Cognition 167 (2017) 172–190

Contents lists available at ScienceDirect

Cognition

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

Original Articles

Moral Learning: Conceptual foundations and normative relevance

Peter Railton

Department of Philosophy, University of Michigan, 2215 Angell Hall, 435 South State Street, Ann Arbor, MI 48109-1003, United States

ARTICLE INFO

Article history:
Received 15 June 2016
Revised 6 August 2016
Accepted 25 August 2016
Available online 3 September 2016

Keywords:
Moral judgment
Moral development
Causal model
Evaluation
Simulation
Empathy
Bayesian
Reinforcement learning
Dual-process
Model-free and model-based learning and control
Trolley problem

ABSTRACT

What is distinctive about a bringing a *learning* perspective to moral psychology? Part of the answer lies in the remarkable transformations that have taken place in learning theory over the past two decades, which have revealed how powerful experience-based learning can be in the acquisition of abstract causal and evaluative representations, including generative *models* capable of attuning perception, cognition, affect, and action to the physical and social environment. When conjoined with developments in neuroscience, these advances in learning theory permit a rethinking of fundamental questions about the acquisition of moral understanding and its role in the guidance of behavior. For example, recent research indicates that spatial learning and navigation involve the formation of non-perspectival as well as egocentric models of the physical environment, and that spatial representations are combined with learned information about risk and reward to guide choice and potentiate further learning. Research on infants provides evidence that they form non-perspectival expected-value representations of agents and actions as well, which help them to navigate the human environment. Such representations can be formed by highly-general mental processes such as causal and empathic simulation, and thus afford a foundation for spontaneous moral learning and action that requires no innate moral faculty and can exhibit substantial autonomy with respect to community norms. If moral learning is indeed integral with the acquisition and updating of casual and evaluative models, this affords a new way of understanding well-known but seemingly puzzling patterns in intuitive moral judgment—including the notorious "trolley problems."

1. Introduction

A query to Google Books requesting an Ngram from 1950 onwards for the phrase *moral development* reveals that this expression underwent a dramatic growth in frequency from 1960 to 1980, before declining gradually to 2008 (the latest year for which results are given). Adding an Ngram for *social learning* shows that this expression followed essentially the same trajectory, climbing yet more dramatically to its 1980 peak before drifting downward in recent years. But request an Ngram for *moral learning* during the same period, and the Ngram Viewer draws a blank. Which leads to the question: If there already are well-established research literatures in moral development and social learning, what might a moral learning perspective add?

The existing literatures in moral development and social learning are far too varied and extensive, and the field of moral learning far too undeveloped, to permit more than a preliminary comparison and contrast. Certainly there is much by way of overlap. A moral learning approach shares with social learning theory the idea that much of our learning takes place by observing others, rather than through direct external reward or punishment. And it shares with moral development theory the idea that our capacity for moral thought emerges over time, drawing upon the development of capacities in other domains.

However, a moral learning approach sees the acquisition of moral understanding as the result of domain-general learning processes, and thus as an integral part of our modeling of the physical and social world. Such modeling generates expectations continuously that guide perception, thought, and action, and permit learning from discrepancies with expectation throughout life. Moral learning therefore can go beyond the acquisition of known moral concepts or internalization of prevailing social norms, and can extend to the formation of novel moral concepts and evaluations, resulting in dramatic personal and social change even within one lifetime.

In this paper, I will examine a series of issues, consideration of which makes it possible to give more substance to a moral learning perspective. Section 2 will present criteria for distinctively *moral* learning. Section 3 will look at causal and evaluative learning as exemplars of the kind of *learning* moral learning might be, and ask why *now* is a particularly apt moment for asking about the power of learning. Section 4 will then apply the model-based

E-mail address: prailton@umich.edu

http://dx.doi.org/10.1016/j.cognition.2016.08.015
0010-0277/© 2016 The Author. Published by Elsevier B.V.

picture of learning developed in Section 3 to the moral case, presenting evidence for the acquisition of non-perspectival evaluative representations that satisfy the criteria presented in Section 2. Section 5 will look into the phenomenon of "intuitive judgment," and use informal student polling data to ask how a "deep" moral learning perspective might account for the puzzling patterns of intuitive moral judgment found in "trolley problems." And Section 6 will conclude by briefly considering how explicit and implicit processes interact in moral learning.

2. Identifying the subject matter

Learning is a success term, and if moral learning is to be an integral part of the knowledge we gain in representing ourselves and the world, then it must be subject to some notion of representational success. Does this require a theory of moral learning to take a stand on which moral theory is correct—seemingly in violation of David Hume's celebrated distinction between *is* and *ought* (1738/1978)?

It is possible, however, for a theory of moral learning to bracket many controversial moral questions by focusing instead on *criteria* of moral evaluation that are shared across a wide range of normative moral theories. Just as we can speak of criteria characteristic of a *scientific point of view* that are implicitly or explicitly followed by those pursuing competing theories, we can speak of criteria characteristic of a *moral point of view*. It is thanks to such shared criteria that there can be a scientific or moral "community," with shared methods and questions, and meaningful disagreement over answers.

Scientific and moral inquiry both aspire to a kind of objectivity that overcomes the limitations of subjective or sectarian perspectives or interests by following methods, and seeking understanding and justification, that are (i) *impartial*, (ii) *general*, (iii) *consistent* (or, more broadly, *coherent*), and (iv) *independent of appeals to special authority*. For example, both require that *like cases be treated alike*, and that the evidence or grounds given in defense of particular positions be in principle *shareable*. Moreover, competing parties to moral and scientific disputes agree that their disputes are not merely speculative. That is, they see themselves as seeking to answer questions about what to believe and how to apply this in practice—whether this is a matter of accepting a scientific hypothesis, following a methodological norm, or deciding upon an ethical course of action. Let us call this the criterion of (v) *thought- and action-guidingness*. One could hardly make sense of the intensity of scientific and moral disputes if one thought that making up one's mind in scientific or moral disputes were a merely notional matter, with no relevance to how we should think and act.

Of course, moral disputes also differ from scientific disputes in a number of respects. For example, morality has a proprietary, non-instrumental concern with questions of (vi) the *harm or benefit of those actually or potentially affected*. Scientists of course are not indifferent to such questions, but they are not treated as an essential part of the evidence or grounds for scientific judgment. Criterion (vi) does not say that impartial concern with harm or benefit is the entire basis of morality, as some utilitarians maintain, but rather that harm and benefit have direct relevance to moral judgment across the full array of major ethical traditions—including deontologies (which typically include duties not to harm and to render assistance to those in need) and virtue theories (which typically connect virtue with human flourishing, and identify beneficence and generosity as central virtues).

To study moral learning or scientific learning, then, it is not necessary to embrace a particular substantive theory or to provide a definition of *morality* or *science*—it is enough to study how individuals or groups develop, and treat as normatively important, forms of inquiry or ways of regulating thought and action that exhibit such features as (i)–(v) or (i)–(vi).

From an evolutionary standpoint, it can appear quite extraordinary that people would impose upon themselves the limitations of forms of inquiry and practice that would meet criteria (i)–(v) or (i)–(vi). Why would natural selection favor the development of mental processes or social dispositions that can be so independent of the reproductive interests of individuals and their kith and kin? Answering this question is one of the key challenges faced by accounts of scientific or moral learning—and we will have something to say about it, below.

3. Causal and evaluative learning

3.1. Philosophical background

Hume framed one of the foundational texts of modern philosophy, *A Treatise of Human Nature* (1738/1978), in terms of the joint problem of understanding *how* we arrive at the attitudes we do on the basis of experience, and whether these attitudes are *warranted*. Hume focused especially on causal and moral beliefs, and perhaps surprisingly, the author of the *is*/*ought* distinction emphasized the fundamental similarities of these two forms of domains of thought. Hume saw that there is a general problem of bridging the gap between sensory impressions, which are particular, concrete, actual, and transient, and what we come to believe on their basis, which is general, abstract, modal, and temporally-extensive. How, he asked, do we come to form causal and moral beliefs which *logically* outstrip all our evidence, and what does this tell us about how or why they might nonetheless be justified?

Hume concluded that *we* must add something to sensation to bridge this gap. Earlier philosophers had often invoked innate ideas, yet these could not really solve the problem he had identified. After all, innateness is not validity, and even if we were endowed with valid general, abstract ideas or rules, we would still have to figure out how to apply these to particular, transient, unruly experiences, or to decisions or actions in concrete contexts. Abstract concepts and rules do not apply themselves, and to appeal to yet other innate concepts or rules to tell us how and when to apply them would be to launch a regress—and "it is impossible for us to carry on our inferences *in infinitum*" (1738/1978; sect. I. iii.4).

Hume's answer is that *imaginative projection* effects the bridge that strictly logical inference cannot. He posited general, default psychological dispositions to respond to certain regularities in sensory experience by mentally extending these patterns to novel experiences and abstract relations of similarity and difference. Forming expectations on the basis of such default projective dispositions might seem to be epistemically reckless, but Hume argued that, by "spreading itself over the world" in this way, belief could make experience into trial-and-error experimentation. Belief for Hume is an active *sentiment* rather than a mere idea, and its projective "initial impulse" will be "broke into pieces" in response to the proportion of success or failure in expectation (1738/1978; sect. I. ii.12). Although Hume is often considered an outright skeptic, on a more plausible interpretation he combined skepticism about the powers of pure reason with realism about the ways sentiments such as belief can ground us in reality and attune our thought and action to the world. Indeed, he claimed, logical reasoning itself can avoid regress only because belief projects spontaneously along the network of the "association of ideas" via relations of similarity and analogy—if such default mental operations cannot be trusted, then reasoning cannot be trusted either (1738/1978; conclusion of Book 1).

Other sentiments contribute to other forms of grounding and attunement. In the imaginative projections that guide our lives as continuing agents situated in a social world, *sympathy* presents to us simulated experience of what it is like to occupy perspectives other than our own current point of view. This gives experiential reality and force both to our own possible fate in the future and to the potential weal or woe of others. As a result, we tend spontaneously to approve of actions, practices, and states of character not only insofar as these benefit our current selves and those closest to us, but also in light of their tendencies toward longer-term or more general benefit or harm (1738/1978; sect. III.iii.6). No innate first principles are needed for us to discover the distinction between virtue and vice, and the mechanisms by which we make this discovery connects this abstract distinction to concrete experience and circumstances, and to what *moves* us to action (1738/1978, Conclusion of Book III).

3.2. A contemporary convergence: Hume 2.0

Today the successors to Hume's epistemology are found in such theories as *reinforcement learning* and *Bayesian updating*. These learning processes are by their nature expectation-based or projective—and recent years have shown us that they are much more powerful in generating discrimination, intelligence, abstract generalization, and even creative action than they were previously imagined to be (Le et al., 2012; Mnih et al., 2015; Silver et al., 2016; Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

Probabilistic models of spontaneous belief formation and revision had been long dismissed as psychologically unrealistic, and at odds with people's actual "intuitive" judgments (Kahneman, 2011; Kahneman & Tversky, 2000). But in the last two decades, developments in neuroscience have provided evidence that these models might be realistic after all, as actual neural processes are observed to approximate such models in sensory coding, visual discrimination and search, and reinforcement learning (Dayan & Daw, 2008; Lee & Mumford, 2003; Ma, Navalpakkam, Beck, van den Berg, & Pouget, 2011; Schultz, 2002).

Simultaneously, theoretical work on hierarchical neural nets and Bayesian causal graphs provided plausible pictures of how the brain might use expectation-based feedback to build abstract generative models of perceptual inputs (Hinton, 2007; Holyoak & Cheng, 2011; Pearl, 2009). And theoretical and applied work in motor control showed how such models could guide behavior optimally, in ways that yielded a good approximation of actual movement (Berthier, Rosenstein, & Barto, 2005; Liu & Todorov, 2007). To complete the picture, these same models used inversely exhibited a good fit for how humans perceive similarity, attribute causation and intention, "break belief into pieces" to guide behavior, and use variation in outcomes to assess risk and adjust learning rate (Behrens, Woolrich, Walton, & Rushworth, 2007; Knill & Pouget, 2004; Kording & Wolpert, 2006). (For a critical perspective on probabilistic and optimizing approaches, see Marcus & Davis, 2013.)

It was the challenge of spatial navigation that first gave rise to the idea that intelligent animals construct persistent *models* of the world around them, and use these "cognitive maps" to guide choice and action (Tolman, 1948). Recently, developments in neuroimaging have made it possible to study in detail the formation and use of "cognitive maps" as an animal learns a maze (Langston, Ainge, & Covey, 2010)—combining information from vision, touch, motion, and head direction to create perspectival and non-perspectival representations of place and space. These representations can guide navigation flexibly by dynamic updating (Moser, Kropff, & Moser, 2008), and can promote "off-line" learning as well. When the rat rests or sleeps, repeated activations occur in these representations (Foster & Wilson, 2006; Ji & Wilson, 2007). During these episodes of re-activation, the trajectories simulated include directions of motion the rat did not experience as well as those it did, and activation focuses especially in regions of the maze less frequently explored—a pattern typical of a learner seeking to extract maximum information from a body of data, rather than a creature of habit caught in the mental "grooves" of the most-entrenched past patterns (Gupta, van der Meer, Touretzky, & Redish, 2010).

Moreover, during these bouts of re-activation, something surprising occurs: "short cut" paths never taken begin to be constructed (Gupta et al., 2010). It perhaps should not be surprising that mammals, shaped over hundreds of millions of generations to be successful foragers under conditions of scarcity, would be built to exploit the energy-saving strategy of building up and exploring a landscape *mentally* on the basis of partial information, leveraging experience to prepare novel responses that might prove more effective or efficient. Such simulation and comparison of possibilities appears to be an intrinsically-motivated activity of mind, enriching spatial representations even in the absence of new external cues or rewards.

Simulation and comparison of pathways also takes place "on line," as the rat negotiates the maze. Once it has had a chance to explore a maze, and before it reaches the point of overtraining, as a rat approaches an important choice-point, activation in its mental map spreads transiently *forward* down the alternate paths ahead (Johnson & Redish, 2007). And if the chosen action fails to live up to expectations—for example, if the rat does not hear the click of the food-tray window opening up ahead—activation in the mental map can then spread *backward* down the other arm of the maze, and the animal reverses direction (Johnson & Redish, 2007).

