Cognition 167 (2017) 1–10

Contents lists available at ScienceDirect

Cognition

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

Moral learning: Psychological and philosophical perspectives

Fiery Cushman [a,*], Victor Kumar [b], Peter Railton [c]

[a] Department of Psychology, Harvard University, United States
[b] Department of Philosophy, Boston University, United States
[c] Department of Philosophy, University of Michigan, United States

ARTICLE INFO

Article history:
Available online 16 June 2017

ABSTRACT

The past 15 years occasioned an extraordinary blossoming of research into the cognitive and affective mechanisms that support moral judgment and behavior. This growth in our understanding of moral mechanisms overshadowed a crucial and complementary question, however: How are they learned? As this special issue of the journal Cognition attests, a new crop of research into moral learning has now firmly taken root. This new literature draws on recent advances in formal methods developed in other domains, such as Bayesian inference, reinforcement learning and other machine learning techniques. Meanwhile, it also demonstrates how learning and deciding in a social domain—and especially in the moral domain—sometimes involves specialized cognitive systems. We review the contributions to this special issue and situate them within the broader contemporary literature. Our review focuses on how we learn moral values and moral rules, how we learn about personal moral character and relationships, and the philosophical implications of these emerging models.

1. Introduction

Between 2001 and 2005, *Cognition* doubled its rate of publication on one topic. By 2009 it doubled again. Then it doubled a third time by 2014—an eightfold increase in little over a decade (Fig. 1; Priva & Austerweil, 2015). The topic, of course, is moral psychology.

During this period of exponential growth, psychologists devoted considerable effort to understanding the cognitive and affective mechanisms responsible for moral judgment and behavior. As a result, we now have a sophisticated understanding of what people consider wrong (e.g., Alicke, 2000; Baron & Ritov, 2009; DeScioli & Kurzban, 2009; Graham et al., 2011; Gray, Young, & Waytz, 2012; Malle, Guglielmo, & Monroe, 2014; Mikhail, 2011; Pizarro, 2011), the kinds of psychological mechanisms we use to make those judgments (e.g., Cushman, Young, & Hauser, 2006; Greene, 2008; Haidt, 2001; Janoff-Bulman, Sheikh, & Hepp, 2009; Rand, Greene, & Nowak, 2012), their neural basis (e.g., Blair, Marsh, Finger, Blair, & Luo, 2006; Greene, 2004; Moll, De Oliveira Souza, & Zahn, 2008; Young, Cushman, Hauser, & Saxe, 2007; Young & Dungan, 2012), their disruption by disorder, injury or pharmacology (e.g., Crockett, Clark, Hauser, & Robbins, 2010; Koenigs, Adolphs, Cushman, & Damasio, 2007; Moran, Saxe, O'Young, & Gabrieli, 2011; Young et al., 2010), and much more.

One area of research, however, remained notably underdeveloped: Where do these mechanisms come from, in the first place?

Current theories of moral judgment tend to posit that they are a product of our innate, evolved psychology. Our capacity for moral judgment has been described as the product of an innate "universal moral grammar" (Hauser, 2006; Mikhail, 2011), as organized around a template "delineating roughly those violations that chimpanzee can appreciate" (Greene, 2004), as arising from evolved "taste buds" giving rise to distinct foundations of moral concern (Haidt & Joseph, 2004), and so on. Indeed, research documents that young children and even infants show remarkably sophisticated moral understanding (Hamlin, Wynn, & Bloom, 2007; Sloane, Baillargeon, & Premack, 2012) and behavior (Warneken & Tomasello, 2006).

None of these theories was antagonistic to the proposal that learning plays a role in moral judgment and behavior. To the contrary, each acknowledged that learning must play a crucial role. Yet, each also grants innate psychological capacities the more central position in constructing moral intuitions, and none advances a detailed account of how moral intuitions might be learned.

This is remarkable, because convergent evidence from multiple fields of academic inquiry shows that learning of some kind must play an essential role in shaping moral judgment and behavior. Anthropologists (Henrich, Heine, & Norenzayan, 2010), economists (Herrmann, Thoni, & Gachter, 2008) and social psychologists (Graham, Haidt, & Nosek, 2009; Nisbett & Cohen, 1996) have documented extensive cross-cultural variability in morality that

* Corresponding author at: Department of Psychology, Harvard University, 1484 William James Hall, 33 Kirkland St., Cambridge, MA 02138, United States.
E-mail address: cushman@fas.harvard.edu (F. Cushman).

http://dx.doi.org/10.1016/j.cognition.2017.06.008

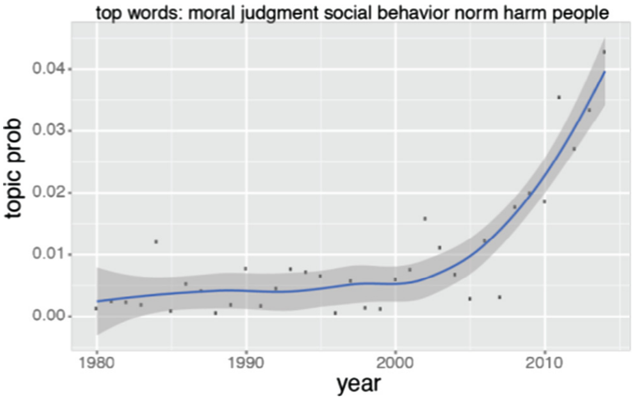

Fig. 1. Proportion of articles published in *Cognition* on the topic of moral psychology, by year. Reprinted with permission from Priva & Austerweil, 2015.

