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
0010-0277/

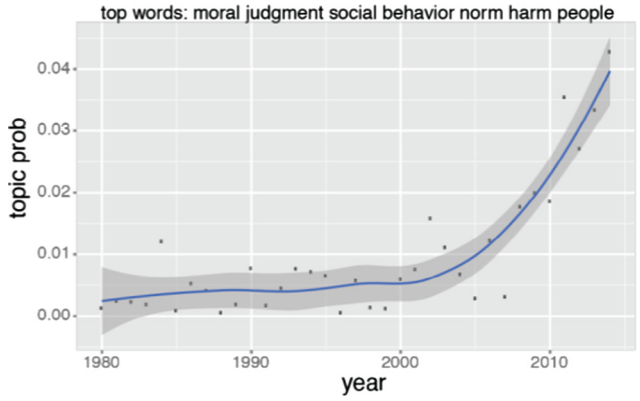

Fig. 1. Proportion of articles published in *Cognition* on the topic of moral psychology, by year. Reprinted with permission from Priva & Austerweil, 2015.

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).

2. Learning value

Value representations are a pervasive feature of morality. We assign moral value to people, to actions, to states of affairs, to rules, and so forth. Thus, a fundamental challenge for any theory of moral learning is to explain how representations of moral value are learned.

An obvious starting point is to analogize from the learning of values that are not specifically moral (Morris & Cushman, in press; Ruff & Fehr, 2014; Seymour, Singer, & Dolan, 2007). As noted above, this has been an area of intense research and notable success over the past 20 years, spanning both psychology and neuroscience (Dolan & Dayan, 2013; Glimcher, 2011; Platt & Glimcher, 1999). Prompted by this success, many studies have asked whether decision-making guided by social or moral value is represented in a similar manner, and by similar neural mechanisms, as decision-making guided by non-social preferences. The predominant answer has been "yes" (Ruff & Fehr, 2014; Shenhav & Greene, 2010; Young & Dungan, 2012; Zaki & Mitchell, 2011).

Much less prior research, however, asks whether moral value is acquired through the same mechanisms. As documented by several of the contributions to this issue, this is a matter of uncertainty and perhaps even controversy.

2.1. Instrumental value: Moral habits and heuristics

At one extreme, perhaps moral values are acquired largely through the same reward-maximization mechanisms widely implicated in non-moral and especially non-social decision-making. The most extreme version of this thesis claims that morality is, in fact, just the ordinary assignment of value to the things we ordinarily find rewarding (i.e., food, companionship, physical security, etc.). Few models of moral learning commit to such an extreme position, although several may be compatible with it (Crockett, 2013; Cushman, 2013; Rand, Kraft-Todd, Wurzbacher, & Greene, 2014).

This basic idea animates the "Social Heuristics Hypothesis" (Rand et al., 2014), which attempts to explain how self-interested motives might ultimately produce apparently altruistic acts. The essence of the model is that we develop heuristic responses to social dilemmas based on the predominate form of the dilemmas and behaviors of our social partners. Thus, for instance, if we grow up in a culture where opportunities for cooperation typically occur in repeated interactions with partners who are conditionally cooperative, then cooperation is in our own self-interest. As a result, we develop the heuristic response to cooperate. Crucially, this heuristic operates even in contexts where cooperation is no longer favored by self-interest, such as one-shot laboratory settings. Consistent with this theory, cooperation in one-shot games is increased when time pressure or other manipulations favor heuristic responding (Rand et al., 2012), but this effect is diminished when people have extensive experience with such games and thus would favor defection by default (Rand et al., 2014).

An obvious implication of these findings is that the quality of a culture's social institutions—i.e., the degree to which they align social interests with individual interests—will predict the degree of cooperativeness of the individuals in that culture. Stagnaro et al. (2017) provide striking data in favor of this hypothesis. First, they show that quality of civic institutions experienced by ordinary people influences their willingness to give away money in a one-shot, anonymous laboratory interaction (i.e., to act "altruistically"). Next, they show that they can manipulate this effect by exposing people to relatively higher or lower quality institutional arrangements in a laboratory public goods game.

2.2. The intrinsic value of moral rewards

A theory of reward learning requires a specification of what counts as a "reward". Although it is possible that a single, common set of rewards underwrites both moral and non-moral learning, many theories posit distinctive sources of reward that contribute especially to moral learning. What distinctive forms of "reward" might guide the learning of moral values? Past discussions have repeatedly revolved around two candidates: empathy and conformity. In other words, we take intrinsic pleasure in (1) seeing other people do well, and (2) acting the way they do.

2.2.1. Empathy

Blair (2017) suggests that a deficit precisely with respect to the ability to affectively simulate negative experiences may play a role in morally problematic behaviors associated with psychopathic disorders, along with a deficit in linking such simulations to novel stimuli and actions. In other words, he conceptualizes psychopathy as a kind of "learning disability"—aversive information about outcomes to others is not being learned from past experience. Consider, for example, actions that cause suffering to others (hitting, biting, humiliating, etc.). Normally children acquire a learned aversion to these actions based on their learned association with suffering. But if there is a failure to register that suffering in others, then the emotional learning will not take place, and such actions will not be directly aversive. In its initial form this hypothesis focused on the role of stimulus-reinforcement learning (Blair, 1995; Blair, Jones, Clark, & Smith, 1997); in addition to reviewing current support for this model, Blair extends it here to encompass response-outcome learning.

The claim that empathy functions as a kind of distinctive "moral reward" dovetails with several distinct literatures. There is much evidence that empathy is correlated with prosocial behavior (Decety, 2015; Eisenberg, Losoya, & Spinrad, 2003; Marsh, 2016; Vaish, Carpenter, & Tomasello, 2009). A distinct literature, focused specifically on children's moral development, investigates the relative value of empathy and punishment in establishing prosocial behavior (Hoffman, 2000). It is argued that punishment is an inferior technique because the child tends to externalize the motive for care (Gershoff et al., 2010), whereas empathy tends to lead to internalization of this value.

2.2.2. Conformity

Empathy can register the goodness or badness of actions and outcomes, but many bad things are not moral transgressions (for instance, tornados). How might we acquire a specific awareness of, and concern with, moral violations? A highly influential tradition of research in social psychology posits that moral behavior learned largely by exposure to "norms". This literature typically contrasts the influence of descriptive norms (what most people do) with that of injunctive norms (what people say you should do). A consistent finding of this research is that humans do as others do, not as they say—in other words, providing descriptive information about norm compliance is more effective in shaping behavior than directly stipulating the norm in moral terms (Goldstein et al., 2008; Rushton, 1976; Schultz, Nolan, Cialdini, Goldstein, & Griskevicius, 2007). Bear and Knobe (2017) suggest that the especially powerful role of descriptive norms may arise in part because of an intuitive, undifferentied concept of "normality" that encompasses both normative and descriptive content (see Section 5.1.3 for further discussion). Consistent with this, some research indicates that learned expectations about typical resource distributions play a key role in guiding attitudes about fair distributions, and ultimately punishment for allocations perceived to be unfair (Chang & Sanfey, 2013).