Ethologists had observed as early as the 1970s that animals in natural or controlled environments manage to develop near-optimal foraging strategies, as determined by the animals' energetic and nutritive needs, the distances traveled, the likelihood of finding a given resource at a location, the marginal rate of return from a resource, and the need to explore as well as exploit (Dugatkin, 2004; Krebs, Ryan, & Charnov, 1974). It would appear, then, that reinforcement learning is capable of providing action-guiding information sufficient for this complex task. Evidence suggests that rats build mental models of causal structure (Blaisdell, Sawa, Leising, & Waldmann, 2006), and that monkeys use mental models to "work backwards" from desired outcomes to planned actions, and to form predictive expectations regarding novel cases via abstraction and analogy (Tanji, Shima, & Mushiake, 2007). Expectation-based action-guidance via imaginative projection and simulation turns out to be just as central to intelligence as Hume imagined, and even more pervasive. Evidently, the metabolically-expensive brains needed for effective "prospection" paid their way evolutionarily (Seligman, Railton, Baumeister, & Sripada, 2013, 2016).

3.3. Learning and nativism

As in Hume's day, talk of the power of learning naturally gives rise to questions about the relative contribution of experience-based learning vs. innate knowledge or "modules." Of course, any learning system depends upon some unlearned structures—for example, the default dispositions and "priors" needed to start error-based or Bayesian learning, or the neural structures that underlie the general capacity for reinforcement learning and model-based control. The answer to "nature vs. nurture" is always *both*, but what recent advances have placed in question, however, are long-standing assumptions about the need to posit more extensive and substantive innate knowledge or "modules" to explain how the mind bridges the gap between actual, limited sensation and open-ended, generative competencies and abstract knowledge structures. Perhaps more generic learning processes with revisable priors can do the job.

For example, it has long been argued that human infants must have something akin to an innate "linguistic faculty" or "language module," since they manage almost universally to acquire an open-ended syntactic and semantic competence in their native tongue on the basis of relatively limited amounts of data or instruction—the "poverty of the stimulus" argument (for a review, see Pullum & Scholz, 2002). Such explanations, though, still face the Humean problem mentioned at the outset: even to engage such a module the infant must develop very considerable open-ended, abstract, generalizing capacities using equally "impoverished" stimuli. In the case of language, the infant must already have achieved considerable auditory competence in identifying certain streams of overheard sound as communicative, parsing these streams into significant, repeatable units despite wide speaker-to-speaker variation in actual sound patterns, distinguishing some of these abstract unit-types as expressions of agreement or disagreement, identifying adult communicative and referential intentions, and so on. Children do, in fact, acquire these competencies during the period when they are first learning language, also without much by way of explicit instruction. (How much explicit instruction could an infant receive *without* such capacities?) We might additionally posit a "theory of mind module" with a special "social pedagogy sub-module" to help explain this, but then engaging *these* modules experientially likewise requires that the infant have developed capacities to parse the stream of experience in projectable ways, distinguishing continuing objects and imputing similarities in patterns of motion and causal relations, and so on—again, without extensive explicit instruction in these more generic tasks. Perhaps elaborating modules is skirting the basic problem.

What we can observe is that these competencies in language, theory of mind, and causation are all developing during the same period, and that progress in one seems to depend upon progress in the others. We also have neuroimaging evidence that, in the more fully-developed mind, the brain areas involved in these different forms of cognition overlap substantially, and seem to make use of common, algorithm-like learning processes (Buckner, Andrews-Hanna, & Schacter, 2008). And we have recently learned that artificial systems built upon fairly generic learning algorithms and memory structures, of kinds broadly similar to those found in the human mind, are capable of tackling artificial intelligence tasks previously thought to require extensive "purpose-built" engineering of features and heavy dependence upon already-developed human expertise (Le et al., 2012; Mnih et al., 2015; Silver et al., 2016). While research in these areas is still young, still, the capacity of such "deep learning" systems to solve open-ended arrays of problems using a generic architecture, including the development of novel value functions, causal hypotheses, and strategies, is a glimpse of a "proof of possibility" for an enlarged understanding of the contribution learning can make to the development and structure of abstract representational systems like the human mind.

A common feature of these generic, but powerful, learning systems is that they operate by trying to *generate* their input—that is, by finding and modeling general, projectable patterns in the input on the basis of which to predict subsequent input, exploiting error-based learning to "train" themselves without requiring extensive explicit instruction. Are human infants like this? Although very young human infants might not seem highly attentive the statistics of the world around them, in the first days of life they can already detect differences in rhythms between languages (Nazzi, Bertoncini, & Mehler, 1998), and by 8 months their sensitivity to conditional probabilities in overheard speech is sufficient to enable them to segment fluent speech into words, and to focus attention on verbal stimuli with the greatest information value (Aslin, Saffran, & Newport, 1998; Kidd, Piantadosi, & Aslin, 2012). Studies of anticipatory looking suggest that infants have developed generative causal models by 8–9 months (Sobel & Kirkham, 2006, 2007), and by 15 months they pay attention to sampling methods in making these causal generalizations (Gweon, Tenenbaum, & Schulz, 2010). At 12 months, infant looking times at complex motion displays approximate the predictions of a Bayesian "ideal observer" model, apparently using underlying abstraction-based inference (Teglas et al., 2011), and by 16–18 months infants appear able to use causal models inversely, inferring from "informal experiments" the most likely hypothesis, and using this to "test" alternative causal pathways (Gopnik & Schulz, 2004)—e.g., using behavior to infer underlying intentions, even when the action is unsuccessful (Gweon & Schulz, 2011).

The *habituation* paradigm used in many of these experiments is, in effect, a *projection* paradigm—indicating how infants actively model incoming data to increase sensitivity to error. Drawing upon two decades of studies, Gopnik and Wellman (2012; see also Wellman, 2014) argue that Bayesian hierarchical causal learning is sufficient to account for many aspects of the infant's development of a "theory of mind." As we would expect from the generic character of reinforcement learning and Bayesian probabilistic inference, these capacities persist into adulthood and show up across such domains as vision (Geisler, 2011), motor control (Liu & Todorov, 2007), causal inference (Holyoak & Cheng, 2011), and imputation of intention or detection of change using limited information (Diaconescu et al., 2014; Gallistel, Liu, Krishan, Miller, & Latham, 2014). The job of learning is never done, and expertise can improve over decades (Yarrow, Brown, & Krakauer, 2009).

While this approximation of normatively appropriate probabilistic learning, representation, and decision- and action-guidance would appear to be contradicted by decades of research on cognitive biases and heuristics (Kahneman, 2011; Kahneman & Tversky, 2000), that body of research is heavily based upon people's explicit responses to word problems in settings of conscious judgment or choice, rather than settings in which implicit learning with feedback takes place. For example, children's implicit learning in the well-known "false-belief" task, can run months or years ahead of their ability to give accurate verbal responses to questions addressed to them (Luo, 2011), and the acquired causal models that guide a child's motor behavior yield much more accurate expectations than the "folk physics" children produce in response to queries (Gelman & Legare, 2011). Even for adults, underlying probabilistic models will contain many "hidden layers" of relations among features, which can guide choice implicitly, yet cannot be directly accessed introspectively (Kolling, Behrens, Mars, & Rushworth, 2012). Interestingly, when a number of the classic experiments from the heuristics and biases literature are redone in a setting where implicit learning is possible and collateral implicit statistical knowledge can be used, some of the well-known anomalies don't persist (see Friedman, 1998; Kolling et al., 2012; Pleskac & Hertwig, 2014; Shanks, Tunney, & McCarthy, 2002). Humans are far from perfect in their reasoning, but the underlying problem doesn't seem to be an implicit or "intuitive" system that wasn't built for statistics, probabilistic inference, or abstract generalization. What might be most important, in the end, is not how much might or might not be contributed by a native endowment or social conditioning—but rather how much *flexibility and depth of understanding in response to evidence and inference* is possible. That is, how much *learning*—wherever one starts.

3.4. Foraging for value

In optimal foraging behavior, animals behave "as if" they were guided by evaluative assessments and rational choice theory, learning the magnitude of costs, benefits, and risks, and selecting actions or policies on the basis of expected value. We have seen that Tolman's early observation that animals behave "as if" guided by internal cognitive maps has received powerful support from subsequent neuroscience, even if many questions remain unanswered (Moser et al., 2008). But what about the *evaluative* or *decision-theoretic* "as if" story for observed optimal patterns in foraging behavior? Is there an "inner mapping" of the expected-value landscape of choice, which is updated dynamically and actually guides choice and behavior?

Two decades of study of the neural mechanisms underlying reinforcement learning in monkeys and other foraging mammals support a positive response. Behaviorists and neoclassical economists thought it was in principle impossible to factor choice behavior into determinate evaluative vs. risk components, since all reinforcement or selection takes place at the level of external behavior rather than inner mechanisms. Yet as with spatial and causal mapping, it seems to pay to have representational capacities that can acquire the *structure* of the world, and magnitude of risk vs. magnitude of reward correspond to importantly different structural features of the world. Single-neuron recordings indicate that the mind of intelligent mammalian foragers does indeed keep separate track of risk vs. reward, and combines these as decision weights in order to assess and compare the expected value of outcomes, and to choose (Fiorillo, Tobler, & Schwartz, 2003; Grabenhorst & Rolls, 2011; Preuschoff, Bossaerts, & Quartz, 2006; Tobler, O'Doherty, Dolan, & Schultz, 2006). These risk and reward signals, moreover, exhibit many of the formal features of probability and utility functions (Lak, Stauffer, & Schultz, 2014; Stauffer, Lak, & Schultz, 2014), and the common pathways to action shared by diverse kinds of risk and reward suggest that something like cardinal utility comparisons guide choice (Quartz, 2007).

To be sure, the claim is not that these animals consciously follow the principles of rational decision theory—that would require not only the ability to represent decision weights, but to use meta-representations of these weights in self-conscious deliberation about how to act. Rather, like human kindergarteners whose comprehension and speech fluently follow norms of grammar and conversation, primates presumably think and act in accord with norms of rational choice implicitly, via first-order representations and processing. While we do not have neuron-by-neuron evidence for humans the way we do for primates, the brain structures involved in primate decision-making have functional homologues in humans, and observations of metabolic activity in the human brain are consistent with underlying processing that has similar algorithmic structure. Moreover, when humans are given simulated foraging tasks their performance, like animal performance, tends to approximate optimality (Behrens et al., 2007; Kolling et al., 2012).

4. Moral learning

4.1. Non-perspectival evaluative models

If moral learning in humans is to take place spontaneously, then it should issue spontaneously in first-order evaluative representations that meet, or approximately meet, appropriate criteria for *moral* assessment, such as those adumbrated in Section 1: (i) impartiality, (ii) generality; (iii) consistency or coherence; (iv) independence from authority or convention; and (vi) representing benefits or harms to those affected in such a way as to be (v) non-instrumentally thought- and action-guiding for the agent.

And if domain-general learning processes are truly to lie at the heart of the development of moral understanding, then this should not depend upon the triggering of an innate "moral module" or external socialization into a set of norms. Rather, evaluative *content* meeting moral criteria should be acquired through experience itself. In fact, recalling our earlier discussion of language learning, a dose of moral learning might actually help solve the "application" problem for nativist and socializing accounts, since a set of moral rules cannot apply itself any more than a grammar can, and so the infant must already have begun to develop discriminative abilities to factor experience and behavior into such categories as harm, benefit, risk, cooperation, and intention vs. accident. Otherwise violations of moral norms will not be appropriately discerned, and principles of helping others or punishing transgressors will not be appropriately engaged. Yet this moral "pre-learning" already involves organizing experience and regulating behavior in ways that conform to (i)–(vi). For example, *imitation* has often been invoked as an "automatic" path into moral development, but evidence increasingly suggests that imitation is a flexible skill that emerges along with other forms of perceptual, causal, and evaluative learning (Williamson, Meltzoff, & Markham, 2008), and is itself applied selectively by infants as a reflection of their experience of the epistemic or moral qualities of potential models (Heyes, 2016; Zmyj, Buttelmann, Carpenter, & Daum, 2010). These capacities for implicit discrimination and choice along morally-relevant dimensions appear to develop integrally with causal, conceptual, intentional, and evaluative learning in general. As a result, maturing individuals become increasingly able to give *explicitly moral* expression to what they think and feel, and thus to participate more fully in individual and shared moral deliberation.

We saw in the case of our foraging animal ancestors that natural selection appears to have favored a spatial navigation system that constructs non-perspectival as well as perspectival representations of the physical environment. These representations are used prospectively to locate the self with regard to the rest of the world, to link proximate action with more distant goals, to compare alternate paths of action, and thus to guide decision-making. Mental mapping, in other words, appears to have evolved to represent the abstract "impartial" geometry of space within which to embed concrete individual locations and pathways. Such representations constitute models of past experience but also serve as "test beds" for projectively simulating the trajectories of oneself and others, facilitating individual planning and social coordination and cooperation—even avoiding collisions as we walk or drive. Notably, when engineers build autonomous robots and vehicles from scratch, they find it efficient and effective to design them to model spatial relations and simulate possible trajectories in similar ways (Katrakazas, Quddus, Chen, & Deka, 2015). Is there evidence in the case of humans—and perhaps other highly social animals as well—of a similar fundamental capacity for non-perspectival as well as perspectival evaluative "mapping" of social space and its possibilities, which likewise models experience and subserves simulation to enhance capacities for action, coordination, and cooperation—including the avoidance of social collisions?

As a first step, we should note that prediction-error signals in the brain track expectation violation not only for such "natural" or concrete values as food and mating, but also for abstract values like uncertainty, conventional values like money, and social values like trustworthiness (Behrens, Hunt, Woolrich, & Rushworth, 2008; Fiorillo et al., 2003). Evaluative expectations come from an array of sources within the affective system broadly understood, with the result that the brain can respond to multiple kinds of values simultaneously (Lak et al., 2014). At the same time, pathways to action collect together these multiple streams of evaluative information, including uncertainty, to permit comparisons of expected value and risk to guide decision-making—and set the stage for the next round of error-based learning (Grabenhorst & Rolls, 2011; Johnson, van der Meer, & Redish, 2007; Stauffer et al., 2014). This appears to be a system with sufficient range and flexibility to represent a wide array of morally-relevant considerations, and with a learning capacity to promote fine-grained attunement of attitude and behavior—indeed, this system appears to be shared by moral and non-moral decision-making alike (Buckner et al., 2008; Decety & Porges, 2011; Shenhav & Greene, 2010).

