Cognition 167 (2017) 66–77

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 23 March 2017

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

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