McAuliffe et al. (2017) extend this program of research to norm learning in young children. They offer young children the opportunity to give money to social partners in a dictator game under the influence of either stingy or generous norms, which are either descriptive in nature (e.g., "most children give 80%") or injunctive in nature (e.g., "you ought to give 80%"). Their participants are sensitive to both norms, and to a roughly equally degree. This comports with some prior research showing modest but significant effects of both descriptive norms (often in the form of parents modeling behavior, reactions, or both) and also injunctive norms on children's moral judgments (Rottman & Kelemen, 2012; Rottman, Young, & Kelemen, 2017; Rushton, 1976).

2.3. Social learning: Inference & internalization

A longstanding theme in the literature on social and moral development is "internalization": The tendency of people—and especially children—to adopt novel preferences from their cultural milieu. The premise of internalization is that people experience actual change in the objects to which they assign primary reward: Avocados come to taste better (when you see your parents eat them), faces look more attractive (when you've watched others swoon over them), and stinginess seems more reprehensible (when you've seen it punished).

Borrowing key concepts from the reinforcement learning literature, Ho et al. (2017) offer a formal account of what it means to internalize moral values, and then analyze why we should ever do it. In ordinary non-social settings, reinforcement learning consists in taking an innate reward function and estimating the value of actions that maximize over it. Put simply: What you like never changes; what you learn is how to get it. Ho et al. explain, however, that in a social context it can be rational to learn not only how to get the stuff you like, but also to like (or dislike) new things—formally, to learn a new reward function. The consequence is a fundamental reorganization of the standard reinforcement learning framework specifically for social settings.

Magid and Schulz (2017) offer a different and intriguing method for the construction of novel moral rewards that depends on deep interpersonal attachments such as love and friendship. The key idea is that one person's non-moral preferences can become another person's moral preferences when the second person has a sufficiently strong social bond to the first. Consider a specific case: A father discovers that his toddler daughter desperately wants a fire truck for Christmas. The daughter's preference for a fire truck is not moral, yet the father may feel a moral obligation to provide her with a truck.

2.4. Self-taught values: Imagination & dissonance

It is a remarkable feature of human psychology that we can learn not only from the world, and from each other, but also from ourselves. Through processes of imagination and reasoning we can come to appreciate new facts (Lombrozo, in press), adopt new perspectives (Epley, 2014; Tamir & Mitchell, 2010), and assign new values (Barron, Dolan, & Behrens, 2013; Gershman, Markman, & Otto, 2014).

Some prior research shows the important ways in which vivid imaginative processes can shape moral values. For instance, when people vividly imagine events in which a person benefits from generosity, their own subsequent prosocial intentions increase (Gaesser & Schacter, 2014). Conversely, when people vividly imagine harm, this increases their moral condemnation of it (Amit & Greene, 2012; Caruso, 2010; Caruso & Gino, 2011). Why does imagination matter? One intriguing possibility is that certain processes of value assignment are ordinarily changed by direct experience but not by abstract conceptual knowledge (Paul, 2014); the role of imagination, in this case, is to transform abstract knowledge into "virtual" experiences sufficient to update value (Barron et al., 2013; Gershman et al., 2014).

Theories such as cognitive dissonance (Festinger, 1962) and balance theory (Heider, 1946) emphasize that individuals attempt to achieve consistence among their attitudes, beliefs and behaviors. Paluck et al. (2017) devised an experiment in which apparent learning about the self in the moral domain could lead to such consistency effects. Subjects who were attempting to carry out a task assigned to them by the experimenter were exposed to a television reporting on famine, positioned as an incidental feature of their situation. Thus subjects "observed" themselves ignoring information about suffering, and Paluck and colleagues found that, when tested after this experience, subjects manifested reduced motivation to alleviate suffering.

3. Learning rules

Morality consists not only of values, but also of rules. Many familiar moral rules are explicit, or even institutionally codified (e.g., criminal laws and religious commandments). Certain consistent patterns of implicit, automatic influence on moral judgment may also be described as a form of "moral rule", although it is less clear whether they are represented in a propositional form. (An example is the much studied "doctrine of double effect".) Whether explicit or implicit, structured patterns of moral judgment must come from somewhere. For any candidate moral rule we may ask, "Was it learned?" and, if so, "How?".

Rhodes and Wellman (2017) build upon a now classic perspective in cognitive development, according to which we understand particular things in light of broader framework theories—elephants, for instance, in light of a theory of biology (Carey, 1985), cannons in light of a theory of physics (Caramazza, McCloskey, & Green, 1981; McCloskey, Caramazza, & Green, 1981) and hide-and-seek in light of a theory of mind (Baron-Cohen, Leslie, & Frith, 1985; Wimmer & Perner, 1983). Applying these principles to the moral domain, they identify two framework theories that are especially fundamental to morality: theory of mind (i.e., a causal model linking unobservable mental states to overt action) and a naive theory of sociology (i.e., a theory of the relationships among individuals, especially in light of their group identities).

This perspective emphasizes the role that non-moral learning may play in shaping moral behavior and guiding moral knowledge. But, it also serves as an important reminder of the scope and complexity of the learning problem that we face when attempting to learn moral rules. Explicit instruction of concrete principles (e.g., "Do not eat pork!") occasionally arise, but humans also exhibit rule-like structure in the moral domain with little explicit instruction (Wright, 2008).

3.1. Social learning by Bayesian inference

Contemporary methods of Bayesian inference offer a powerful model of how complex structure can be accurately learned from sparse data (Tenenbaum et al., 2011). Indeed, this approach has sought to explain how both particular knowledge and abstract theories can be learned in domains including biology (Tenenbaum et al., 2006), physics (Gerstenberg et al., 2012; Hamrick & Tenenbaum, 2011) and theory of mind (Baker, Saxe, & Tenenbaum, 2009; Hamlin et al., 2013). It is no surprise, then, that there is growing interest in applying Bayesian methods to understand inference and learning in the moral domain (Cushman, 2009; Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum; Nichols, Kumar, Lopez, Ayars, & Chan, 2016).

Ayars and Nichols (2017) applies insights from the Bayesian perspective to explore how people might generalize from properties of known moral rules to properties of novel moral rules. For instance, they demonstrate that when people have been exposed to three moral rules that prohibit both accidental and intentional acts they assume that a fourth rule will also prohibit both types, and significantly more often than people exposed to three rules that merely prohibit intentional acts. This method of inference depends upon the concept of an "overhypothesis", and can be modeled in a hierarchical Bayesian setting (Griffiths, Kemp, & Tenenbaum, 2008).

Similarly, Kleiman-Weiner, Saxe, and Tenenbaum (2017) deploy Bayesian methods to model how people might infer rule-like structure in others' social preferences, ultimately in service of adopting the same social preferences themselves. Their model assumes several candidate rules for social preferences, such as kin favoritism, in-group favoritism, direct reciprocity and indirect reciprocity. For an observer, then, these preference structures constitute a hypothesis space that could explain the social behavior of their peers. Even under uncertainty about social relationships (e.g., uncertainty about kinship, group membership or prior cooperative behavior), it is possible for agents to perform joint inference over both the relationships and the moral rules that refer to them.