And there is no area where it is more important for the human infant to develop fine-grained, well-attuned evaluative representations than the interpersonal, given an infant's nearly complete dependence on help from others and vital need to learn from them. Grasping the dynamics of persons is central to interacting with them reciprocally and successfully, and, here again, the world the infant mind is striving to predict is more predictable if that mind models things in ways that go beyond its own perspective. The behavior of others, for example, is more projectable from *their* mental states than one's own, and social dynamics are more projectable if one takes into account general causal and structural relations that are inherently non-perspectival. A representational system adequate to the challenge of building generative causal models will possess a capacity to correct for perspective or personal interests (i), to generalize across cases (ii), to treat like cases alike (iii), to operate in some measure independently of received opinion or authority relations (iv), and to regulate thought and action accordingly (v). And if that model is to generate the behavior of *agents*, it will also need to be able represent benefits and harms to others as *they* experience them, not just as one would personally benefit from seeing them (vi).

While these considerations are plausible, for the hypothesis of moral learning to be credible, we need more direct evidence that infants take a spontaneous interest in modeling and evaluating agents' intentions, actions, and outcomes from non-perspectival as well as perspectival standpoints, and that such representations are for them non-instrumentally thought- and action-guiding.

That infants pay close attention to normative aspects of third-party behavior, and use this information to regulate their own conduct and choices in normatively-relevant ways, is perhaps most readily seen in the *epistemic* case, e.g., learning whom to trust about what. As we saw in discussing theory of mind, beginning sometime in the first year and continuing progressively as greater physical, cognitive, and social competence develops, infants use contextual and social cues derived from third-person observation to adjust their attention, learning, and behavior (Wellman, 2014). By 12 months infants are able to distinguish reliably between "unable" and "unwilling" adult behavior (Woodward, Sommerville, Gerson, Henderson, & Buresh, 2009). By 16 months, they show heightened attention to mistaken labeling and labelers in learning word use (Koenig & Echols, 2003), and by 36–48 months, children are using observation of third-party adult behavior and its outcomes to make discriminations of an adult's accuracy, knowledgeability, competence, reliability, deceptiveness, and quality of will, and to use these discriminations in deciding whether to trust the adult in learning a new word, the location or identity of a hidden object, the tastiness or healthiness of a novel food, or a counter-evident fact (Doebel & Koenig, 2013; Lane, Harris, Gelman, & Wellman, 2014; Nguyen, Gordon, Chevalier, & Girgis, 2016; Sobel & Corriveau, 2010). An example of the non-perspectival aspect of these epistemic evaluations is the fact that, with age, infants become increasingly willing to rely upon information from an *unfamiliar* individual who displays greater epistemic reliability than a familiar caregiver (Harris & Corriveau, 2011). As they move into their fourth year, children exhibit more fine-grained and well-modulated epistemic sensitivities, for example paying increased attention to the domain-relevance of imputed adult traits in making decisions about what to learn from whom (Sobel & Corriveau, 2010). It is clear why non-perspectival evaluation of a third-party's capacities and motivations has special importance for the child's modeling of the social world. For example, even if a newly-encountered adult is playing favorites in a way that happens to benefit the child at the moment, the fact that this adult plays favorites remains important for the child to learn and take into account, since favorites change.

Infants could hardly begin to learn at all if they did not start life with "priors" that amount to default trust in their own sensory system and memory—trust without evidence of reliability. After all, without reliance upon sensation and memory, how could they gather evidence to test reliability? Yet priors are neither destiny nor blunt heuristics. With growing experience and perceptual and cognitive development, infants become more discriminating about when, and how much, to trust their eyes, ears, or memory. For example, 3–6 year-olds with firmer grip upon the (abstract) appearance/reality distinction show greater willingness to lend some credence to the testimony of an informant who indicates something contrary to what they took themselves to have (concretely) seen (Lane et al., 2014).

Similarly, infants could hardly begin to have successful interaction with, or learning from, adults if they did not start life with "priors" that amount to according adults some measure of default trust. For example, during the first three years of life, infants tend to require more evidence for a negative as opposed to positive trait attribution to an adult (Bosevoski & Lee, 2006). Yet trust-based expectations also drive feedback learning from the consequences of relying upon a given adult, so that an infant's causal/evaluative social models become better attuned to the actual distribution of trustworthiness in the adult world. For example, while infants from early on appear to be especially wary of adults who show signs of harmful behavior, by 8 months they can excuse adult harmful behavior that takes the form of punishing an anti-social individual (Hamlin, Wynn, Bloom, & Mahajan, 2011). And by the fourth year, infants calibrate their own responses to adults to the *degrees* of good will evidenced by an adult's third-party interactions (Bosevoski & Lee, 2006).

4.2. Empathic distress and empathic concern

Thus we have evidence that infants attend actively to epistemically-relevant behavior among third parties, and use this information to guide their own behavior. Such non-perspectival mapping of the epistemic landscape, however, might be thought to be a case where a clearly *instrumental* motive is at work—it typically pays to be mindful of who is trustworthy. Is there evidence of *intrinsic* motivation to make and act on non-perspectival assessments, as required by criterion (v)? And especially, is there evidence that such motivation is responsive to morally-relevant features of agents or actions, such as the pain or harm others suffer, considered in its own right (vi)?

Studies of infants in the first year of life have found multiple strands of evidence that very young infants seem intrinsically motivated to detect agency in the world around them, attending preferentially to movements that appear to be intentional even in the absence of external incentive or reward. By 6–9 months infants appear to infer agential goals from behavior, even for "agents" that are no more than moving shapes on a screen. They spontaneously engage emotionally with such "agents," reacting not only to success or failure in goal pursuit, but also to the nature of the imputed goal (for a review of infant intention attribution, see Woodward et al., 2009; see Csibra, 2003 for an alternative interpretation of such first-year cognition).

Arguably, there is instrumental advantage to attending to the goals of the agents in one's environment, and surely such information has general epistemic and explanatory value. Consistent with this, Kiley Hamlin and colleagues provide evidence that an infant's modeling of the intentional or narrative structure of third-party interactions conforms to the predictions of Bayesian causal inference (Hamlin, Mahajan, Liberman, & Wynn, 2013; Hamlin, Ullman, Tenenbaum, Goodman, & Baker, 2013). At the same time, however, throughout this period, infants also show a marked *preference* for third parties whose behavior exhibits morally-favored patterns (see Hamlin, 2013, for a summary, and a possible nativist account of this "prior" in infant preferences). Infants as young as 4–6 months follow with interest "morality plays" involving googly-eyed geometric puppets, in which one puppet seems to struggle to climb a hill while a second puppet intervenes either to help or hinder its efforts. Infants with high reliability exhibit favorable attitudes and preference for "pro-social" individuals, and by 8 months of age that they show a preference for a puppet who hinders, rather than helps, a hinderer (Hamlin et al., 2011). By 19–21 months, infants pay attention to whether rewards are distributed equitably among third-parties (Sloane, Baillargeon, & Premack, 2012).

The mechanisms underlying such early cognitive and affective engagement with apparent benefits and harms to third-parties are not known, but we do know that, by 9–10 months, infant response to others' distress is shifting from a focus on their own distress to an orientation toward the distressed individual, and showing signs of caring rather than being upset (Geangu, Benga, Stahl, & Striano, 2011). And as infants' capacity to model others and to take action themselves grows through the second year and beyond, they tend increasingly to attempt to console or help someone in distress (Roth-Hanania, Davidov, & Zhan-Waxler, 2011).

We also know from neuroimaging studies of adults by Jean Decety and others that experiencing a mild shock, imaging oneself receiving such a shock in the future, watching another receive such a shock, and imagining another receiving this shock, activate similar, though not entirely overlapping, elements of the affective system (for recent reviews see Decety, Michalsha, & Kinzler, 2012 and Bernhardt & Singer, 2012). In particular, empathic simulation of another's pain appears to focus primarily upon the aversiveness of pain, rather than its sensory features (Bernhardt & Singer, 2012)—as such, it focuses the empathizing individual upon the intrinsic disvalue of pain as such.

Debate continues over the nature of empathy, the role of cognitive vs. affective components, and the importance of *empathic distress* (affective "resonance") in response to the distress of others, as opposed to *empathic concern* with relieving their distress. However, evidence suggests that, by 12–16 months, infants attempt to decipher the cause of others' distress, and engage spontaneously in helping even in the absence of external reward or encouragement (Warneken & Tomasello, 2006; Zahn-Waxler, Radke-Yarrow, Wagner, & Chapman, 1992). If it were to turn out that all spontaneous empathic responses are, at the most fundamental level, mediated by empathic distress, then empathic motivation would be self-centered and instrumental, and fail to meet criteria (v) and (vi) of moral appropriateness. Immanuel Kant's well-known rejection of the moral relevance of sympathy was based on a psychological theory of this kind (1797/1996). For Hume, by contrast, "extensive sympathy" takes the other's suffering as its direct object. Evidence suggests (see Brown, Nesse, Vinokur, & Smith, 2003) that we experience satisfaction in relieving a stranger's suffering, but this does not show that the effort to relieve the pain was itself instrumentally motivated. On the contrary, Hume argued, such satisfaction depends upon the fact that we had relieving the other's suffering as our goal in the first place (1751/2000). Hence for Hume, unlike Kant, empathic concern or sympathy can be the cornerstone of moral motivation.

Careful investigation of the time-course of empathy in adults indicates that a distress-like response to witnessing another's pain does indeed take place almost immediately, within the first 60–400 ms after exposure. But in normal adults it is soon followed (at 330–420 ms) by cognitive and affective responses associated with taking other perspectives (Thirioux, Mercier, Blanke, & Berthoz, 2014). Akin to other fast affective responses such as fear, empathic distress might serve as an "alarm signal" that something needs attending to, priming the mind and body to reorient (Blair, 2007). But typically, the other, rather than the self, is the *object* of this reorientation—just as fear typically reorients the individual toward the *source* of the threat, rather than toward her own internal state.

Interestingly, there seems to be a range of variation across individuals in the extent to which empathic concern predominates over empathic distress, and their evaluative attitudes or attributions of responsibility can mediate this process (Decety, Echols, & Correll, 2010). Such differences in individual responsiveness may help explain why hypothetical moral scenarios seldom receive unanimous verdicts. For example, experimental subjects with a more pronounced disposition to experience empathic distress (as opposed to empathic concern) were found to be more likely to exhibit emotional distancing and to withdraw rather than assist when costly help is needed (FeldmanHall, Dalgleish, Evans, & Mobbs, 2015; Paciello, Fida, Cerniglia, Tramontano, & Cole, 2013). The notorious "trolley problems" we will consider below involve the prospect of very costly interventions—killing an innocent human being—when one could simply walk away. Great attention has been focused on asymmetries in willingness to intervene across different versions of the trolley problem, but one point is often lost sight of: Why would individuals express willingness to intervene in *any* such cases—why take on such a serious offense to save a group of strangers, given that one could reduce distress in such situations simply via distancing and rationalization? Yet cross-culturally, in hypothetical cases, virtual simulations, and monetized versions involving real effects upon assistance to actual, needy children, a significant majority of people are willing to intervene in a variety of trolley cases (Gold, Colman, & Pulford, 2014; Navarrette et al., 2012; though see also Gold, Pulford, & Colman, 2014). It seems difficult to explain this robust pattern in behavior without positing a widely-distributed tendency to accord intrinsic weight to the fate of strangers.

If a Humean account is right, then we should expect that those deficient in the ability to simulate accurately the affective states of others would tend exhibit a range of difficulties or dysfunctions in their social conduct and personal lives. Diminished ability to simulate *affectively* "what it is like" for others, or "what it would be like" for others or for one's own future self were one to take certain actions, leaves one at a systematic disadvantage in successful navigation of the human landscape. Blair (2006, 2007) has used behavioral, neuroimaging, and clinical evidence to argue that something like this may be the case for individuals on the psychopathy spectrum, who show specific deficits in processing aversive affective information in reinforcement learning. In consequence such individuals fail to learn from past negative experience to accurately project future harms, and thus lose an important source of guidance in avoiding behaviors harmful to themselves as well as others. More than we now realize, psychopathy might be a learning disorder.

Relatedly, we should expect that damage to brain regions or circuits involved in the affective component of simulating perspectives other than one's own current point of view, would tend to undermine both moral and prudential learning and behavior. Ventromedial prefrontal cortex (vmPFC) and frontopolar cortex (FPC) appear to play a key role in affective simulation and evaluation, and early damage to these regions can cause serious impairments in moral learning, while frontopolar dementia can sharply reduce moral sensitivity, social inhibition, and prudential restraint, even without change in the content of normative beliefs (Baez et al., 2014; Mendez, Anderson, & Shapira, 2005).

4.3. Beyond the in-group/out-group distinction?

One impressive sign that neuro-typical children actively construct non-perspectival evaluative models in the social domain, and can be intrinsically motivated thereby, is the evidence that infants normally acquire tacit mastery of the distinction between a moral transgression and the violation of a social rule in the first 3–4 years of life, and treat the former as more serious (Smetana, 1989). Such moral learning appears to be spontaneous—it is found across an array of cultures (Turiel, 2002), and it seems unlikely that most parents will have explicitly reinforced their children for recognizing a distinction between what morality requires and what is required by the adults in authority. Evidence also indicates that children at this age place moral violations in a separate category from questions of personal taste and "pragmatic" norms of convenience, and treat moral concerns as less optional and more serious (Dahl & Kim, 2014).

Moreover, children not only make these discriminations, but they tend to use the same criteria for distinguishing the moral from other norms, and for explaining its relative importance for action. For example, in line with the criteria we have used here, preschoolers see moral requirements as authority- and convention-independent, general in scope, concerned with harm or benefit to others, and more serious to violate (Turiel, 2002). This discriminative competence is manifest motivationally as well, since children at this age will resist taking actions that inflict harm upon others simply in virtue of a new rule imposed by an authority figure. And they appear to be intrinsically motivated to enforce moral norms in ways they are not motivated to enforce demands of conventional authority, and will spontaneously console victims of moral transgressions even when the victims themselves do not show physical distress (Vaish, Carpenter, & Tomasello, 2009; Vaish, Missana, & Tomasello, 2011).

As a sophisticated implicit "model" of the moral, this competency can be used to innovate as well. Some children, starting around age 6, become "independent vegetarians"—resisting eating meat even though this has been the norm in their family and social context (Hussar & Harris, 2009). More so than children who grow up in vegetarian families, "independent vegetarians" cite reducing the suffering of animals as the ground for refraining from eating meat. Thus, these children seem capable of appreciating, and being intrinsically motivated to act upon, morally-important concerns grounded in the interests of those wholly outside their own "in-g roup"—and in the face of life-long evidence that "people like me" (e.g., members of one's own family) do *not* act this way.

