). Value, pleasure and choice in the ventral prefrontal cortex. *Trends in Cognitive Sciences, 15*(2), 56–67.

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

Greene, J. (2003). From neural 'is' to moral 'ought': What are the moral implications of neuroscientific moral psychology? *Nature Reviews Neuroscience, 4*(10), 846–850.

Greene, J. D. (2007). In W. Sinnott-Armstrong (Ed.), *Moral psychology: The neuroscience of morality: Emotion, disease, and development* (Vol. 3). Cambridge, MA: MIT Press.

Greene, J. (2013). *Moral tribes: Emotion, reason and the gap between us and them*. New York: Penguin Press.

Greene, J. D. (2014a). Beyond point-and-shoot morality: Why cognitive (neuro)science matters for ethics. *Ethics, 124*(4), 695–726.

Greene, J. D. (2014b). The cognitive neuroscience of moral judgment and decision-making. In: M. S. Gazzaniga (Ed.), *The cognitive neurosciences V*. Cambridge, MA: MIT Press.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition, 111*(3), 364–371.

Greene, J., & Haidt, J. (2002). How (and where) does moral judgment work? *Trends in Cognitive Sciences, 6*(12), 517–523.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science, 293* (5537), 2105–2108.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814.

Haidt, J. (2003). The emotional dog does learn new tricks: A reply to Pizarro and Bloom (2003). *Psychological Review, 110*(1), 197–198.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. New York: Vintage.

Haidt, J., Bjorklund, F., & Murphy, S. (2000). *Moral dumbfounding: When intuition finds no reason*. Unpublished manuscript. University of Virginia.

Hardin, G. (1968). The tragedy of the commons. *Science, 162*(3859), 1243–1248.

Hare, R. M. (1981). *Moral thinking: Its method, levels, and point*. Oxford University Press.

Hauser, M., Cushman, F., Young, L., Kang-Xing Jin, R., & Mikhail, J. (2007). A dissociation between moral judgments and justifications. *Mind & Language, 22* (1), 1–21.

Henrich, J. (2015). *The secret of our success: How culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economics: Behavioral experiments in 15 small-scale societies. *The American Economic Review, 91*(2), 73–78.

Jordan, J. J., Hoffman, M., Bloom, P., & Rand, D. G. (2016). Third-party punishment as a costly signal of trustworthiness. *Nature, 530*(7591), 473–476.

Kahan, D. M., Wittlin, M., Peters, E., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. N. (2011). The tragedy of the risk-perception commons: Culture conflict, rationality conflict, and climate change. *Temple University legal studies research paper* (2011-26).

Kahan, D. M., Peters, E., Wittlin, M., Slovic, P., Ouellette, L. L., Braman, D., & Mandel, G. (2012). The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nature Climate Change, 2*(10), 732–735.

Kahneman, D. (2003). A perspective on judgment and choice: Mapping bounded rationality. *American Psychologist, 58*(9), 697.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Macmillan.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515.

Kant, I. (1785/2002). *Groundwork for the metaphysics of morals*. New Haven, CT: Yale University Press.

Kant, I. (1930). *Lectures on ethics*. Indianapolis, IN: Hackett.

Kant, I. (1983). On a supposed right to lie because of philanthropic concerns. In J. W. Ellington (Ed.), *Ethical philosophy* (pp. 162–166). Indianapolis, IN: Hackett.

Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446*(7138), 908–911.

LeDoux, J. E. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience, 23*, 155–184.

Lieberman, D., Tooby, J., & Cosmides, L. (2007). The architecture of human kin detection. *Nature, 445*(7129), 727–731.

Mikhail, J. (2000). *Rawls' linguistic analogy: A study of the 'generative grammar' model of moral theory described by John Rawls in 'a theory of justice'*. PhD Dissertation. Cornell University.

Mikhail, J. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. Cambridge University Press.

Mill, J. S., & Bentham, J. (1863/1987). *Utilitarianism and other essays*. Harmondsworth, UK: Penguin.