3.2. The influence of descriptive norms on rule inference

One of the principal reasons to frame morality in terms of *rules*, and not just in terms of *value*, is that rules directly capture the notion of an exclusionary constraint on behavior–some things are not just bad but forbidden (Cushman, 2015). In contrast, a key premise of value-guided decision-making is that diverse goods trade off flexibly against each other in a common currency format (Becker, 1996)—a scheme that forecloses the possibility of exclusionary constraints on choice. Insofar as there is ample evidence that human moral judgment and behavior involves elements of both value representation and rule representation (Nichols, 2002), a key question is how people decide when to treat a moral issue as relatively more value-like (e.g., "I'm willing to eat meat, but I try not to do it too much") versus rule-like (e.g., "It is wrong to eat meat").

Heiphetz and Young (2017) take an important step towards answering this question. Their point of departure is an emerging literature on adult moral judgment that demonstrates a simple point: When we observe that everybody around us agrees on a moral issue, we tend to treat it more like a "moral fact" or inviolable rule. In contrast, when we observe pervasive disagreement on a moral issue, we tend to treat it as a matter of "moral opinion", and thus in some sense more like value (Theriault, Waytz, Heiphetz, & Young, 2017; Theriault, Waytz, Heiphetz, Young, & Theriault, 2017). Heiphetz and colleagues demonstrate the same pattern of judgement in young children. Collectively, these studies add an important nuance to the literature on the role of descriptive norms in the moral domain. Prior research tends to assume that there is "one kind" of norm representation, and that the more people are perceived to follow a norm, the stronger its grip on the mind of the perceiver. In contrast to this assumption, Heiphetz and colleagues demonstrate that there are diverse types of norm representation favored under conditions of divergent versus convergent descriptive norms.

3.3. Self-taught rules: Reasoning and reflection

People also internally update rules by engaging in principled reasoning and reflection. This is a remarkably understudied topic in contemporary moral psychology. Walker and Lombrozo (2017) provide a vivid demonstration of the crucial role that explicit, reflective reasoning plays in extracting rule-like representations in order to foster moral learning. Consistent with some prior research (Lee, Talwar, Ross, Evans, & Arruda, 2014; Paluck, 2009), they find that stories are an important vector for moral learning. But their principal aim is to investigate how a child's interaction with the content of a story supports this transformation. Their experiments indicate that when children are asked to identify and explain the moral of a story, this increases their ability to apply that moral to subsequent stories. They find evidence that self-generated explanations can be even more powerful than direct instruction, in this case from the experimenter.

3.4. The relationship between values and rules

Values and rules are different kinds of representations, but they both clearly contribute to moral cognition. Thus it is essential to understand how they interact. There are at least two possible forms that this interaction might take that are not mutually exclusive. One form of interaction is for the rule to specify an assignment of value. For instance, a rule could be, "Every life must be valued equally"; or, "You must value your family members above strangers". This form of interaction could be thought of as a "value-internal rule", insofar as the content of the rule is stated directly in terms of a representation of value (or relation between values). Alternatively, another form of interaction is for the rule itself to be valued. For instance, a rule could be, "Waiters should be tipped 20% for good service." This form of interaction could be thought of as a "value-external rule", insofar as the content of the rule is not stated in terms of a representation of value. These forms can be combined: Imagine, for instance, a person who says, "I think that it is important to value animals as much as humans [value-internal], but not important enough to stop eating meat [value-external]!".

4. Learning about people

Our moral psychology is not designed exclusively to regulate our own behavior; It also allows us to make judgments of others (Uhlmann et al., 2015). The capacity to evaluate third party behavior has several crucial functions: To predict how they will act in the future, to decide whether to forge, maintain or dissolve social relationships with them, and to decide whether they ought to be punished or rewarded for their conduct. These functions require their own form of learning, not about rules or values, but about social partners. This includes their personal character, and also their affiliation with other individuals and with social groups.

4.1. Learning personal character

Much research in social psychology emphasizes the tendency of humans to form spontaneous impressions of others (Heider, 1958; Winter and Uleman; 1984, Fiske, Cuddy, Glick, & Xu, 2002). A key new frontier for theories of character inference is to establish how we integrate potentially inconsistent information across time in service of a unified and accurate model of personal character (Behrens et al., 2008; Chang et al., 2010; Kliemann et al., 2008; Koster-Hale & Saxe, 2013; Pizarro & Tannenbaum, 2011; Zaki et al., 2016).

Steckler, Woo, and Hamlin (2017) begin by evaluating whether this capacity is present during infancy, as the capacity for social evaluation is just emerging. Prior research indicates that by 5 months infants have the capacity to distinguish between unambiguously "helpful" and "harmful" actors (Hamlin et al., 2007), and this capacity becomes increasingly sophisticated over the following year (Hamlin, 2013; Hamlin, Wynn, Bloom, & Mahajan, 2011; Hamlin et al., 2013). Building on these methods, Steckler and

Hamlin asked how infants would compare an unambiguously helpful or harmful agent with one who exhibits a mixture of behaviors: Some helpful, others harmful. They find that infants cannot reliably distinguish unambiguous from ambiguous actors.

Siegel, Crockett, and Dolan (2017) explore a related set of issues in the character inferences of adults. Their research grafts contemporary methods of modeling trial-by-trial incremental learning onto classic approaches to person perception and moral judgment. When the history of prior choices indicates that a person has a bad moral character, subsequent judgments of that individual are highly sensitive to information about their personal incentives—in other words, they will be judged especially harshly for apparently self-interested choices. In contrast, when the history indicates a good moral character, subsequent judgments are less sensitive to information about personal incentives. In essence, this is a "rich get richer, poor get poorer" scheme: Immoral actors are kept on a tight leash, and observers are attuned to any suggestion of wrongdoing, while moral actors are granted the benefit of the doubt.

4.2. Learning group boundaries

As Rhodes and Wellman (2017) emphasize, moral cognition requires not just a theory of others' minds, but also a theory of the relations, alliances and conflicts between individuals—what they describe as a 'folk sociology'. They review the crucial role that group membership plays in structuring our moral judgments. But how do we determine group membership in the first place? Two contributions to this special issue share a common hypothesis regarding the cognitive genesis of group psychology: The scope of our concern for the wellbeing of others (Graham et al., 2017; Kleiman-Weiner et al., 2017).

Graham et al. (2017) systematize the set of psychological forces that expand and contract the set of individuals whose wellbeing we show concern for. They adopt the metaphor of a moral circle (Singer, 1981), which does double duty. First it captures the simple idea that there is a set of individuals included in the scope of concern and others excluded. Second, it captures the hypothesis that a primary organizing dimension of inclusion is the social "closeness" of the entity in question—i.e., family lies close to the circle, followed perhaps by friends, affiliates, compatriots, humans, animals, etc. This metaphor has been used extensively to describe conflict *between* individuals—e.g., contemporary political liberals tend to favor a broader moral circle than contemporary political conservatives (Waytz, Iyer, Young, Haidt, & Graham, in prep). Yet, Graham et al. point out that this reflects an internal conflict *within* individuals, driven by psychological forces that promote an expansion of the moral circle and those that promote its restriction.