Such a spontaneous willingness to take into account the suffering of strangers—even members other species—at some expense to the self might seem incompatible with an evolution of the human psyche based upon genetic relatedness and small social groups. A significant body of psychological evidence does attest to the tendencies of humans, from early months onward, to favor those perceived as similar to themselves (see for example Hamlin, Mahajan, et al., 2013; Hamlin, Ullman, et al., 2013). However, various behavioral studies have also found that generosity or cooperation in one-off interactions with strangers is the *default* response of most individuals, while self-interested behavior emerges only with second and third thoughts (Rand, Greene, & Nowak, 2012). In a carefully-run experiment in using actual money and actual (small but real) shocks, Molly Crockett and colleagues found that adults on the whole were willing to pay a higher price to prevent a shock to a stranger than to themselves (Crockett, Kurth-Nelson, Siegel, Dayan, & Dolan, 2014). And Ernst Fehr and colleagues, also using real money, have found that people were willing to pay to punish someone who was being unfair at the expense of a third party (Fehr & Fischbacher, 2004). At the neural level, fMRI imaging has been used to argue that empathic distress responses remain intact toward stigmatized victims (Decety et al., 2010), and that, in many subjects, achieving cooperation in Prisoner's Dilemma games induces greater reward activity than even self-advantageous non-cooperative outcomes (Rilling et al., 2002).

Why might this be? The ethnographic record indicates that many hunter-gatherer bands practice exogamy (i.e., marriage takes place with individuals outside the group), and field observations suggest that individuals not infrequently shift from one band to another, whether as an outcome of warfare, social exclusion, migration, decay of group size, or attempts to secure better prospects (Marlowe, 2003). Hunter-gatherer social and trading networks can be extensive, and "functional social proximity" through shared activity or exchange is often more influential than actual "kin proximity" in shaping behavior (Apicella et al., 2012). Engaging effectively in these more flexible and less directly group-centric ways of life puts a premium on the ability to "size up" and interact with strangers and other groups in light of assessments of *general*, *modal*, *morally-relevant* characteristics such as cooperativeness, trustworthiness, competence, knowledgeability, aggressiveness, or tendencies to help or share. Group selection and sexual selection likewise involve such traits, and have been identified as processes that can favor the emergence of significantly altruistic behavior that cuts across genetic relatedness (Nesse, 2007).

4.4. Explicit and implicit bias

Still, an inevitable problem for such a learning-based architecture is that it will function only as well at providing an objective representation of social reality as the experiential "sample" to which it is exposed is itself broadly representative of that reality. And unfortunately, human social experience tends to be unrepresentative in many ways—one is born into a particular family in a particular society, and into a particular social and cultural location within society, skewing the sample from which one learns from early on. This appeal to skewed learning to explain bias is distinct from explanations involving purportedly atavistic "us/them" or "tribal" attitudes. In this light, it is notable that implicit bias does not appear to be fully established until approximately the age of six (Dunham, Baron, & Banaji, 2008). That is, even though younger children tend to be essentializing (Gelman, 2009), and to favor those perceived as like themselves (Hamlin, Mahajan, et al., 2013; Hamlin, Ullman, et al., 2013), they appear to need to *learn* which social groups have which "essential" features, who in the wide world is alike with or different from themselves, and what the social hierarchy is. Unlike an "us/them" bias, moreover, implicit bias tends to favor those of higher social standing, and thus often does not favor one's group if one belongs to a stigmatized subpopulation. Further, dominant-culture negative stereotypes of a marginalized group are often found in those belonging to this group (Dunham et al., 2008). It would seem that children living in the same society tend to learn similar *implicit casual/evaluative models* of that society, reflective of the underlying social hierarchy and power relations, even if these children also differ in a host of other group-related attitudes.

If much social prejudice is learned in these ways, then it should in principle be possible to *unlearn* it by changing people's experiential sample. Perhaps the most relevant literature here is research on effective ways of overcoming implicit bias. An extensive review of this literature supports the hypothesis that the most effective ways of overcoming such bias involve "contact" processes that go beyond mere exposure of groups to one another, and include activities in which individuals from different groups co-participate in activities that have a common goal, draw upon the contributions of each, and involve taking the perspective of others (Blair, 2002; Dasgupta & Rivera, 2008; Pettigrew, 1998; Pettigrew & Tropp, 2006). Changing people's experiential sample in ways that provide more extensive and representative feedback seems to be the most reliable way to reduce implicit bias—and can have this effect even in those individuals who continue to express explicit prejudice.

A historic "natural experiment" illustrates this on the social scale. Explicit and implicit bias against homosexuals and homosexual relations has been strong across a wide array of societies and religious traditions for centuries, and some psychologists had suggested that such bias arises from "automatic" disgust reactions rooted in natural selection. Yet disgust and related "basic" mechanisms like gustatory taste appear to be considerably more complex and construal-dependent than "automatic" accounts would suggest (Tybur, Kursban, Lieberman, & DeScioli, 2013). And "core disgust" may have less role in moral judgment than previously thought (Landy & Goodwin, 2015; Yu et al., 2013). In recent decades, gay individuals' greater openness about their sexual orientation—which began a great personal expense in the face of strong social sanctions—has meant that the heterosexual population is now *aware* that it is engaged in meaningful, shared activity with gay individuals in virtually every area of life. In consequence, attitudes toward questions like gay marriage, and measures of implicit bias toward gays, have undergone a dramatic change—especially among the young, who have lived their entire lives in an atmosphere of more open recognition of sexual orientation (Westgate, Riskind, & Nosek, 2015). That a centuries-old form of prejudice—an "us" vs. "them" bias thought to be deeply rooted in our psyches—could undergo such rapid change is a tribute to the power of learning, even from "old" evidence that has been re-categorized, if given a chance.

Here we have largely emphasized implicit and individual learning processes, but not because social and self-consciously reflective processes are unimportant in moral learning—on the contrary, humans are social, discursive, ratiocinating beings, and human moral thought and practice profoundly reflect this. Rather, our focus arises from the need to contrast a moral learning approach as clearly as possible with "hard-wired" or nativist accounts, on the one hand, or external "socialization" into morality, on the other. Innate modules and external socialization could well play a role in moral learning, but, we have argued, their operation may *presuppose* children's fundamental abilities to be attentive to, discriminating about, and internally motivated by, morally-relevant features of the world—and implicitly competent at distinguishing these from matters of authority, convention, taste, or convenience.

5. Some applications of a moral learning perspective

5.1. The nature of intuition and intuitions

Our challenge has been to develop a picture of how individuals can acquire through experience evaluative models of situations, agents, actions, and practices that meet such moral criteria as impartiality, generality, authority-independence, a non-instrumental concern with benefits and harms to those affected, and thought- and action-guidingness. These evaluative models are not assumed to belong to a dedicated "moral faculty," but rather to be components of the more comprehensive causal/evaluative models we form of the world.

As currently understood, such mental models take the form of many-layered neural networks with large numbers of parameters, constantly engaged in the projection of novel expectations and subsequent updating of the model's parameters via the "back-propagation" of any discrepancies with actual outcomes. Neural networks embody a hierarchy of abstraction and generality, and we have limited ability to introspect their workings beyond the upper-most levels where they connect with conscious experience, language, and thought. Much, perhaps most, of the evaluative assessment they furnish is thus likely to be "intuitive"—in essentially the same way that native speakers' judgments of whether a sentence is "grammatical" or "odd" are largely "intuitive"—even though they may carry highly nuanced information about an underlying model of syntactic regularities or semantic features.

Intuition in this sense is a summative or holistic conscious "mode of presentation" of the outputs of a complex underlying model. It appears to operate across such diverse domains as perceptual image recognition, athletic or game-playing skill, artisanal or professional expertise, artistic creativity, and "social intelligence"—in all of these domains, a given interpretation, situation, or action may "feel" right or wrong, "look" shaky, "sound" promising, or "seem" urgent. Similarly, recent research suggests that the insula receives many streams of input from different systems of the body and mind, as well as sensory information about the external environment, and yields a summative "intuitive" sense of one's mental or physical condition, or of how well one's life is going (Craig, 2009). Even though intuitions are typically *experienced* as spontaneous and non-deliberative, they in fact can represent the output of on-going, information-intensive, high-dimensional calculations of brain networks capable of millions of neural firings per second. In all these examples, intuitions are not "automatic," inflexible, "preset," stimulus-driven, or "informationally-encapsulated" responses. Moreover, given their deep experiential base, such intuitions can also resist conscious cognitive pressure or rationalization—even after one has constructed a plausible-sounding rationale, an action might still *feel* wrong, though we cannot say why.

Intuition in this sense is also not confined to momentary experiences. An entire suite of thought and action—a fluent conversation, a gracious social response to an awkward situation, a novel musical improvisation—can be guided "intuitively," without conscious reflection. Such intuitive action-guidance is made possible via implicit *model-based control*, which draws upon the generative character of the underlying causal/evaluative models and their capacity to update in response to experience in real time.

If people are largely similar in their basic psychology and typical hopes and fears, and if life presents us with similar problems and opportunities, then learning might account for the fact that people tend to have broadly similar intuitions in a number of core areas in morality: the need to limit personal aggression or defend oneself against it, to build mutual trust and willing cooperation, to stabilize possessions while facilitating beneficial exchanges, to share jointly-created or scarce resources, and to provide for the young and vulnerable (Brown, 2004; Henrich et al., 2004; Rai & Fiske, 2011).

5.2. Dual-process accounts of moral judgment: first generation

This "deep learning" picture of intuition offers a different approach to the nature and potential normative authority of moral intuition from predominant contemporary "dual-process" accounts of morality. Let's explore this contrast in two stages, looking first at the original dual-process accounts (Greene & Haidt, 2002; Haidt, 2001), and then considering a "second generation" of dual-process accounts that has emerged only recently (Crockett, 2013; Cushman, 2013). As our main case study, we will discuss how these different accounts tackle the well-known "trolley problems," which afford the best-studied (some would say, most *over*-studied) examples of moral judgment in the literature.

On predominant dual-process accounts, moral intuitions are attributed to the action of "System 1," an "evolutionarily ancient" affective system we inherit from our animal ancestors. While different characterizations are used by different authors, System 1 is typically described as fast, "automatic," "emotional," "domain-specific," and as based upon "heuristics" that manifest "little understanding of logic or statistics" or upon "gut feelings," "push buttons," or "presets" (Greene, 2013; Greene & Haidt, 2002; Greene et al., 2009; Haidt, 2007; Kahneman, 2011; Prinz, 2004). The automatic responses of System 1 are hypothesized to have evolved because they are of a kind generally useful in the environments in which most adaptation of humans and their ancestors occurred, and because they require little representational or computational complexity to operate in the brain. Neural simplicity makes for speed, but at the cost of flexibility, and as a result System 1 intuitions may be inappropriate in unusual circumstances or in some modern social settings. Evolutionarily newer "System 2," by contrast, is seen as the system of reflective thought and decision-making. It is slower, "cognitive," "domain-general," and calculative—capable of delivering reasoned judgments based upon logical and statistical inference, but also capable of after-the-fact rationalizations or "confabulation" when System 1 issues in responses before System 2 can come into play (Haidt, 2001).

A well-known example used to illustrate this contrast between intuitive and reasoned judgment is Jonathan Haidt's case of Julie & Mark:

> ***Julie & Mark*** *are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried making love. At the very least it would be a new experience for each of them. Julie was already taking birth control pills, but Mark uses a condom too, just to be safe. They both enjoy making love, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that? Was OK for them to make love?*
>
> [Haidt, 2001]

Haidt reports that a majority of his subjects say that this was *not* "OK," though when pressed for reasons they often provide rationales that do not, it seems, apply in this case—that incest leads to birth defects, psychological and familial trauma, and so on. Further pressed, subjects may insist upon their original negative judgment even while admitting they can offer no convincing rationale for it, a phenomenon Haidt and colleagues call "moral dumbfounding" (Haidt, 2001).

Haidt offers an evolutionary explanation: for early humans living in small groups, incest was indeed a serious risk to reproductive fitness, and so an automatic *disgust* response evolved to make the prospect of incestuous relations aversive (Haidt, 2001). This System 1 "flash" of disgust is insensitive to further information, so Julie and Mark's action continues to "just seem wrong," despite the favorable outcome.

The dual-process approach challenged a number of long-standing assumptions about the role of reason vs. emotion in moral judgments. Of special importance was the neuroimaging work of Joshua Greene and colleagues (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001), who studied brain responses to contemplating hypothetical moral dilemmas, most notably, "trolley problems." Trolley problems were first introduced into the philosophical literature by Foot (1967/1978) and Thomson (1976), and are now widely used in psychological testing of moral judgment as well. They owe their influence to a striking asymmetry shown in typical intuitive moral judgments regarding two canonical scenarios with seemingly similar net outcomes, which we will call **Switch** and **Footbridge**.

> ***Switch****. A runaway trolley is speeding down the track, its driver slumped over the controls, apparently unconscious. Ahead on the tracks are five workers, who do not see the trolley coming, and who soon will be struck and killed. You are standing next to a lever that operates a switch lying between the trolley and the workers. Pushing this lever would send the trolley onto a sidetrack. That would save the five workers, but there is a single worker on the sidetrack, who will be struck and killed. Should you push the lever to send the trolley down the sidetrack?*

In a typical sample, a strong majority will say *yes*, and give as their reason the need to minimize the loss of life in such emergencies. But now consider:

> ***Footbridge.*** *A runaway trolley is speeding down the track, its driver slumped over the controls, apparently unconscious. Ahead on the tracks are five workers, who do not see the trolley coming, and who soon will be struck and killed. You are standing on a footbridge over the track, next to a very large man. This man's weight is sufficient to stop the trolley, though your own is not. If he were to fall into the path of the trolley, that would bring it to a halt before hitting the five workers, saving their lives but killing him. Should you push the man off the footbridge into the path of the trolley?*

This time, in a typical sample, a strong majority will say *no*. When queried, people often explain this verdict in terms of the impermissibility of deliberately *using* a person in this way, rather than merely harming a person as an unavoidable side-effect, as in Switch.