Miller, R. M., Hannikainen, I. A., & Cushman, F. A. (2014). Bad actions or bad outcomes? Differentiating affective contributions to the moral condemnation of harm. *Emotion, 14*(3), 573.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... Petersen, S. (2015). Human-level control through deep reinforcement learning. *Nature, 518*(7540), 529–533.

Moser, E. I., Kropff, E., & Moser, M. B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Neuroscience, 31*(1), 69.

Nisbett, R. E., & Cohen, D. (1996). *Culture of honor: The psychology of violence in the South*. Boulder, CO: Westview Press.

Nozick, R. (1974). *Anarchy, state, and utopia*. New York: Basic Books.

Olsson, A., Nearing, K. I., & Phelps, E. A. (2007). Learning fears by observing others: The neural systems of social fear transmission. *Social Cognitive and Affective Neuroscience*.

Olsson, A., & Phelps, E. A. (2007). Social learning of fear. *Nature Neuroscience, 10*(9), 1095–1102.

Paxton, J. M., Bruni, T., & Greene, J. D. (2014). Are 'counter-intuitive' deontological judgments really counter-intuitive? An empirical reply to. *Social Cognitive and Affective Neuroscience, 9*(9), 1368–1371.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*(1), 163–177.

Phelps, E. A., Delgado, M. R., Nearing, K. I., & LeDoux, J. E. (2004). Extinction learning in humans: Role of the amygdala and vmPFC. *Neuron, 43*(6), 897–905.

Pinillos, N. Á., Smith, N., Nair, G. S., Marchetto, P., & Mun, C. (2011). Philosophy's new challenge: Experiments and intentional action. *Mind & Language, 26*(1), 115–139.

Pinker, S. (2011). *The better angels of our nature: Why violence has declined*. New York: Viking.

Pizarro, D. A., & Bloom, P. (2003). The intelligence of the moral intuitions: Comment on Haidt (2001). *Psychological Review, 110*(1), 193.

Plato (1987). *The republic*. London: Penguin Classics.

Quartz, S. R. (2009). Reason, emotion and decision-making: Risk and reward computation with feeling. *Trends in Cognitive Sciences, 13*(5), 209–215.

Railton, P. (2014). The affective dog and its rational tale: Intuition and attunement. *Ethics, 124*(4), 813–859.

Rawls, J. (1971). *A theory of justice*. Cambridge, MA: Harvard University Press.

Ronson, J. (2016). *So you've been publicly shamed*. New York: Riverhead Books.

Sandel, M. J. (2009). *Justice*. New York: Farrar, Straus and Giroux.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275*(5306), 1593–1599.

Shenhav, A., & Greene, J. D. (2010). Moral judgments recruit domain-general valuation mechanisms to integrate representations of probability and magnitude. *Neuron, 67*(4), 667–677.

Shenhav, A., & Greene, J. D. (2014). Integrative moral judgment: Dissociating the roles of the amygdala and ventromedial prefrontal cortex. *The Journal of Neuroscience, 34*(13), 4741–4749.

Sidgwick, H. (1907). *The methods of ethics*. Indianapolis, IN: Hackett.

Singer, P. (2005). Ethics and intuitions. *The Journal of Ethics, 9*(3–4), 331–352.

Singer, T., Critchley, H. D., & Preuschoff, K. (2009). A common role of insula in feelings, empathy and uncertainty. *Trends in Cognitive Sciences, 13*(8), 334–340.

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R. J., & Frith, C. D. (2004). Empathy for pain involves the affective but not sensory components of pain. *Science, 303*(5661), 1157–1162.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Mach. Learn., 3*(1), 9–44.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Thomson, J. J. (1985). The trolley problem'. *Yale Law Journal, 94*, 1395.

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2007). Reward value coding distinct from risk attitude-related uncertainty coding in human reward systems. *Journal of Neurophysiology, 97*(2), 1621–1632.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review, 55*(4), 189.

Wright, R. (1994). *The moral animal: Why we are, the way we are: The new science of evolutionary psychology*. New York: Vintage.

Please cite this article in press as: Greene, J. D. The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition* (2017), http://dx.doi.org/10.1016/j.cognition.2017.03.004