On this analysis there is a dual relationship between welfare concern and group identity. On the one hand, group membership determine moral concern: group affiliations can help to define the dimension of social closeness that determines whether one individual fits within or beyond the scope of another's moral concern. On the other hand, moral concern defines group membership: Whatever forces move us to value or devalue the lives of others will influence how we conceive of our "community".

Kleiman-Weiner et al. (2017) take an important step towards formalizing this latter process. The crux of the problem is to render a representation of the *value of individuals* into a representation of the *identity of the group*. In their model, this is accomplished through a logic so simple that it is summarized by lyrics of a children's song: "Your friends are my friends, and my friends are your friends". This property falls out of the recursive nature of interpersonal utility: Peter values Paul, and Paul values Mary, then Peter must also value Mary. By adding to this property the constraint that people assign greatest value to individuals who share their values, Kleiman-Weiner et al. (2017) demonstrate a dynamic in which people naturally arrange themselves into groups of individuals with shared values and mutual interpersonal concerns.

5. Practical and philosophical implications

Implicit in the study of moral learning is the promise of change. First, and most obviously, by understanding how people change their moral values we are positioned to facilitate that change. These lessons may be applied at large scales, such as policymaking; they may be applied at intermediate scales, such as moral education in homes and schools; and they may also be applied at intimate scales, such as when an individual embarks on program of self-improvement. Each of these is a potential practical consequence of research into moral learning.

The goal of changing morals is, of course, to improve them. But this raises a vexing question: what constitutes "moral improvement"? This is not an easy question to answer, but several of the contributions to this volume argue that it becomes easier to answer when we have a clear understanding of moral learning. After addressing the practical consequences of moral learning ("How *can* we change morals?"), we finally turn to consider the philosophical consequences ("How *should* we change them?").

5.1. Practical implications

Following the contours of this issue's contents, we consider several broad answers to the question, "how can we change moral values"? These are neither mutually exclusive nor (we presume) complete.

5.1.1. Manipulate incentives

One answer is that we can manipulate moral values by the application of incentives that are not intrinsically moral themselves—i.e., by "material" punishments and rewards. There are several reasons this approach might be effective. According to the social heuristics hypothesis, this is possible because moral values often reflect a habit-like assignment of intrinsic value to behaviors that maximize "material" self-interest. Alternatively, on the analysis offered by Ho and colleagues in this issue, humans are designed to internalize the inferred communicative intent that underlies acts of reward and punishment. They propose that this form of internalization constitutes an important mechanism of moral learning, due principally not to the incentive value of the rewards and punishments, but rather to the social acts themselves.

5.1.2. Co-opt empathy

Another possibility is that we can change moral values by reorienting peoples' focus or experience in a way that draws upon existing incentives intrinsic to the moral domain, such as empathy. Blair (2017) proposes that an innate empathic response to victim distress is the engine driving much moral thought and behavior. If so, then an important avenue for fostering moral change is to illustrate to people the way that their behaviors will tend to cause or alleviate distress among others. This has been an influential idea in the field of moral development and education (Hoffman, 2000), although hardly a universally accepted one, as Graham et al. (2017) review.

Railton (2017) argues that empathy can play a role in "unlearning" acquired social prejudices. The quality of one's learning depends upon the quality of one's evidential sample, and social exclusion and stigmatization often mean that our personal experiences with other groups are limited. If institutions can promote more genuinely inclusive personal experience—that is, make available a more representative sample, especially one that involves

shared activities with joint goals—empathy-based learning can undermine implicit and explicit bias (Dasgupta & Rivera 2008; Pettigrew & Tropp 2006). The recent history of dramatically increased acceptance of equal rights for gay couples provides another example: Acceptance rose in proportion to the size of the population became aware of the sexual orientation of valued members of the communities or families (Westgate, Riskind, & Nosek, 2015).

5.1.3. Shape descriptive norms

A third way to change moral values is to alter people's perceptions of others' behavior—in other words, to manipulate descriptive norms. The power of descriptive norms to influence adults' behavior has a long history of research (Rushton, 1976; Schultz et al., 2007), and McAuliffe et al. (2017) show that it can also be effective in changing children's behavior. Heiphetz and Young (2017) additionally show that children draw rich normative inferences based on the descriptive distribution of norm endorsement. Specifically, they tend to represent controversial moral norms as more preference-like, while representing uncontroversial moral norms as more fact-like.

Bear and Knobe (2017) propose that the relationship between descriptive norms (i.e., a representation of how people tend to behave) and prescriptive norms (i.e., the representation of a moral rule or value) are more intimately connected that has been previously assumed. Specifically, they show that people apply a concept of "normality" that encompasses both prescriptive and descriptive norms. For instance, when asked what a "normal" weight is, people tend to choose a number intermediate between their perception of the average weight (a descriptive norm) with their perception of the ideal weight (a prescriptive norm). Thus, they construe normality as an "undifferentiated" concept in which both statistics and ideals are integrated. Emerging evidence suggests a parallel "undi fferentiated" concept of possibility that applies in the domain of modal cognition (Phillips, 2017). A key area for future research is to see whether this undifferentiated concept can help to explain the causal influence of descriptive norms on moral beliefs and behaviors.

5.1.4. Promote reasoning and reflection

Several contributions to this volume emphasize the potential role for reasoning processes in fostering moral change. The potential for reasoning processes to effect meaningful moral change has been subject to much dispute (Greene, 2008; Haidt, 2001; Paxton & Greene, 2011; Pinker, 2011; Pizarro, 2003; Schwitzgebel & Cushman, 2015). There are several reasons to doubt whether reasoning often contributes to change in moral values or moral behavior: People often exhibit consistent patterns of moral judgment without awareness (Cushman et al., 2006; Hauser, Young, & Mikhail, 2007), their reasoning is subject to predictable self-interested biases (Ditto, 2011; Nickerson, 1998; Uhlmann, Pizarro, Tannenbaum, & Ditto, 2009), and even moral philosophers (who are putatively experts in moral reasoning; Crosthwaite, 1995; Føllesdal, 2004; Grundmann, 2010; Haidt, 2001) tend to think and act similarly to others (Rust, in preparation; Schwitzgebel, 2009; Schwitzgebel, in preparation; Schwitzgebel & Cushman, 2012; Schwitzgebel & Cushman, 2015; Schwitzgebel & Rust, 2010; Schwitzgebel, Rust, Moore, Huang, & Coates, in preparation).

On the other hand, several other bodies of evidence suggest a potential role for reasoning and reflection in altering moral judgment and behavior. There is some evidence that intelligence positively correlated with prosocial behavior (Jones, 2008; Proto, Rustichini, & Sofianos, 2014; but see Rand et al., 2012). Several lines of evidence suggest a link between controlled cognitive processes and utilitarian moral judgment (Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008; Greene, Nystrom, Engell, Darley, & Cohen, 2004; Moore, Clark, & Kane, 2008). And, qualitative analysis of major historical changes in moral attitudes and behaviors is consistent with a role for argument, reflection and reasoning (Pinker, 2011).