So, the next scenario to consider is:

> ***Loop.*** *A runaway trolley is speeding down the track, its driver slumped over the controls, apparently unconscious. Ahead on the tracks are five workers, who do not see the trolley coming, and who soon will be struck and killed. You are standing next to a lever that operates a switch lying between the trolley and the workers, and pushing the lever would send the trolley onto a side loop. This loop rejoins the main track just before the location of the workers. However, on that loop stands a single large worker, who would be struck and killed if you switched the trolley. His weight would bring the trolley to a halt before the loop rejoins the main track, saving the five workers. Should you push the lever to send the trolley onto the side loop?*

In Loop, the man on the side-track, like the man on the footbridge, is deliberately used as a means to stop the trolley. Yet a strong majority will give the same answer to Loop as Switch—*yes*, one should push the lever to send the trolley down the sidetrack.

Now, when asked to explain why their verdicts differ in Loop and Footbridge, most cannot articulate a clear rationale. Despite this, for the majority, the intuitive verdicts in all three cases remain fairly strong, and the asymmetry between Switch and Loop, on the one hand, and Footbridge, on the other, persists. This, then, appears to be another case of "moral dumbfounding."

What *does* explain this resilient pattern of intuitive judgments?[1] Some philosophers offer a normative response in terms of more complex underlying moral principles (e.g., Kamm, 2007), but the neuroimaging results of Greene and colleagues suggested that looking for ever more intricate principles was looking in the wrong place (Greene et al., 2001). They had found that contemplating Footbridge-like scenarios excited relatively greater activation in brain areas linked to emotional responses, while contemplating scenarios involving more indirect ways of bringing harm to the victim, as in Switch or Loop, excited relatively greater activation in areas concerned with working memory and controlled cognition. This suggested that the asymmetry was attributable to dual-processing. In Footbridge, a fast, strong, negative, "emotional" System 1 response

[1] We should keep in mind, of course, that the same pattern is not shown by all subjects—there are non-trivial variations that also need explaining, as we will discuss briefly below.

to the "personal" harm of pushing the man to his death pre-empts or preponderates over a slower, weaker, favorable, "rational" System 2 response to saving the five workers' lives. In Switch and Loop, by contrast, the more "impersonal" character of the harm done to the lone victim means that no strong emotional response is triggered in System 1, and so System 2's reasoned, harm-minimizing response preponderates.[2] Greene and colleagues subsequently made this interpretation more precise through a series of experiments that strongly suggested that it is the direct use of one's own muscular force in Footbridge, rather than "personal" harming as such, that is driving the asymmetry—a parameter that seemed even more likely to reflect an "automatic" emotional response rather than a credible underlying moral principle (Greene et al., 2009). Other experiments showed that increasing cognitive load selectively diminishes cost-benefit-oriented judgment in trolley problems, while cognitive priming and interfering with visualization increases it (Amit & Greene, 2012; Greene, Morelli, Lowenberg, Nystrom, & Cohen, 2008). In a recent formulation, Greene attributes the System 1 response in Footbridge to an emotional "alarm" in an evolved "myopic module" that cannot see ahead to the greater number of lives that would be saved by pushing the man (Greene, 2013). In Switch and Loop, by contrast, harm is done without direct exertion of muscular force on the victim, and so does not trigger the same "myopic" System 1 response as Footbridge.

However, as we have seen in previous sections, an impressive body of evidence in neuroscience and behavioral psychology suggests that the affect and reward system we inherit from our foraging ancestors does not resemble these characterizations of System 1—instead, it seems designed to learn complex statistical relationships, subserving the building of abstract casual/evaluative models that guide attention, perception, and action along expected-value maximizing lines. Indeed, the long-standing idea of dividing the brain and mind into "rational" and "emotional" regions or functions is losing currency, since the two are increasingly being found to be inextricably intertwined—one of the key claims of Hume's *Treatise* (1738/1978). Reasoning makes heavy use of the affective system in forming the decision weights that guide deliberation and choice, and affective responses evolve dynamically in response to cognitive construal and appraisal (Nesse & Ellsworth, 2009; Pessoa, 2008).

Moral emotion and judgment in particular appear to be grounded in large-scale, functionally-integrated, domain-general brain networks that recruit information widely and overlap extensively with such functions as episodic and semantic memory, theory of mind, hypothetical mental "construction" and simulation, and planning (Buckner et al., 2008; Hassabis & Maguire, 2009; Moll & Oliveira-Souza, 2007; Moll, Zahn, Oliveira-Souza, Krueger, & Grafman, 2005; Shenhav & Greene, 2010). As the evidence stands, there seems to be nothing like a domain-specific "moral module" or set of moral "push buttons."

5.3. Dual-process accounts of moral judgment: second generation

However, there has recently emerged a new dual-process approach to the asymmetry between Footbridge and Switch that does not depend upon problematic assumptions about "reason" vs. "emotion." Moreover, it focuses precisely on the role of *learning* in moral psychology.

In the past two decades, learning theorists have delineated a distinction between "model-free" and "model-based" reinforcement learning, and shown its importance in explaining a variety of behavioral phenomena (Daw & Doya, 2006; Sutton & Barto, 1999). Recently, Molly Crockett, Fiery Cushman, and others have shown how this distinction can be used to account for otherwise puzzling patterns in moral judgment, including trolley problems (Crockett, 2013; Cushman, 2013).

Thus far, we have largely focused on model-based learning and control—how do model-free learning and control differ? In *model-free learning*, prediction-error learning is used, not to build causal/evaluative models that generate and evaluate multiple options during choice, but to develop "cached" expectation values for individual actions, or, more properly, for ⟨situation, action⟩ pairs. This is done by telescoping into a single expectation value assigned to an action in a situation all the estimated reward information about future value paths that could issue from that action. In *model-free control*, a simple comparison of cached value estimates makes it possible to select the action of those available with the higher expected value, without needing to "look ahead" to future consequences and calculate back to the alternative choices—that information is already "contained" in the cached expected value. In an environment with stable choices and rewards, model-free control that has been fully trained up can achieve optimal choice with fewer on-line computational demands than model-based control (Le et al., 2012). Thus, in such a static environment, control will tend to shift from model-based to model-free control as an animal becomes "overtrained." However, if the available actions or reward values subsequently change, then, although the cached values for ⟨situation, action⟩ pairs will no longer be accurate, the "overtrained" animal will initially continue to follow these cached values and choose sub-optimally, since learning new model-free values requires gradual retraining. In model-based control, by contrast, new information can enter directly into the model, allowing an animal to adapt rapidly even to large changes in the environment or reward values.

The classic demonstration of model-free vs. model-based control of behavior is thus "devaluation," in which a previously-valued stimulus is experimentally devalued, yet the behavior continues as before (Dickinson, Balleine, Watt, Gonzalez, & Boakes, 1995). For example, since I always flip the light switch when starting down the cellar stairs, and have always been rewarded by being able to see where I'm going, I have acquired a cached positive expected value for ⟨starting down the cellar stairs, flip the light switch⟩. This saves me mentally projecting possible future outcomes of a decision that always yields the same, favorable result. However, this also means that, even if I am heading down into the cellar for the express purpose of replacing the burned-out lightbulb, I will still flip the switch as I start down the stairs. My overall casual/evaluative model of the situation, which includes the recently burned-out bulb, the devaluation of flipping the switch in the current circumstance, the disvalue of being unable to see my way in the cellar, and the existence in the cupboard of a supply of replacement bulbs, enables me to imaginatively project and positively evaluate the action sequence of stopping by the cupboard, picking up a replacement bulb, and heading down into the cellar without flipping the switch—a suite of behaviors I might never before have performed. Yet this model-based plan doesn't stop the stimulus of starting down the stairs from triggering the cached model-free reward value of flipping the switch—so I do so, pointlessly.

Model-based and model-free learning and control thus can operate in tandem, giving us a new way of conceiving dual-processing: two potentially "rational" forms of learning and control are shaping my behavior, and the impact of the proximate stimulus of starting down the cellar stairs can trigger a well-established cached ⟨situation, action⟩ value for flipping the switch even though my more comprehensive model has devalued this act.

[2] In the literature, resistance to pushing in Footbridge is called a "deontological" judgment (since it seems to exclude mere cost/benefit calculation), while pushing the man is called "utilitarian" (since it seems to rely exclusively upon cost/benefit calculation). However, these are not entirely happy labels, for reasons that will be discussed below, so I will avoid them here.

For obvious reasons, model-free processing is associated with habitual, stimulus-response behavior, and model-based processing with goal-directed, flexible, and adaptive action.

Cushman (2013) has suggested that the intuitive rejection of pushing the large man off the footbridge might be an instance of model-free learning and control at work. For example, my personal history of direct experience and indirect observation of the disvalue of violently shoving someone in situations other than self-defense leaves me with a cached, strongly negative value for acts of this kind. The stimulus of imaginatively contemplating pushing the man off the bridge triggers this cached disvalue, which is not sensitive to the highly unusual "revaluation" that connects this act to the saving of five workers' lives. Model-free processing (think of this as System 1) thus results in a powerful "intuitive" sense that I should not push, which does not simply yield in the face of the more abstract cost-benefit analysis issuing from my overall causal/evaluative model of the situation (think of this as System 2). In Switch and Loop, by contrast, since I do not have a similar history of direct or indirect negative experience with pushing levers that activate a switch, no cached model-free negative value is triggered by the imagined stimulus of addressing myself to the lever, so there is no similar competition with model-based control that looks ahead to the positive remoter consequences of pulling the lever.

Cushman's (2013) account thus offers a compelling explanation of the trolley asymmetries that is grounded in current thinking about learning and control. It also coheres well with much of the collateral evidence for the original dual-processing account, e.g., concerning the effects of cognitive load, cognitive priming, disrupting visualization, and varying degrees of directness in exerting muscular force upon the victim (see Amit & Greene, 2012; Greene et al., 2008, 2009). And Cushman adds to this a new body of data indicating that people are averse to performing actions that superficially resemble normally harmful actions, even when it is clear that no harm is actually involved—e.g., hammering a person's trouser leg when it is known to contain only a plastic pipe or pulling the trigger of heavy fake gun pointed at the experimenter's face (Cushman, Gray, Gaffey, & Mendes, 2012).

5.4. Model-based trolleyology

The model-free/model-based dual-process explanation of trolley asymmetries is highly promising. However, this approach does make the persistence and force of these asymmetries a bit puzzling. Model-free values are slow to respond to revaluations, but they *do* respond. If I can't find a replacement bulb for the cellar light, so that it remains non-functional for several weeks, then a number of frustrated attempts will eventually lead me to lose the habit of flipping the switch as I go down the steps, and perhaps acquire the habit of picking up a flashlight before heading to the cellar. Philosophers and psychologists who have been contemplating trolley problems for years, even running various kinds of simulations, still seem to share the sense that pushing in Footbridge is markedly more problematic morally than pulling the lever in Switch or Loop, even as they struggle to explain why.

A model-based explanation of the asymmetries, if grounded in robust causal/evaluative features that differentiate the cases, could account for such persistence. We already have seen reasons for thinking that underlying causal/evaluative models can manifest themselves in immediate "intuitions" as well as more deliberative judgments. And since we typically lack direct insight into the deeply-layered underlying models, even robustly-based distinctions may be difficult to fathom based upon surface features. To test these ideas, we need to fill out our sample of cases in a way that is not customary in the trolley-problem literature. Consider the 2 × 2 matrix, Fig. 1.

Can we find cases that occupy quadrants *X* and *Y*? Let's start with *X*:

> ***Bus.*** *You are visiting a city where there have recently been terrorist suicide bombings. The terrorists target crowded buses or subway cars. To prevent anyone stopping them, they run up at the last moment when the bus or subway doors are closing, triggering their bomb as they enter. You are on a crowded bus at rush hour, just getting off at your stop. Next to you a large man is also getting off, and the doors are about to close behind the two of you. You spot a man with an overcoat rushing at the doors, aiming to enter just behind the exiting man. Under his coat you see bombs strapped to his chest, and his finger is on a trigger. If you were to push the large man hard in the direction of the approaching man, they both would fall onto the sidewalk, where the bomb would explode, killing both. You would have fallen back onto the bus, and the closing doors would protect you and the other occupants of the bus from the bomb. Alternatively, you could continue exiting the bus, and you and the large man would be on sidewalk, protected by the closing doors, as the bomb goes off inside the bus, killing the terrorist and five passengers. Either way, then, you will not be hurt. Should you push the large man onto the bomber?*

Bus differs from Footbridge in far too many ways to constitute a "minimal test case." But an extended history of looking at minimal test cases has left the asymmetry still quite opaque to most of us, so perhaps it is worthwhile trying another methodology.

Using hand-held remote devices, students in my introductory ethics classes are able to respond rapidly and anonymously to hypothetical scenarios like Switch, Footbridge, Loop, and Bus. Moreover, I am able to pose a series of subsequent questions, and also to re-sample their intuitive judgments over time. This is far from any approved experimental protocol, but I am far from being an experimenter, and I thought experimenters (as well as theorists and philosophers) might find it worthwhile to see what this informal method of questioning and sampling yields.

When my students are given Switch, Footbridge, and Loop, their responses follow the well-known, asymmetric patterns. When I pose Bus at a subsequent class meeting, a strong majority answers *yes*, one should push the large man onto the bomber (most recently, 70% answered *yes* and 28% answered *no*; n = 42).[3] Somehow, in this scenario, any model-free aversion to use of one's muscles to forcefully push a man to a violent death does not show up as a preponderant, immediate rejection of the act.

Now consider a candidate for the *Y* quadrant of our matrix:

> **Beckon.** *A runaway trolley is speeding down the track, its driver slumped over the controls, apparently unconscious. Ahead on the tracks are five workers, who do not see the trolley coming, and who soon will be struck and killed. You are standing at some distance from the track, with no ability to turn the train or warn the men. A large man, whose weight is sufficient to stop the trolley,*

[3] Percentages may fail to add to 100%, because students may accidentally push a button that does not correspond to a valid answer. In the text I will give student percentages from the most recent time the course was taught. There has been no notable variation in the pattern of student responses from year to year, even though in some cases the course was taught for first- and second-year students with no background in philosophy, while in other cases it has been taught for upper-level students who have already taken some philosophy. The wording of the questions has changed somewhat from year to year, without notable effect on student voting. Students were told, truthfully, that they would receive 1 point for every class session in which they entered a response to at least one question, but that individual responses would never be consulted. Normally, sampling lasted about 10–30 s after the question was posed, at which time the result of the polling was displayed as a bar graph for all to see. The fact that several scenarios or questions might be posed in the same class session might have had the effect of implicitly encouraging students to seek to give more "consistent" responses, though they were given no instruction concerning which cases were most similar to one another in moral terms. Approximately one-quarter of the students on a given day were women.