corresponds to differences in social contexts, suggesting learning. Evolutionary theorists argue that such variability was implicated in the cultural evolution of morality, as cultures that developed more effective cooperative norms gained an edge in intergroup competition (Boyd, 2005; Henrich, 2015). Laboratory experiments confirm that individuals adjust their moral behavior to the standards set by peers (Gino, Ayal, & Ariely, 2009; Goldstein, Cialdini, & Griskevicius, 2008; Peysakhovich, 2013). Learning is also crucial on a more fine-grained timescale, as people construct evaluations of social partners on the basis of their unfolding behavior (Behrens, Hunt, Woolrich, & Rushworth, 2008; Chang, Doll, van't Wout, Frank, & Sanfey, 2010; Kliemann, Young, Scholz, & Saxe, 2008; Koster-Hale, 2013; Pizarro & Tannenbaum, 2011; Zaki, Kallman, Wimmer, Ochsner, & Shohamy, 2016). And, of course, there is a long tradition of interest in moral learning in the developmental psychology tradition (reviewed in Kohlberg, 1969; Piaget, 1965/1932; Rushton, 1976; Turiel, 2005).

So there is ample evidence that learning does play a crucial role in morality; the next challenge is to understand how. What are the computations and representations that support the acquisition or formation of new moral thoughts and actions? This question animates the articles contributed to this special issue of *Cognition*. Below, we highlight these contributions and situate them within the broader contemporary literature.

The study of moral learning is timely because of recent breakthroughs in our understanding of learning. This revolves around three major areas of research—Bayesian inference, reinforcement learning, and artificial intelligence—each of which involved novel applications of computational methods and cognitive structures to solving problems of longstanding concern.

The "Bayesian" revolution in learning comprises several distinct but related elements—for instance, showing that human inference is probabilistic, that it operates over generative causal models, and that hypotheses can be arranged hierarchically (Tenenbaum, Griffiths, & Kemp, 2006; Tenenbaum, Kemp, Griffiths, & Goodman, 2011). These elements enable impressive feats of learning based even when data is limited or biased. In addition, they are well suited to learn abstract rules that generalize over diverse cases (Goodman, Ullman, & Tenenbaum). This is an appealing property for learning in the moral domain (Darley & Shultz, 1990; Kohlberg, 1969; Mikhail, 2011). Finally, Bayesian methods can enable individuals with different assumptions to converge on common conclusions (Good, 1967), which may foster cooperation among diverse individuals and groups.

The revolution in theories of value-guided learning and decision-making was prompted largely by the application of reinforcement learning (RL) methods, a family of computational models that subsequently chooses contextually appropriate actions by estimating their value—i.e., the long-term prospect of reward (Sutton, 1998). A key feature of reinforcement learning mechanisms is that they learn based on an error-driven update mechanism, a feature shared with older and influential theories of learning, such as the Rescorla and Wagner (1965) model and Thorndike's (1898) "Law of Effect". A second key feature of reinforcement learning models is their elegant encapsulation of the distinction between habitual and planned (or goal-directed) action (Dolan & Dayan, 2013). Formalizing this distinction has catalyzed a burst of new research on the psychological and neural basis of decision-making.

Finally, the last few years have seen a spectacular growth in the capabilities of artificial learning systems built on neural network models that replicate some of the features of cortical architecture, and that rely upon generic learning algorithms similar to those studied in Bayesian and RL research (Tenenbaum et al., 2011). These systems afford a "proof of possibility" of the power of general-purpose learning to learn rules and generate novel evaluative structures that promote successful behavior. This "proof" gains special relevance in light of the substantial body of neuroscientific evidence that moral decision-making implicates neural substrates widely shared among other cognitive functions (Buckner, Andrews-Hanna, & Schachter 2008; Young & Dungan, 2012; Reniers et al., 2013; Shenhav and Greene, 2014).

As this special issue reflects, many contemporary models of moral learning seek to combine these computational approaches with insights from a wide array of other traditions and literatures: The classic studies of children's moral learning that emerged in the cognitive development literature (Kohlberg, 1969; Piaget, 1965/1932), more recent studies of social and moral evaluation in infancy (Hamlin, 2013; Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013; Hamlin, Wynn, & Bloom, 2008), the embrace of social preferences that vitalized a decade of research in behavioral economics (Gintis & Boyd, 2005), the social psychological literature on norm learning (Gino et al., 2009; Goldstein et al., 2008; Peysakhovich & Rand, 2013) and the concurrent development of formal models of the cultural evolution of social norms (Boyd, Richerson, & Henrich, 2011; Henrich, 2007).

Driven by these forces, new theories of moral learning are emerging on three broad fronts. Two of these are easily anticipated: The learning of moral values (drawing especially from RL methods), and the learning of moral rules (drawing especially from Bayesian methods). A third area of development is less obvious but no less important: Learning about *people* (Uhlmann, Pizarro., & Diermeier, 2015). This comprises several interrelated challenges: Figuring out who you should care about or trust, what attitudes or motives others have toward you or toward one another, what to expect from someone and what others will expect of you, and how these networks of interpersonal valuation influence and react to social group boundaries. As we review below, each of these areas has seen recent activity, and all three are well represented in this special issue.

One of the most exciting consequences of a theory of moral learning is that it naturally suggests mechanisms both for innovation in moral thought and for practical ways of bringing about moral changes. Several of the chapters take up the practical question of asking how moral change might be promoted (Graham, Waytz, Meindl, Iyer, & Young, 2017; McAuliffe et al., 2017; Stagnaro, Arechar, & Rand 2017; Walker & Lombrozo, 2017), and also discover its potential limits (Graham et al., 2017; McAuliffe et al., 2017; Paluck, Shafir, & Wu, 2017). Finally, several contributions to this issue explore the philosophical implications of recent research into moral learning (Railton, 2017; Campbell, 2017; Kumar, 2017; Greene, 2017).