Several contributions to this special issue add further evidence in support of the power of reflection and reasoning to prompt moral change for better (Walker & Lombrozo, 2017) or for worse (Paluck et al., 2017), and also suggest potential mechanisms for this influence (Campbell, 2017). Campbell (2017) argues that reasoning spurs moral change but not primarily through the application of explicit moral principles (like the principle of utility). Instead, reasoning is employed to identify conflicts between judgments about similar cases and then revise or modify judgments to restore consistency (see also Campbell & Kumar, 2012). "Treating like cases alike" can ameliorate bias, set priorities between conflicting norms, and help to extinguish pernicious attitudes. As Campbell suggests, learning from moral inconsistency may have played a role in progressive social movements, including the recent revolution in attitudes toward homosexuality.

5.2. Interactions between normative moral philosophy and empirical moral psychology

An exciting aspect of the moral learning approach is that attention to learning mechanisms opens up the possibility of new explanations of well-known phenomena that have proven persistently puzzling in moral philosophy and psychology. Greene (2017) and Railton (2017) present contrasting accounts of intuitive moral judgment in certain moral dilemmas—including the well-known "trolley problems"—each deploying a distinctive learning mechanism.

Greene (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Greene & Haidt, 2002) had earlier explored a "dual-process" account of apparent conflicts among people's judgments about whether it is permissible to harm one save multiple others, arguing that, in cases of "personal" harm, a fast, emotional, domain-specific "System 1" response will, in cases of more "personal" harm, predominate over more rational, domain-general "System 2" calculation of costs and benefits. In Greene (2017), he suggests—following Cushman (2013) and Crockett (2013), and drawing upon his own earlier work (Shenhav & Greene, 2014)—that the dual-process account of conflicts in moral judgment might best be formulated in terms of two different, domain-general forms of learning and decision-making—model-free vs. model-based.

Railton (2017) agrees about the domain-general character of the processes underlying intuitive moral judgment, but presents evidence that variation in when harming one to save many is viewed as permissible may be attributable model-based learning, involving a simulation and assessment (cf. Buckner et al., 2008) of the *trustworthiness* of someone who would perform the action in question, a character attribution for which we have independent empirical evidence (Conway & Gawronski, 2013; Everett, Pizarro, & Crockett, 2016; Gleichgerrcht & Young, 2013; Kahane, Evertt, Earp, Farias, & Savulescu, 2015). This would speak in favor of the credibility of these intuitions, and also suggest that patterns of judgment in trolley-like scenarios may have a greater connection to moral views like virtue theory (Annas, 2004) or motive-utilitarianism (Adams, 1976; Railton, 1988) than to the familiar opposition between act-utilitarianism and deontology.

The difference between Greene's and Railton's account is of special interest to assessing the credibility of moral intuitions, since it concerns whether a given intuitive judgment is, or is not, responsive to the morally-relevant values at stake in a given scenario. Given our increased understanding of the neural basis of model-free and model-based learning and their role in choice (Daw, Niv,

& Dayan, 2005; Dayan & Berridge, 2014; Glimcher, 2011; Redish, 2016), we are reaching a point where we can begin to test for evidence of which kind of learning—if either—might be at work in intuitive judgments in such moral dilemmas.

Moral learning appears to involve both rational and emotional elements, and thus promises to help overcome the dualism of older philosophical and psychological approaches to morality. This idea is further developed in Kumar (2017) and Campbell (2017). A number of philosophers and psychologists have sought to "debunk" certain moral perspectives by tracing their etiology to morally-irrelevant considerations. Kumar argues, however, that if rational learning processes—like Bayesian inference—are involved in the acquisition and evolution of moral attitudes, then an empirical explanation of their origins may *vindicate* rather than debunk them. Kumar suggests that model-based learning vindicates attitude changes related to moral luck, honor, and disgust, which have sometimes been dismissed as mere rationalizations—philosophical accounts of moral attitudes may be fertile ground for empirical hypotheses about how the changes occurred. And Campbell (2017) investigates the ways in which *consistency reasoning* grounded in the assessment of actual or potential cases plays an important role in moral life, and can yield a distinctive kind of pressure for rational revision of moral views—a form of reflective, higher-order learning that combines the roles of experience, affect, and reason.

6. Conclusion

Philosophers and psychologists have pondered the origins of moral thought and action since the inception of their fields. Never, however, has the topic engendered such a palpable shared sense of excitement and progress. The current renaissance of moral learning promises to deepen the mutual engagement of these disciplines yet further—it is not tethered to one theory of morality, one theory of learning, or one set of phenomena. Rather, as this special issue attests, it encompasses diverse traditions of theory, numerous methods, and explananda ranging from values to rules to persons. In other words, the unifying theme of this contemporary research is not any one answer, but the common recognition of a pivotal question: How do we learn right from wrong?

Funding

F.C. is supported by grant N00014-14-1-0800 from the Office of Naval Research and by the Transformative Experience project funded by the Sir John Templeton Foundation.