	Intervention harming one to save five should *not* be done, according to most subjects	**Intervention harming one to save five *should* be done, according to most subjects**
Use of direct muscular force to inflict harm	Footbridge	*X*
No use of direct muscular force to inflict harm	*Y*	Switch, Loop

Fig. 1. Trolley problem "intuitive" judgments.

> *is standing on the other side of the track, facing in your direction. He is unable to see the oncoming trolley owing to a traffic signal box that blocks his view up the track. If you would beckon to him [I pantomime an encouraging beckoning gesture], he would step forward onto the track, and be immediately struck and killed. This would halt the trolley and save the five workers. Should you beckon to the large man?*

When given the Beckon scenario, a fairly strong majority of my introductory students will answer *no* (most recently, 58% *no* vs. 42% *yes*; n = 38; for comparison, in Footbridge 59% answered *no* vs. 41% answering *yes*; n = 44).

Might there be a cached, model-free strong negative valuation of making a gesture that lures a distant person into mortal danger? Consider for comparison another case I offer to my students:

> **Wave.** *A runaway trolley is speeding down the track, its driver slumped over the controls, apparently unconscious. Ahead on the tracks are five workers, who do not see the trolley coming, and who soon will be struck and killed. A wall prevents them from moving to their left to avoid the trolley, but there is space to their right. You are standing at some distance from the track, with no ability to turn the train. The workers are facing in your direction, and if you were to wave to their right with your arms [I pantomime an encouraging waving gesture], the five workers on the track would step off and escape injury. However, a single worker who is closer to you and standing to the left of the track, and who also does not see the trolley, will see you wave, and he will step onto the track, and immediately be hit and killed. Should you wave to the workers?*

In Wave, a strong majority answer *yes* (most recently, 87% *yes* vs. 13% *no*; n = 38; for comparison, in Switch, 85% *yes* vs. 13% *no*; n = 45). In other words, Beckon and Wave exhibit the same intuitive asymmetry as Footbridge and Switch, yet the immediate action (an encouraging hand gesture) is virtually the same in each, and there is no personal violence in either. To be sure, Beckon does use a man as a means, like Footbridge. But it does so at a distance, like Loop. Why, then, is beckoning disapproved by the majority while switching the trolley in Loop is approved by the majority (most recently, 90% answered *yes* to switching the trolley in Loop, 7% answered *no*; n = 41)?

An advantage of using anonymous sampling of intuitive reactions in a classroom setting is that it permits exploration of further dimensions of the questions with the same students, who moreover have shared the experience of seeing the overall results of previous polling. Nothing is hidden. This approach loses a great deal of the experimental control typical of moral psychology experiments, and is no substitute for them, but it does provide a different kind of data, since students in effect become co-investigators into their own moral intuitions. This, they tell me, gives them a strong motivation to respond to questions in earnest—they have a keen curiosity about finding out what people, themselves included, "really think," and would like very much to understand this thinking better (for some studies supporting the idea of electronic polling as a useful source of information, see Stowell & Nelson, 2007 and Kennedy & Cutts, 2005).

Pushing our shared exploration further, I ask the students, "If you were to learn that your roommate or friend had pulled the lever in Switch [or pushed the man in Footbridge, etc.], would you trust him or her the same, more, or less?" This might seem a very vague and impressionistic question, but students readily responded. In the case of pulling the lever in Switch, students as a whole expressed no net change in level of trust (in the most recent year, 26% answered *more*, 56% answered *the same*, and 18% answered *less*; n = 39). When asked the same question about pushing the man in Footbridge, students as a whole expressed a sharp loss of trust (most recently, 0% answered *more*, 21% answered *the same*, and 78% answered *less*; n = 33). If I subsequently pose the same questions about Wave and Beckon, and the same asymmetry emerges (most recently, in Wave, 20% answered *more*, 68% answered *the same*, and 13% answered *less*; n = 40; in Beckon, 0% answered *more*, 21% answered *the same*, and 79% answered *less*; n = 42). Thus, despite the differences in the *concrete* kinds of acts performed in Switch vs. Footbridge as opposed to Wave vs. Beckon, students might be modeling an *abstract* similarity in the two pairs of cases. A strong majority associate performance of the lethal act in Footbridge and Beckon with *untrustworthiness*, but not in the case of the lethal act in Switch or Wave. I'm tempted to say, that given my students' implicit causal/evaluative model of the cases, learning the fact that someone pushed in Footbridge or beckoned in Beckon supported an inverse inference attributing to this person a less trustworthy motivational structure than before.

Were the students picking up on anything real, or was this a generic effect of distrusting a person who would perform an act they judged wrong? Yet in the class in question, 41% of students had earlier given the answer that they thought they *should* push the man in Footbridge, while *no* students indicated increased trust in a person who performed this act, and 78% indicated increased distrust. Similarly, 42% had answered that they *should* beckon in Beckon, but *no* students indicated increased trust in someone who has beckoned, and 79% indicated increased distrust. See Table 1 for a summary (the most frequent response is **underlined**).

If we turn to the experimental literature, it seems there is evidence supporting the students' distrust response. Kahane, Everett, Earp, Farias, and Savulescu (2015) found that disposition to push the man in Footbridge correlated with a higher score on a psychopathy scale and a lower score on an altruism scale. Bartels and Pizarro (2011) and Gao and Tang (2013) also found a positive correlation between giving the "push" answer in Footbridge and higher score on a psychopathy scale. A number of studies have found decreased levels of empathy, harm-aversion, and perspective-taking in those giving push-like responses in Footbridge-style scenarios (Conway & Gawronski, 2013; Gleichgerrcht & Young, 2013; Weich et al., 2013). And Duke and Begue (2015) found that rate of giving the push verdict in

Table 1
Trolley problems and trustworthiness.

	Would trust more	Would trust the same	Would trust less
What if you learned your roommate or friend had . . .			
Pushed lever in Switch? (n = 39)	26%	**56%**	18%
Pushed man Footbridge? (n = 33)	0%	21%	**78%**
Waved in Wave? (n = 40)	20%	**68%**	13%
Beckoned in Beckon? (n = 42)	0%	21%	**79%**
Pushed man in Bus? (n = 44)	**41%**	39%	20%

Footbridge-like scenarios increased with increasing blood alcohol concentration, an effect they attributed to alcohol's impairment of social cognition and empathy, combined with its reduction of executive function. At a more abstract level, Uhlmann, Zhu, and Tannenbaum (2013) found evidence that an implicit imputation of personality characteristics mediated judgments in sacrificial dilemmas. And in a similar vein, Sripada (2012) used structural equations analysis to develop evidence that an implicit imputation of underlying personality characteristics (a "deep self") mediated the well-known asymmetry in judgments of intentionality in the "Knobe Effect" (Knobe, 2010).

What happens in subjects' or students' minds when they are posed hypothetical moral scenarios? In effect, they are being asked to do what their default mode does for a living—to project and evaluate an imagined situation and array of possible responses, drawing upon cognitive and affective resources to simulate the consequences and the states of mind of those involved. Neuroimaging studies indicate that such simulation takes place spontaneously when hearing hypothetical scenarios, whether or not the subject is specifically instructed to place *herself* in the agential role (Decety et al., 2012). My speculation is that when students attempt to mentally simulate pushing the man off the footbridge, they typically encounter significant affective resistance—it is harder to "see themselves" pushing the man in Footbridge as opposed to pulling the lever in Switch or Loop. This effect may arise in part from a rapid empathic distress response, discussed above, which is experienced by neuro-typical people when they imagine taking certain kinds of harmful actions (Blair, 2007). But this distress reaction is usually not the whole of the empathic response, which, as we saw, tends to evolve quickly to include perspective change and affective concern (Thirioux et al., 2014). My speculation is that, in mentally simulating moral scenarios, we view them from multiple perspectives in space and time—implicitly asking how the act would feel, what it would look like to others, what one would feel afterwards, how one would explain oneself, and so on. All very rapidly—this sort of personal and social "mental construction" and evaluation is, after all, one of the main tasks our mind was built to perform, and perform reasonably well (Buckner et al., 2008; Hassabis & Maguire, 2009; Seligman et al., 2013, 2016). I would speculate that, in the case of Footbridge, simulating pushing the man "feels" aversive, "looks" surprising and callous, "seems" as if it would be hard to defend; in Switch or Loop, I further speculate, simulating pulling the lever does not generate these negatively-valenced "seemings"—indeed, simulating standing by and doing nothing while five die rather than one is likely to "feel" callous. One can imagine, in Switch, bystanders shouting, "Quick—throw the switch!" Can one imagine them shouting in Footbridge, "Quick—throw the man!"?

A person who, when simulating Footbridge, doesn't feel a strong resistance and sense of alarm or callousness in contemplating pushing, or who doesn't "sense" the shock and appearance of callousness to onlookers, or who doesn't imaginatively "project" the difficulty of coming to peace with, or defending, the act after the fact, is someone whose underlying affect and motivation system is displaced from the normal range. And if the evidence and my students are to be believed, the displacement is more in the direction of moral indifference than moral altruism, and more in the direction of untrustworthiness than assurance.

The often-noticed feature of "impersonality" in Loop and Switch emerge in this context in a different light. It is a terrible thing to sacrifice one to save five, but an emergency can make this necessary. What we learn about someone's motivational structure from her performing such an act depends upon the *likelihood* that someone with typical feelings would willingly perform the action. And "impersonality" makes it more likely that someone with typical feelings could bring herself to do so. So a person who pulls the lever in Switch or Loop inspires neither greater nor lesser trust. (On the brain's construction of personality models, see Hassabis et al., 2014.)

"Impersonality," however, is an abstract characteristic, mediated by a model of agents in situations, and not a merely concrete feature of the action. If I ask my students, "When you visualized this scenario, which potential victim seemed to you most proximate?," this, too, seems like a vague and impressionistic question. Yet once more they readily responded. And again an interesting pattern emerged (see Table 2; the most frequent response is **underlined**).

Imaginative proximity to the single victim is as high in Beckon as it is in Footbridge, despite the greater physical distance and the lack of any contact or force. Indeed, in Beckon all the force comes from the victim—and that may help explain the scenario's effect, since inducing his agential force requires that one engage with the victim as an intentional agent, upon whose trust one is relying in getting him to understand and follow one's gesture. This brings that victim imaginatively close in a way such that a willingness to act in Beckon inspires a loss of trust in my students equal to a willingness to act in Footbridge. The same surprising *sang froid* and seeming callousness would be needed.

Isn't there a similar imaginative engagement with the victim in Wave? Not quite. In Wave one is counting on the group to see and respond to one's waving, but not counting on the *victim* in particular doing so. Hence all six are, according to most of my students, imaginatively most proximate. (Interestingly, though perhaps not in a statistically significant way, fewer students say the individual victim is imaginatively most proximate in Wave than in Switch, 16% vs. 28%, and more students say that the five who would be saved are most proximate, 25% vs. 9%. After all, in Wave one is counting upon *the five* to understand and follow one's gesture, but *not* upon the sixth.)

As a test, let us return to Bus, which shows a strikingly different pattern from either Footbridge and Beckon, on the one hand, or Switch and Wave, on the other. In the matter of trust, only in Bus does a roommate or friend who takes action to intervene *gain* in net trustworthiness (see Table 1). And only in Bus are the five people potentially saved imaginatively most proximate (see Table 2). These two facts are, I believe, related to what it is like to simulate Bus using a causal/evaluative model of the scenario as a whole. Bus is unlike the other cases because it involves a background threat that is shared by the whole population, since all are at risk when terror-bombing is common—"we are all in this together," so to speak. That changes the framing of pushing the one individual who could block the bomber's entry into the bus. A simulation of pushing the man in this setting can be "felt" to be a form of group self-defense, a desperate and bloody act, but the kind of response needed to foil terrorists and save as many as possible in a society under threat. Note that only in Bus does

Table 2
Trolley problems and imaginative proximity.

	All six are equally proximate	The single individual potentially killed is most proximate	The five individuals potentially saved are most proximate
Switch (n = 31)	**59%**	28%	9%
Footbridge (n = 31)	39%	**58%**	3%
Wave (n = 32)	**59%**	16%	25%
Beckon (n = 31)	32%	**65%**	3%
Bus (n = 30)	30%	17%	**53%**

one enter the scenario imaginatively "among" the potential victims, and only in Bus are the five who would be saved highly imaginatively proximate as a group. Someone who can act as required to minimize the loss of life under the pressure of such dire circumstances may inspire increased, not decreased, trust. Related evidence comes from the experimental result of Lucas and Livingston (2014), who found that inducing a greater sense of "social connectedness" among subjects who are considering trolley-like dilemmas also increased the rate approving action to protect the five potential victims, even when this involved giving "push"-type responses in Footbridge-like dilemmas.

5.5. Normative issues

The persistent intuitive "sense" that pushing in Footbridge is morally problematic could thus be grounded in a model-based representation that includes agents as well as actions, a modeling that would support an inverse inference—which appears to be accurate—from a stronger disposition to push to a greater likelihood that the agent in question is more callous and less empathic than normal, hence less trustworthy.

Numerous historically-important moral theories are centered in the first instance upon the evaluation of agents and their attitudes, not acts. According to many virtue theorists, for example, whether an act ought to be performed in a given circumstance is to be answered indirectly, by understanding how a virtuous person would see and think about the situation and the act, and whether she would be motivated to perform it (for discussion, see Annas, 2004). And many of the most important figures in the utilitarian tradition, including Hume and John Stuart Mill (1863/2001), have placed questions of cultivating moral sentiments and fellow feeling, encouraging general types of actions, and promoting better social relations—rather than a theory of optimal individual acts—at the center of their thinking about how a concern for the common good is to be applied in practice. It could well be true that the kinds of sensibilities, attitudes, motivational structures, and interpersonal relations it would be best overall for people to develop, would also dispose them to push the lever in Switch but not to push the man off the Footbridge. Or to wave at the workers in Wave but not to beckon the man in Beckon. Or to push the man in Bus.

There is nothing irrational about this. Most human goods, and especially such life-sustaining goods as friendship, family, social solidarity, mutual respect, and humane caring, depend not only upon the acts performed but upon the attitudes, affect, and will of those involved. Our implicit causal/evaluative models are developed over a life-experience that begins with a long period of dependence upon others in which, as we've seen, sensitivity to the good or ill will others, and motivation to understand and help them, emerge early and remain important. When simulating these hypothetical scenarios, such implicit models could well generate a sense that there is likely something amiss in the underlying psychology of a person who would spontaneously throw the man off the bridge in Footbridge or beckon the man in Beckon, and that a society composed of people with such psychologies would not be a better place.

My students needn't mistakenly think their own motivational structure would *change* if they pushed in Footbridge, or that the general undesirability of a motivational structure compatible with willingly pushing in Footbridge is a decisive reason to let five people die "impersonally" in order to avoid "personally" killing one. Indeed, by the end of a term of introductory ethics, typically, fewer students are confident that they *should not* perform this bare, singular act (in a recent instance, 56% answered *yes* to pushing by the end of term, 44% *no*; n = 54). From this we cannot conclude that they are right. But if my students—even those who have concluded that this bare, singular act is what the situation might require—were to sense that there is more than unthinking habit involved in reluctance to embrace this conclusion, would they be wrong?

5.6. Realistic trolley problems?

If deeper causal/evaluative modeling of agents and actions explains the persistence and force of the intuitive asymmetries in trolley problems, we should be able to get different intuitive dynamics in otherwise similar cases that *remove* direct agency. Interestingly, there are now such cases before us as a society: Should we legislate how self-driving vehicles are to be programmed to respond in emergency situations?

When I first ask my students whether self-driving vehicles should be programmed to swerve to the side to avoid five pedestrians in a cross-walk, even though this would cause the death of one pedestrian on a side walkway, a majority typically answers *yes*, in a response reminiscent of Switch (most recently, 82% answered *yes*, while 18% answered *no*; n = 33). Yet when I ask whether such cars should be programmed to swerve to avoid five pedestrians in a cross-walk, even though this would cause the vehicle to collide with a side wall, killing the *rider* in the car, the majority verdict flips (most recently, only 38% of the students answered *yes*, while 62% answered *no*; n = 37), a response reminiscent of Footbridge. In effect, this pair of realistic "trolley problems" reproduces the asymmetry in verdicts found in Switch and Footbridge (for a similar result, see Bonnefon, Shariff, & Rahwan, 2016).

Or does it? Let's test the robustness of this asymmetry. Typically, at the next class session, I ask the students whether, in imagining the original emergency scenarios, they put themselves in the point of view of someone riding *in* the car or someone outside the car. And here we find an interesting asymmetry paralleling the first: a majority had placed themselves imaginatively in the car (most recently, 66% vs. 34%; n = 35). Once they've seen this result, I ask again about programming the car to swerve into a wall to avoid killing five pedestrians, even if this kills the rider. The answer typically changes markedly (most recently, 57% answered *yes*, 43% *no*; n = 35). If I ask the same question again a week later, the original asymmetry is typically almost gone (most recently, 63% answered *yes*, 37% *no*; n = 33). In other words, an asymmetry that at first seemed as stark as Switch vs. Footbridge has largely disappeared, as students continue to think about the problem, encouraged only to vary their perspective.

Contrast this rapid re-evaluation with the original trolley problems, which, after many years and thousands of articles, continue for most who study them to exhibit the same intuitive asymmetry. The initial asymmetry in the case of self-driving cars is not "deep" or persistent, I conjecture, because the locus of agency has been moved to the question of what general algorithmic principles we, as a society of sometime-drivers and sometime-pedestrians, would legislate in light of a mutual exposure to a novel technological risk—we do not have to worry about the "psychology" of the driver.

That model-based simulation and evaluation might be at the bottom of the trolley asymmetry coheres with the evidence offered in Amit and Greene (2012). They note that fMRI studies of trolley problems (Greene, Nystrom, Engell, Darley, & Cohen, 2004; Greene et al., 2001) have consistently found *two* key patterns. The first is the now-familiar fact that considering "personal harm" dilemmas like Footbridge is associated with greater activation in certain areas associated with affective processing, and lower activation in certain areas concerned with task-based controlled cognition. And the second is that considering "personal harm" dilemmas also elicited greater activity in the default mode network, which, as we saw earlier, appears to be the core system by which the brain draws upon episodic and semantic memory to attribute mental states to others and to simulate and evaluate real or hypothetical actions (Buckner et al., 2008). This suggests that model-based simulation and evaluation might be relatively *more* active in assessing "personal harm" dilemmas like Footbridge and Beckon, rather than less. In "impersonal harm" cases like Switch, simulation makes the fates of all six potential victims imaginatively proximate (see again Table 2), facilitating mental processing in terms of a straightforward comparison of overall benefits and costs, rather than vivid simulation of actions and outcomes.

Moreover, Amit and Greene (2012) found that introducing a task that interferes with *visualization* tends to increase the rate of "cost/benefit" responses in Footbridge cases and their ilk. Visualization appears to draw upon a neural network that overlaps extensively with default mode simulation of actions and imputation of mental states (Hassabis & Maguire, 2009). Thus interference with visualization would likely disrupt deeper simulation and favor shallower processing and straightforward cost/benefit calculation. Similarly, patients with vmPFC damage or frontotemporal dementia are reported to have difficulty imagining or empathically simulating the affective states of others, and so they, too, might be more likely to use relatively shallower processing to give a simple cost/benefit response in Footbridge-like cases—and indeed they do have a higher rate of push-like verdicts (Koenigs et al., 2007). Moll and Oliveira-Souza (2007) suggest that an intact vmPFC and frontopolar cortex might be essential for the distinctive combination of cognition and affect that constitutes appropriate moral sentiment.

What, then, about "moral dumbfounding"? Given the relative opacity of implicit processing, it might be unsurprising that subjects find it difficult to explain their persistent intuitive resistance to Julie and Mark's incestuous act. But what model-based explanation would account for this persistence, given the absence of malicious intent or bad consequences? Consider:

> ***Janet & Matt*** *are brother and sister. They are traveling together in France on summer vacation from college. One night they are staying alone in a cabin near the beach. They decide that it would be interesting and fun if they tried playing Russian Roulette. At very least it would be a new experience for each of them. Fortunately, when the spin the revolver's chambers, neither of them lands on the bullet. They both enjoy playing Russian roulette, but they decide not to do it again. They keep that night as a special secret, which makes them feel even closer to each other. What do you think about that, was it OK for them to play Russian roulette?*

The case of Janet & Matt has the same benign consequences as the case of Julie & Mark, but now the reason why their action wasn't OK is very evident to us. Though their intent was not malicious, their thinking was reckless, and our causal/evaluative modeling of actions is modal, incorporating risks as well as actual outcomes. If the gun had gone off, either Janet or Matt would be dead, the other would have a lifetime of regret, and family and friends would suffer a grievous loss for no purpose. An imaginative simulation of their decision-making, as viewed from multiple perspectives in light of multiple possible outcomes, would be unlikely to generate an intuitive "sense" that this is an appropriate way to enliven a dull evening. That the subjects in Haidt (2001) pointed to such potential consequences of incest as psychic harm or traumatic pregnancy to explain their verdicts might not be mere confabulation—these were, in fact, serious risks, to which Julie and Mark gave insufficient weight as they cast about for something fun to do.

Reinforcement learning takes place at many levels and degrees of abstraction and generality. In any given case, elements of both model-free and model-based learning are likely to play a role. Indeed, recent work suggests that the systems involved in model-free and model-based overlap extensively, and that each may play various roles in shaping how the brain uses the other, even in classic Pavlovian conditioning (Dayan, 2012; Dayan & Berridge, 2014; Doll, Simon, & Daw, 2012). Model-free learning and control might help model-based deliberation avoid the regresses we discussed at the outset, and model-based learning and control might monitor habitual behaviors to prevent us from becoming oblivious to change. For complex imaginative tasks like simulating and evaluating hypothetical scenarios, it is quite likely that both are at work, as parts of a larger causal/evaluative modeling competency which generates intuitive assessments of acts, but also of agents, practices, and traits of character—as it must if it is to be adequate to moral thought and practice (cf. Crockett, 2013; Moll & Oliveira-Souza, 2007).

6. Explicit and implicit moral learning

Shared consideration of our intuitive assessments—whether in the ethics classroom, among friends, within families, through public debates, or in psychology journals—reveals a special role for *explicit* forms of moral learning which is not limited to socialization into prevailing norms. As in the case of implicit learning of a language or of causal relations, the implicit moral understanding at which one arrives will be hostage to the quality of one's learning environment, and all particular learning environments have limitations and biases—just as no method of learning is free of liabilities (Dayan & Niv, 2008). Thus, it is especially important that, by discussing our intuitive responses together and trying to understand their origins and import, we can share our experience and gain some hope of reducing bias and expanding our knowledge.

Hume's project, of trying to understand how abstract, general causal and moral cognition could arise and acquire justification on the basis of concrete, particular, shared experience, is a bit closer to being realized thanks to a convergence of results across a range of fields that center on the power of learning. And it is fitting that the resulting approach to moral psychology and neuroscience gives a prominent role, as Hume himself did, to imaginative projection, empathic concern, non-perspectival evaluation, and the modeling of agents as well as actions. These processes are not the whole of moral psychology, as Hume realized. "General sympathy" is the "chief source of moral distinctions" and of "most of the virtues" (1738/1978; Conclusion of Book III), yet morality as we find it also includes equity, rules, and the "limited generosity" that binds together families, friends, and affiliates. But by following Hume in trying to see how the core of morality could be acquired experientially in ways parallel to such paradigms of knowledge as causal relations, we begin to glimpse what moral learning can be.

References

Amit, E., & Greene, J. D. (2012). You see, the ends don't justify the means: Visual imagery and moral judgment. *Psychological Science, 20*, 1–8.

Annas, J. (2004). Being virtuous and doing the right thing. *Proceedings and Addresses of the American Philosophical Association, 78*, 61–75.

Apicella, C. L., Marlowe, F. W., Fowler, J. H., & Christakis, N. A. (2012). Social networks and cooperation in hunter-gatherers. *Nature, 481*, 497–501.

Aslin, R. N., Saffran, J. R., & Newport, E. L. (1998). Computation of conditional probability statistics by 8-month-old infants. *Psychological Science, 9*, 321–324.

Baez, S., Manes, F., Huepe, D., Torralva, T., Fiorentino, N., Richter, F., ... Ibanez, A. (2014). Primary empathy deficits in frontotemporal dementia. *Frontiers in Aging Neuroscience, 6*, 1–11.

Bartels, D. M., & Pizarro, D. A. (2011). The mismeasure of morals: Antisocial personality traits predict utilitarian responses to moral dilemmas. *Cognition, 121*, 154–161.

Behrens, T. E. J., Hunt, L. T., Woolrich, M. W., & Rushworth, M. F. S. (2008). Associative learning of social value. *Nature, 456*, 245–250.

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

Bernhardt, B. C., & Singer, T. (2012). The neural basis of empathy. *Annual Review of Neuroscience, 35*, 1–23.

Berthier, N. E., Rosenstein, M. T., & Barto, A. G. (2005). Approximate optimal control as a model for motor learning. *Psychological Review, 112*, 329–346.

Blair, I. V. (2002). The malleability of automatic stereotypes and prejudice. *Personality and Social Psychology Review, 6*, 242–261.

Blair, R. J. R. (2006). The emergence of psychopathy: Implications for the neurophysiological approach to developmental disorders. *Cognition, 101*, 414–442.

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

Blaisdell, A. P., Sawa, K., Leising, K. J., & Waldmann, M. R. (2006). Causal reasoning in rats. *Science, 311*, 1020–1022.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*, 1573–1576.

Bosevoski, J. J., & Lee, K. (2006). Children's use of frequency information for trait categorization and behavioral prediction. *Developmental Psychology, 42*, 500–513.

Brown, D. E. (2004). Universals, human nature, and human culture. *Daedalus, 133*, 47–54.

Brown, S. L., Nesse, R. M., Vinokur, A. D., & Smith, D. M. (2003). Providing social support may be more beneficial than receiving it: Results from a prospective study of mortality. *Psychological Science, 14*, 320–327.

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *New York Academy of Sciences, 1124*, 1–38.

Conway, P., & Gawronski, B. (2013). Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. *Journal of Personality and Social Psychology, 104*, 216–235.

Craig, A. D. (2009). How do you feel—now? The anterior insula and human awareness. *Nature Reviews Neuroscience, 10*, 59–70.

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

Crockett, M., Kurth-Nelson, Z., Siegel, J. Z., Dayan, P., & Dolan, R. J. (2014). Harm to others outweighs harm to self in moral decision making. *PNAS, 111*, 17320–17325.

Csibra, G. (2003). Teleological and referential understanding of action in infancy. *Philosophical Transactions of the Royal Society: B, 358*, 447–458.

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

Cushman, F. A., Gray, K., Gaffey, A., & Mendes, W. (2012). Simulating murder: The aversion to harmful action. *Emotion, 12*, 2–7.

Dahl, A., & Kim, L. (2014). Why is it bad to make a mess? Preschoolers' conceptions of pragmatic norms. *Cognitive Development, 32*, 12–22.

Dasgupta, N., & Rivera, L. M. (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., & Doya, K. (2006). The computational neurobiology of learning and reward. *Current Opinion in Neurobiology, 16*, 199–204.

Dayan, P. (2012). How to set the switches on this thing. *Current Opinion in Neurobiology, 22*, 1068–1074.

Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience, 14*, 473–492.

Dayan, P., & Daw, N. D. (2008). Decision theory, reinforcement learning, and the brain. *Cognitive, Affective, & Behavioral Neuroscience, 8*, 429–453.

Dayan, P., & Niv, Y. (2008). Reinforcement learning: The good, the bad and the ugly. *Current Opinion in Neurobiology, 18*, 185–196.

Decety, J., Echols, S., & Correll, J. (2010). The blame game: The effect of responsibility and social stigma on empathy for pain. *Journal of Cognitive Neuroscience, 22*, 985–997.

Decety, J., Michalska, K. J., & Kinzler, K. D. (2012). The contribution of emotion and cognition to moral sensitivity: A neurodevelopmental study. *Cerebral Cortex, 2*, 209–220.

Decety, J., & Porges, E. C. (2011). Imagining being the agent of actions that carry different moral consequences: An fMRI study. *Neuropsychologica, 49*, 2994–3001.