References

Adams, R. (1976). Motive utilitarianism. *Journal of Philosophy, 73*, 467–481.
Alicke (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*(4), 556–574.
Amit & Greene (2012). You see, the ends don't justify the means visual imagery and moral judgment. *Psychological Science, 23*(8), 861–868.
Annas, J. (2004). Being virtuous and doing the right thing. *Proceedings of the American Philosophical Association, 78*, 61–75.
Ayars, A., & Nichols, S. (2017). Moral empiricism and the bias for act-based rules. *Cognition, 167*, 11–24.
Baker Saxe & Tenenbaum (2009). Action understanding as inverse planning. *Cognition, 113*(3), 329–349.
Baron & Ritov (2009). Protected values and omission bias as deontological judgments. In D. M. Bartels, C. W. Bauman, L. J. Skitka, & D. Medin (Eds.). *Moral judgment and decision making* (Vol. 50). San Diego, CA: Academic Press.
Baron-Cohen Leslie & Frith (1985). Does the autistic child have a "theory of mind"? *Cognition, 21*, 37–46.
Barron Dolan & Behrens (2013). Online evaluation of novel choices by simultaneous representation of multiple memories. *Nature Neuroscience, 16*(10), 1492–1498.
Bear, A., & Knobe, J. (2017). Normality: Part descriptive, part prescriptive. *Cognition, 167*, 25–37.
Becker (1996). *Accounting for tastes*. Harvard University Press.
Behrens Hunt Woolrich & Rushworth (2008). Associative learning of social value. *Nature Materials, 456*(7219), 245–249. http://dx.doi.org/10.1038/nature07538.
Blair (1995). A cognitive developmental approach to morality: Investigating the psychopath. *Cognition, 57*, 1–29.
Blair, J. (2017). Emotion-based learning systems and the development of morality. *Cognition, 167*, 38–45.
Blair, J. R., Jones Clark & Smith (1997). The psychopathic individual: A lack of responsiveness to distress cues? *Psychophysiology, 34*(2), 192–198.
Blair Marsh Finger Blair & Luo (2006). Neuro-cognitive systems involved in morality. *Philosophical Explorations, 9*(1), 13–27.
Boyd, & Richerson (2005). *The origin and evolution of cultures*. Oxford University Press.
Boyd, Richerson, & Henrich (2011). The cultural niche: Why social learning is essential for human adaptation. In Proceedings of the National Academy of Sciences of the United States of America, 108(Suppl. 2), pp. 10918–10925. http://dx.doi.org/10.1073/pnas.1100290108.
Buckner Andrews-Hanna & Schachter (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the New York Academy of Sciences, 1124*, 1–38.
Campbell, R. (2017). Learning from moral inconsistency. *Cognition, 167*, 46–57.
Campbell & Kumar (2012). Moral reasoning on the ground. *Ethics, 122*, 273–312.
Caramazza McCloskey & Green (1981). Naive beliefs in "sophisticated" subjects: Misconceptions about the trajectories of objects. *Cognition, 9*, 117–123.
Carey (1985). *Conceptual change in childhood*. Cambridge: MIT Press.
Caruso (2010). When the future feels worse than the past: A temporal inconsistency in moral judgment. *Journal of Experimental Psychology: General*. http://dx.doi.org/10.1037/a0020757.
Caruso & Gino (2011). Blind ethics: Closing one's eyes polarizes moral judgments and discourages dishonest behavior. *Cognition, 118*(2), 280–285.
Chang Doll van't Wout Frank & Sanfey (2010). Seeing is believing: Trustworthiness as a dynamic belief. *Cognitive Psychology, 61*(2), 87–105.
Chang & Sanfey (2013). Great expectations: Neural computations underlying the use of social norms in decision-making. *Social Cognitive and Affective Neuroscience, 8*(3), 277–284.
Conway & Gawronski (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology, 104*, 216–235.
Crockett (2013). Models of morality. *Trends in Cognitive Sciences*.
Crockett Clark Hauser & Robbins (2010). Serotonin selectively influences moral judgment and behavior through effects on harm aversion. *Proceedings of the National Academy of Sciences, 107*(40), 17433.
Crosthwaite (1995). Moral expertise: A problem in the professional ethics of professional ethicists. *Bioethics, 9*(4), 361–379.
Cushman (2013). Action, outcome, and value: A dual-system framework for morality. *Personality and Social Psychology Review, 17*(3), 273–292. http://dx.doi.org/10.1177/1088868313495594.
Cushman (2015). From moral concern to moral constraint. *Current Opinion in Behavioral Sciences, 3*, 58–62.
Cushman, & Macendoe (2009). *The coevolution of punishment and prosociality among learning agents*. Paper presented at the Proceedings of the 31st annual conference of the cognitive science society, Amsterdam.
Cushman Young & Hauser (2006). The role of conscious reasoning and intuition in moral judgment: Testing three principles of harm. *Psychological Science, 17*(12), 1082–1089. http://dx.doi.org/10.1111/j.1467-9280.2006.01834.x.
Darley & Shultz (1990). Moral rules – Their content and acquisition. *Annual Review of Psychology, 41*, 525–556.
Dasgupta & Rivera (2008). When social context matters: The influence of long-term contact and short-term exposure to admired outgroup members on implicit attitudes and behavioral intentions. *Social Cognition, 26*, 112–123.
Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8*(12), 1704–1711.
Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience, 14*(2), 473–492.
Decety (2015). The neural pathways, development and functions of empathy. *Current Opinion in Behavioral Sciences, 3*, 1–6.
DeScioli & Kurzban (2009). Mysteries of morality. *Cognition, 112*(2), 281–299.
Ditto, & Liu (2011). Deontological Dissonance and the Consequentialist Crutch. In M. Mikulincer & P. R. Shaver (Eds.), *The social psychology of morality: Exploring the causes of good and evil*. APA Press.
Dolan & Dayan (2013). Goals and habits in the brain. *Neuron, 80*(2), 312–325.
Eisenberg Losoya & Spinrad (2003). Affect and prosocial responding. In R. J. Davidson, K. R. Scherer, & H. H. Goldsmith (Eds.), *Handbook of affective sciences* (pp. 787–803). New York: Oxford University Press.
Epley (2014). Mindwise: Why we misunderstand what others think, believe, feel, and want: Vintage.
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.
Festinger (1962). *A theory of cognitive dissonance* (Vol. 2). Stanford University Press.
Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology, 82*(6), 878.
Føllesdal (2004). The philosopher as coach. In E. Kurz-Milcke & G. Gigerenzer (Eds.), *Experts in science and society*. New York: Kluwer.

Gaesser & Schacter (2014). Episodic simulation and episodic memory can increase intentions to help others. *Proceedings of the National Academy of Sciences, 111* (12), 4415–4420.

Gershman Markman & Otto (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General, 143* (1), 182–194. http://dx.doi.org/10.1037/a0030844.

Gershoff Grogan-Kaylor Lansford Chang Zelli Deater-Deckard & Dodge (2010). Parent discipline practices in an international sample: Associations with child behaviors and moderation by perceived normativeness. *Child Development, 81* (2), 487–502.

Gerstenberg, Goodman, Lagnado, & Tenenbaum (2012). Noisy Newtons: Unifying process and dependency accounts of causal attribution. Paper presented at the in proceedings of the 34th.

Gino, Ayal, & Ariely (2009). Contagion and differentiation in unethical behavior. *Psychological Science*.

Gintis, Bowles, Boyd, & Fehr (2005). Moral sentiments and material interests: Origins, evidence, and consequences. *Moral Sentiments and Material Interests: The Foundations of Cooperation in Economic Life*, 3–39.

Gleichgerrcht & Young (2013). Low levels of empathic concern predict utilitarian moral judgment. *PLOS-One, 8*, e60418.

Glimcher (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences, 108*(Suppl. 3), 15647–15654.

Goldstein Cialdini & Griskevicius (2008). A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *Journal of Consumer Research, 35*(3), 472–482. http://dx.doi.org/10.1086/586910.

Good, I. J. (1967). On the principle of total evidence. *The British Journal for the Philosophy of Science, 17*(4), 319–321.

Goodman, Ullman, & Tenenbaum. Learning a Theory of Causality.

Graham, J., Waytz, A., Meindl, P., Iyer, R., Young, L. (2017). Centripetal and centrifugal forces in the moral circle: Competing constraints on moral learning. *Cognition, 167*, 58–65.

Graham Haidt & Nosek (2009). Liberals and conservatives rely on different sets of moral foundations. *Journal of Personality and Social Psychology, 96*(5), 1029–1046. http://dx.doi.org/10.1037/a0015141.

Graham Nosek Haidt Iyer Koleva & Ditto (2011). Mapping the moral domain. *Journal of Personality and Social Psychology, 101*(2), 366.

Gray Young & Waytz (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101–124.

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.

Greene (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*(2), 389–400. http://dx.doi.org/10.1016/j.neuron.2004.09.027.

Greene (2008). The secret Joke of Kant's soul. In W. Sinnott-Armstrong (Ed.). *Moral psychology* (Vol. 3). Cambridge, MA: MIT Press.

Greene, J. D. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition, 167*, 66–77.

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

Greene Morelli Lowenberg Nystrom & Cohen (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*, 1144–1154.

Greene Nystrom Engell Darley & Cohen (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389–400.

Griffiths Kemp & Tenenbaum (2008). Bayesian models of cognition. In Ron. Sun (Ed.), *The Cambridge handbook of computational cognitive modeling*. Cambridge University Press.

Grundmann (2010). Some hope for intuitions: A reply to Weinberg. *Philosophical Psychology, 23*(4), 481–509.

Haidt (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834.

Haidt & Joseph (2004). Intuitive ethics: How innately prepared intuitions generate culturally variable virtues. *Daedalus, 133*, 55–66.

Hamlin (2013). Failed attempts to help and harm: Intention versus outcome in preverbal infants' social evaluations. *Cognition, 128*(3), 451–474.

Hamlin, Wynn, & Bloom (2008). Social evaluation by preverbal infants. *Nature Materials*.

Hamlin Ullman Tenenbaum Goodman & Baker (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental Science, 16*(2), 209–226.

Hamlin Wynn & Bloom (2007). Social evaluation by preverbal infants. *Nature, 450*, 557–559.

Hamlin Wynn. Bloom & Mahajan (2011). How infants and toddlers react to antisocial others. *Proceedings of the National Academy of Sciences, 108*(50), 19931–19936.

Hamrick, Battaglia, & Tenenbaum (2011). *Internal physics models guide probabilistic judgments about object dynamics*. Paper presented at the Proceedings of the 33rd annual conference of the cognitive science society.

Hauser (2006). *Moral minds: How nature designed a universal sense right and wrong*. New York: Harper Collins.

Hauser, Cushman, Young, Jin, & Mikhail (2007). A dissociation between moral judgments and justifications. *Mind and Language*.

Heider (1946). Attitudes and cognitive organization. *The Journal of psychology, 21*(1), 107–112.

Heider, F. (1958). *The psychology of interpersonal relations*. Psychology Press.

Heiphetz, L. & Young, L. (2017) Can only one person be right? The development of objectivism and social preferences regarding widely shared and controversial moral beliefs. *Cognition, 167*, 78–90.

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

Henrich, & Henrich (2007). *Why humans cooperate: A cultural and evolutionary explanation*. Oxford University Press, USA.

Henrich Heine & Norenzayan (2010). The weirdest people in the world? *Behavioral and Brain Sciences, 33*, 61–135.

Herrmann Thoni & Gachter (2008). Antisocial punishment across societies. *Science, 319*(5868), 1362.

Ho, M. K., MacGlashan, J., Littman, M. L. & Cushman, F. A. (2017) Social is special: A normative framework for teaching with and learning from evaluative feedback *Cognition, 167*, 91–106.

Hoffman (2000). *Empathy and moral development*. Cambridge: Cambridge University Press.

Janoff-Bulman Sheikh & Hepp (2009). Proscriptive versus prescriptive morality: Two faces of moral regulation. *Journal of Personality and Social Psychology, 96*(3), 521–537.

Jones (2008). Are smarter groups more cooperative? Evidence from prisoner's dilemma experiments, 1959–2003. *Journal of Economic Behavior & Organization, 68*(3), 489–497.

Kahane Evertt Earp Farias & Savulescu (2015). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134*, 193–209.

Kleiman-Weiner, Gerstenberg, Levine, & Tenenbaum. Inference of Intention and Permissibility in Moral Decision Making.

Kleiman-Weiner, M., Saxe, R., & Tenenbaum, J. D. (2017). Learning a commonsense moral theory. *Cognition, 167*, 107–123.

Kliemann Young Scholz & Saxe (2008). The influence of prior record on moral judgment. *Neuropsychologia, 46*(12), 2949–2957.

Koenigs, Young, Adolphs, Tranel, Cushman, Hauser, & Damasio (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature Materials, 446*(7138), 908–911. http://dx.doi.org/10.1038/nature05631.

Kohlberg (1969). Stage and sequence: The cognitive-developmental approach to socialization. In D. A. Goslin (Ed.), *Handbook of socialization theory and research* (pp. 151–235). New York: Academic Press.

Koster-Hale, & Saxe (2013). Theory of mind: A neural prediction problem. *Neuron*.

Lee Talwar Ross Evans & Arruda (2014). Can classic moral stories promote honesty in children? *Psychological Science, 25*(8), 1630–1636.

Lombrozo (2017). "Learning by thinking" in science and in everyday life. In P. Godfrey-Smith & A. Levy (Eds.), *The Scientific Imagination*. Oxford University Press (in press).

Magid, R. W., Schulz, L. E. (2017) Moral alchemy: How love changes norms. *Cognition, 167*, 135–150.

Malle Guglielmo & Monroe (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186.

Marsh (2016). Neural, cognitive, and evolutionary foundations of human altruism. *Wiley Interdisciplinary Reviews: Cognitive Science, 7*(1), 59–71.

McAuliffe, K., Raihani, N., & Dunham, Y. (2017). Children are sensitive to norms of giving. *Cognition, 167*, 151–159.

McCloskey Caramazza & Green (1981). Curvilinear motion in the absence of external forces: Naive beliefs about the motion of objects. *Science, 210*, 1139–1141.

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

Moll De Oliveira Souza & Zahn (2008). The neural basis of moral cognition. *Annals of the New York Academy of Sciences, 1124*(1), 161–180.

Moore Clark & Kane (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science, 19*(6), 549–557.

Moran Young Saxe Lee O'Young Mavros & Gabrieli (2011). Impaired theory of mind for moral judgment in high-functioning autism. *Proceedings of the National Academy of Sciences, 108*(7), 2688.

Morris, A., Cushman, F. (2017) A common framework for theories of norm compliance. *Social Philosophy and Policy* (in press).

Nichols (2002). Norms with feeling: Toward a psychological account of moral judgment. *Cognition, 84*, 221–236.

Nichols Kumar Lopez Ayars & Chan (2016). Rational learners and moral rules. *Mind & Language, 31*(5), 530–554.

Nickerson (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology, 2*(2), 175.

Nisbett & Cohen (1996). *Culture of honor: The psychology of violence in the south*. Boulder: Westview Press Inc..

Paluck (2009). Reducing intergroup prejudice and conflict using the media: A field experiment in Rwanda. *Journal of Personality and Social Psychology, 96*(3), 574.

Paluck, E. L., Shafir, E. & Wu, J. (2017) Ignoring alarming news brings indifference: Learning about the world and the self. *Cognition, 167*, 160–171.

Paul (2014). *Transformative experience*. Oxford: Oxford University Press.

Paxton, Ungar, & Greene (2011). Reflection and reasoning in moral judgment. *Cognitive Science*.

Pettigrew & Tropp (2006). A meta-analytic test of intergroup theory. *Journal of Personality and Social Psychology, 90*, 757–783.

Peysakhovich, & Rand (2013). Habits of virtue: Creating norms of cooperation and defection in the laboratory. *Available at SSRN 2294242*.

Phillips, & Cushman (2017). Morality constrains the default representation of what is possible. *Proceedings of the National Academy of Sciences*, 201619717.
Piaget (1965/1932). *The moral judgment of the child*. New York: Free Press.
Pinker (2011). *The better angels of our nature: Why violence has declined*. Penguin Books.
Pizarro, & Bloom. (2003). The intelligence of the moral intuitions: Comment on Haidt (2001). *Psychol Rev*, 110(1), 193–196; discussion 197–198.
Pizarro, & Tannenbaum (2011). Bringing character back: How the motivation to evaluate character influences judgments of moral blame. *The Social Psychology of Morality: Exploring the Causes of Good and Evil*, 91–108.
Platt & Glimcher (1999). Neural correlates of decision variables in parietal cortex. *Nature, 400*, 233–238.
Priva & Austerweil (2015). Analyzing the history of Cognition using topic models. *Cognition, 135*, 4–9.
Proto, Rustichini, & Sofianos (2014). Higher intelligence groups have higher cooperation rates in the repeated prisoner's dilemma.
Railton, P. (1988). How thinking about character and utilitarianism might lead to rethinking the character of utilitarianism. *Midwest Studies in Philosophy, 13*, 398–416.
Railton, P. (2017). Moral learning: Conceptual foundations and normative relevance. *Cognition, 167*, 172–190.
Rand, Peysakhovich, Kraft-Todd, Newman, Wurzbacher, Nowak, & Greene (2014). Social heuristics shape intuitive cooperation. *Nature Communications*, 5.
Rand Greene & Nowak (2012). Spontaneous giving and calculated greed. *Nature, 489* (7416), 427–430.
Redish, A. D. (2016). Vicarious trial and error. *Nature Reviews Neuroscience, 17*(3), 147–159.
Rescorla, & Wagner (1965). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement.
Rhodes, M., & Wellman, H. M. (2017). Moral learning as intuitive theory revision. *Cognition, 167*, 191–200.
Rottman, Young, & Kelemen (2017). The impact of testimony on children's moralization of novel actions. *Emotion*.
Rottman & Kelemen (2012). Aliens behaving badly: Children's acquisition of novel purity-based morals. *Cognition, 124*(3), 356–360.
Ruff & Fehr (2014). The neurobiology of rewards and values in social decision making. *Nature Reviews Neuroscience, 15*(8), 549–562.
Rushton (1976). Socialization and the altruistic behavior of children. *Psychological Bulletin, 83*(5), 898.
Rust, & Schwitzgebel (in preparation). The moral behavior of ethics professors: Responsiveness to student emails.
Schultz Nolan Cialdini Goldstein & Griskevicius (2007). The constructive, destructive, and reconstructive power of social norms. *Psychological Science: A Journal of the American Psychological Society/APS, 18*(5), 429–434.
Schwitzgebel (2009). Do ethicists steal more books? *Philosophical Psychology, 22*(6), 711–725.
Schwitzgebel, & Rust. (in preparation). The self-reported moral behavior of ethics professors.
Schwitzgebel & Cushman (2012). Expertise in moral reasoning? Order effects on moral judgment in professional philosophers and non-philosophers. *Mind & Language, 27*(2), 135–153. Doi 10.1111/J.1468-0017.2012.01438.X.
Schwitzgebel & Cushman (2015). Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition, 141*, 127–137. http://dx.doi.org/10.1016/j.cognition.2015.04.015.
Schwitzgebel & Rust (2010). Do ethicists and political philosophers vote more often than other professors? *Review of Philosophy and Psychology, 1*, 189–199.
Seymour, Singer, & Dolan (2007). The neurobiology of punishment. *Nature Reviews Neuroscience*.
Shenhav & Greene (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. *Journal of Neuroscience, 34*(13), 4741–4749.
Siegel, J., Crockett, M. J., & Dolan, R. J. (2017). Inferences about moral character moderate the impact of consequences on blame and praise. *Cognition, 167*, 201–211.
Singer (1981). *The expanding circle*. Clarendon Press, Oxford.
Sloane Baillargeon & Premack (2012). Do infants have a sense of fairness? *Psychological Science, 23*(2), 196–204.
Stagnaro, M. N., Arechar, A., & Rand, D. G. (2017). From good institutions to generous citizens: Top-down incentives to cooperate promote subsequent prosociality but not norm enforcement. *Cognition, 167*, 212–254.
Steckler, C., Woo, B. M., & Hamlin, J. K. (2017). The limits of early social evaluation: 9-Month-olds fail to generate social evaluations of individuals who behave inconsistently. *Cognition, 167*, 255–265.
Sutton, & Barto (1998). *Introduction to reinforcement learning*. MIT Press.
Tamir & Mitchell (2010). Neural correlates of anchoring-and-adjustment during mentalizing. *Proceedings of the National Academy of Sciences, 107*(24), 10827–10832.
Tenenbaum Griffiths & Kemp (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*(7), 309–318. http://dx.doi.org/10.1016/j.tics.2006.05.009.
Tenenbaum Kemp Griffiths & Goodman (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*(6022), 1279–1285.
Theriault, Waytz, Heiphetz, & Young (2017). Examining overlap in behavioral and neural representations of morals, facts, and preferences.
Theriault, Waytz, Heiphetz, Young, & Theriault (2017). Metaethical judgment relies on activity in right temporoparietal junction: Evidence from neuroimaging and transcranial magnetic stimulation. Retrieved from osf. io/j2wxk.
Thorndike (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Monographs: General and Applied*, 2(4), i-109.
Turiel (2005). Handbook of moral development. In M. Killen & J. Smetana (Eds.), *Handbook of moral development*. Lawrence Erlbaum Associates Publishers.
Uhlmann Pizarro Tannenbaum & Ditto (2009). The motivated use of moral principles. *Judgment and Decision Making, 4*(6), 476–491.
Uhlmann Pizarro. & Diermeier (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10*(1), 72–81.
Vaish Carpenter & Tomasello (2009). Sympathy through affective perspective taking and its relation to prosocial behavior in toddlers. *Developmental Psychology, 45* (2), 534.
Walker, C. M., & Lombrozo, T. (2017). Explaining the moral of the story. *Cognition, 167*, 266–281.
Warneken & Tomasello (2006). Altruistic helping in human infants and young chimpanzees. *Science, 311*(5765), 1301.
Waytz, Iyer, Young, Haidt, & Graham. (in prep). *Your Ambit of Concern: Political Ideology, Empathy Distribution, and the Expanse of the Moral Circle*.
Westgate Riskind & Nosek (2015). Implicit preferences for straight people over lesbian and gay men weakened from 2006 to 2013. *Collabra, 1*, 1–10.
Wimmer & Perner (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13*, 103–128.
Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology, 47*(2), 237.
Wright, & Bartsch (2008). Portraits of early moral sensibility in two children's everyday conversations. *Merrill-Palmer Quarterly* (1982), 56–85.
Young Bechara Tranel Damasio Hauser & Damasio (2010). Damage to ventromedial prefrontal cortex impairs judgment of harmful intent. *Neuron, 65*(6), 845–851.
Young Cushman Hauser & Saxe (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences, 104*(20), 8235–8240.
Young & Dungan (2012). Where in the brain is morality? Everywhere and maybe nowhere. *Social Neuroscience, 7*(1), 1–10.
Zaki, Kallman, Wimmer, Ochsner, & Shohamy. (2016). Social cognition as reinforcement learning: Feedback modulates emotion inference. *Journal of Cognitive Neuroscience*.
Zaki & Mitchell (2011). Equitable decision making is associated with neural markers of intrinsic value. *Proceedings of the National Academy of Sciences, 108*(49), 19761–19766.