Diaconescu, A. O., Mathys, C., Weber, L. A. E., Daunizeau, J., Kasper, L., Lomakina, E. I., ... Stephan, K. E. (2014). Inferring the intentions of others by hierarchical Bayesian learning. *PLOS Computational Biology, 10*, e1003810.

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

Doebel, S., & Koenig, M. A. (2013). Children's use of moral behavior in selective trust discrimination versus learning. *Developmental Psychology, 49*, 462–469.

Doll, B. D., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinions in Neurobiology, 22*, 1075–1081.

Dugatkin, L. A. (2004). *Principles of animal behavior*. New York, NY: W. W. Norton.

Duke, A. A., & Begue, L. (2015). The drunk utilitarian: Blood alcohol concentration predicts utilitarian responses in moral dilemmas. *Cognition, 134*, 121–127.

Dunham, Y., Baron, A. S., & Banaji, M. R. (2008). The development of implicit intergroup cognition. *Trends in Cognitive Sciences, 12*, 248–253.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*, 63–87.

FeldmanHall, O., Dalgleish, T., Evans, D., & Mobbs, D. (2015). Empathic concern drives costly altruism. *NeuroImage, 105*, 347–356.

Fiorillo, C. D., Tobler, P. N., & Schwartz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science, 299*, 1898–1902.

Foot, P. (1967/78). *The problem of abortion and the doctrine of double effect. Reprinted in virtues and vices*. Oxford: Oxford University Press.

Foster, D. J., & Wilson, M. A. (2006). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature, 440*, 680–683.

Friedman, D. (1998). Monty Hall's three doors: Construction and deconstruction of a choice anomaly. *American Economic Review, 88*, 933–946.

Gallistel, C. R., Liu, Y., Krishan, M., Miller, R., & Latham, P. E. (2014). The perception of probability. *Psychological Review, 121*, 96–123.

Gao, Y., & Tang, S. (2013). Psychopathic personality and utilitarian moral judgment in college students. *Journal of Criminal Justice, 41*, 342–349.

Geangu, E., Benga, O., Stahl, D., & Striano, T. (2011). Individual differences in infants' emotional resonance to a peer in distress: Self-other awareness and emotion regulation. *Social Development, 20*, 450–470.

Geisler, W. S. (2011). Contributions of ideal observer theory to vision research. *Vision Research, 51*, 771–781.

Gelman, S. A. (2009). Learning from others: Children's construction of concepts. *Annual Review of Psychology, 60*, 115–140.

Gelman, S. A., & Legare, C. H. (2011). Concepts and folk theories. *Annual Review of Anthropology, 40*, 379–398.

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

Gold, N., Colman, A. M., & Pulford, B. D. (2014). Cultural differences in responses to real-life and hypothetical trolley problems. *Judgment and Decision Making, 9*, 65–76.

Gold, N., Pulford, B. D., & Colman, A. M. (2014). The outlandish, the realistic, and the real: Contextual manipulation and agent role effects in trolley problems. *Frontiers in Psychology, 5*(35), 1–10.

Gopnik, A., & Schulz, L. (2004). Mechanisms of theory formation in young children. *Trends in Cognitive Sciences, 8*, 371–377.

Gopnik, A., & Wellman, H. (2012). Reconstructing constructivism: Causal models, Bayesian learning, and the theory theory. *Psychological Bulletin, 128*, 1085–1108.

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

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

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

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

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition, 107*, 1144–1154.

Greene, J. D., Nystrom, L. E., Engell, A., Darley, J. M., & Cohen, J. D. (2004). The neural basis of cognitive conflict and control in moral judgment. *Neuron, 44*, 389–400.

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*, 2015–2018.

Gupta, A. S., van der Meer, M. A. A., Touretzky, D. S., & Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. *Neuron, 65*, 695–705.

Gweon, H., & Schulz, L. (2011). 16-Month-olds rationally infer causes of failed actions. *Science, 332*, 1524.

Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2010). Infants consider both the sample and the sampling process in inductive generalization. *PNAS, 107*, 9066–9071.

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

Haidt, J. (2007). The new synthesis in moral psychology. *Science, 316*, 998–1002.

Hamlin, J. K. (2013). Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Directions in Psychological Science, 22*, 186–193.

Hamlin, J. K., Mahajan, N., Liberman, Z., & Wynn, K. (2013). Not like me = bad: Infants prefer those who harm dissimilar others. *Psychological Science, 24*, 589–594.

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

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

Harris, P. L., & Corriveau, K. H. (2011). Young children's selective trust in informants. *Philosophical Transactions of the Royal Society: B, 366*, 1179–1187.

Hassabis, D., & Maguire, E. A. (2009). The construction system in the brain. *Philosophical Transactions of the Royal Society: B, 364*, 1263–1271.

Hassabis, D., Spreng, R. N., Rusu, A. A., Robbins, C. A., Mar, R. A., & Schacter, D. L. (2014). Imagine all the people: How the brain creates and uses personality models to predict behavior. *Cerebral Cortex, 24*, 1979–1987.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., & Gintis, H. (2004). *The foundations of human sociality: Economic experiments and ethnography evidence from fifteen small-scale societies*. Oxford, UK: Oxford University Press.

Heyes, C. (2016). Who knows? Metacognitive social learning strategies. *Trends in Cognitive Sciences, 20*, 204–213.

Hinton, G. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences, 11*, 428–433.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning as a rational process: The new synthesis. *Annual Review of Psychology, 62*, 135–163.

Hume, D. (1738/1978). In L. A. Selby-Bigge & P. H. Nidditch (Eds.), *A treatise of human nature*. Oxford, UK: Oxford University Press.

Hume, D. (1751/1998). In T. L. Beauchamp (Ed.), *An enquiry concerning the principles of morals*. Oxford, UK: Oxford University Press. Appendix 2, 90–95.

Hussar, K. M., & Harris, P. L. (2009). Children who choose not to eat meat: A study in early moral decision-making. *Social Development, 19*, 627–641.

Ji, D., & Wilson, M. A. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience, 10*, 100–107.

Johnson, A., van der Meer, M. A. A., & Redish, A. D. (2007). Integrating hippocampus and striatum in decision-making. *Current Opinion in Neurobiology, 17*, 692–697.

Johnson, A., & Redish, A. D. (2007). Neural ensembles in CA3 transiently encode paths forward of the animal at a decision point. *Journal of Neuroscience, 27*, 12176–12189.

Kahane, G., Everett, J. A. C., Earp, B. D., Farias, M., & Savulescu (2015). 'Utilitarian' judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134*, 193–209.

Kahneman, D. (2011). *Thinking, fast and slow*. New York: Farrar, Straus, and Giroux.

Kahneman, D., & Tversky, A. (2000). *Choice, value, and frames*. New York: Russell Sage.

Kamm, F. (2007). *Intricate ethics: Rights, responsibilities, and permissible harm*. New York: Oxford University Press.

Kant, I. (1797/1996). In M. Gregor (Ed.), *The metaphysics of morals*. Cambridge, UK: Cambridge University Press.

Katrakazas, C., Quddus, M., Chen, W.-H., & Deka, L. (2015). Real-time motion planning methods for autonomous on-road driving: State-of-the-art and future research directions. *Transportation Research Part C, 60*, 416–442.

Kennedy, G. E., & Cutts, Q. I. (2005). The association between students' use of an electronic voting system and their learning outcomes. *Journal of Computer Assisted Learning, 21*, 260–268.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to sequences that are neither too simple nor too complex. *PLOS-One, 7*, e36399.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences, 27*, 712–719.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences, 33*, 315–365.

Koenig, M. A., & Echols, C. H. (2003). Infants' understanding of false labeling events: The referential roles of words and the speakers who use them. *Cognition, 87*, 179–208.

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

Kolling, N., Behrens, T. E. J., Mars, R. B., & Rushworth, M. F. S. (2012). Neural mechanisms of foraging. *Science, 336*, 95–98.

Kording, K. P., & Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences, 10*, 319–326.

Krebs, J. R., Ryan, J. C., & Charnov, E. L. (1974). Hunting by expectation or optimal foraging? A study of patch use by chickadees. *Animal Behavior, 22*, 953–964.

Lak, A., Stauffer, W. R., & Schultz, W. (2014). Dopamine prediction error responses integrate subjective value from different reward dimensions. *PNAS, 111*, 2343–2348.

Landy, J. F., & Goodwin, G. P. (2015). Does incidental disgust amplify moral judgments? A meta-analytic review of experimental evidence. *Perspectives on Psychological Science, 10*, 518–536.

Lane, J. D., Harris, P. L., Gelman, S. A., & Wellman, H. W. (2014). More than meets the eye: Young children's trust in claims that defy their perceptions. *Developmental Psychology, 50*, 865–871.

Langston, R. F., Ainge, J. A., & Covey, J. J. (2010). Development of the spatial representation system in the rat. *Science, 328*, 1576–1580.

Le, Q. V., Ranzato, M. A., Monga, R., Devin, M., Chen, K., Corrado, G. S., ... Ng, A. Y. (2012). Building high-level features using large scale unsupervised learning. In *Proceedings of the 29th international congress on machine learning. Edinburgh, Scotland, UK*. .

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in visual cortex. *Journal of the Optical Society of America, A, 20*, 1434–1448.

Liu, D., & Todorov, E. (2007). Evidence for flexible sensorimotor strategies predicted by optimal feedback control. *Journal of Neuroscience, 27*, 9354–9368.

Lucas, B. J., & Livingston, R. W. (2014). Feeling socially connected increases utilitarian choices in moral dilemmas. *Journal of Experimental Social Psychology, 53*, 1–4.

Luo, Y. (2011). Do 10-month-old infants understand others' false beliefs? *Cognition, 121*, 289–298.

Ma, W. J., Navalpakkam, V., Beck, J. M., van den Berg, R., & Pouget, A. (2011). Behavior and neural basis of near-optimal visual search. *Nature Neuroscience, 14*, 783–790.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition? *Psychological Science, 24*, 2351–2360.

Marlowe, F. W. (2003). The mating system of foragers in the standard cross-cultural sample. *Cross-Cultural Research, 37*, 282–306.

Mendez, M. F., Anderson, E., & Shapira, J. S. (2005). And investigation of moral judgment in frontotemporal dementia. *Cognitive, Behavioral, and Affective Neuroscience, 18*, 193–197.

Mill, J. S. (1863/2001). In G. Sher (Ed.), *Utilitarianism* (2nd ed.. Indianapolis: Hackett.

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

Moll, J., & Oliveira-Souza, R. (2007). Moral judgments, emotions and the utilitarian brain. *Trends in Cognitive Sciences, 11*, 319–321.

Moll, J., Zahn, R., Oliveira-Souza, R., Krueger, F., & Grafman, J. (2005). The neural basis of human moral cognition. *Nature Reviews Neuroscience, 6*, 799–809.

Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience, 31*, 69–89.

Navarrette, C. D., McDonald, M. M., Mott, M. L., & Asher, B. (2012). Virtual morality: Emotion and action in a simulated "trolley problem". *Emotion, 12*, 364–370.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance, 3*, 756–766.

Nesse, R. M. (2007). Runaway social selection for displays of partner value and altruism. *Biological Theory, 2*, 143–155.

Nesse, R. M., & Ellsworth, P. E. (2009). Emotion, evolution, and emotional disorders. *American Psychologist, 64*, 129–139.

Nguyen, S. P., Gordon, C. L., Chevalier, T., & Girgis, H. (2016). Trust and doubt: An examination of children's decision to believe what they are told about food. *Journal of Experimental Child Psychology, 144*, 66–83.

Paciello, M., Fida, R., Cerniglia, L., Tramontano, C., & Cole, E. (2013). High-cost helping scenario: The role of empathy, prosocial reasoning and moral disengagement on helping behavior. *Personality and Individual Differences, 55*, 3–7.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). New York: Cambridge University Press.

Pessoa, L. (2008). On the relationship between emotion and cognition. *Nature Reviews Neuroscience, 9*, 148–158.

Pettigrew, T. F. (1998). Intergroup contact theory. *Annual Review of Psychology, 49*, 65–85.

Pettigrew, T. F., & Tropp, L. R. (2006). A meta-analytic test of intergroup contact theory. *Journal of Personality and Social Psychology, 90*, 751–783.

Pleskac, T. J., & Hertwig, R. (2014). Ecologically rational choice and the structure of the environment. *Journal of Experimental Psychology: General, 143*, 2000–2019.

Preuschoff, K., Bossaerts, P., & Quartz, S. R. (2006). Neural differentiation of expected reward and risk in human subcortical structures". *Neuron, 51*, 381–390.

Prinz, J. J. (2004). *Gut feelings: A perceptual theory of emotion*. Oxford, UK: Oxford University Press.

Pullum, G. K., & Scholz, B. C. (2002). Empirical assessment of stimulus poverty arguments. *Linguistic Review, 19*, 9–50.

Quartz, S. R. (2007). Reason, emotion, and decision-making: Risk and reward computation with feeling. *Trends in Cognitive Sciences, 13*, 209–215.

Rai, T. S., & Fiske, A. P. (2011). Moral psychology is relationship regulation: Moral motives for unity, hierarchy, equality, and proportionality. *Psychological Review, 118*, 57–75.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature, 489*, 427–430.

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A neural basis for social cooperation. *Neuron, 36*, 395–406.

Roth-Hanania, R., Davidov, M., & Zhan-Waxler, C. (2011). Empathy development from 8 to 16 months: Early signs of concern for others. *Infant Behavior and Development, 34*, 447–458.

Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron, 36*, 241–263.

Seligman, M. E. P., Railton, P., Baumeister, R., & Sripada, C. S. (2013). Navigating into the future or driven by the past? *Perspectives in Psychological Science, 8*, 119–141.

Seligman, M. E. P., Railton, P., Baumeister, R. A., & Sripada, C. (2016). *Homo prospectus*. New York: Oxford University Press.

Shanks, D. R., Tunney, R. J., & McCarthy, J. D. (2002). A Re-examination of probability matching and rational choice. *Journal of Behavioral Decision Making, 15*, 233–250.

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

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature, 529*, 484–489.

Sloane, S., Baillargeon, R., & Premack, D. (2012). Do infants have a sense of fairness? *Psychological Science, 23*, 196–204.

Smetana, J. G. (1989). Toddlers' social interactions in the context of moral and conventional transgressions in the home. *Developmental Psychology, 25*, 499–508.

Sobel, D. M., & Corriveau, K. H. (2010). Children monitor individual's expertise for word learning. *Child Development, 81*, 669–679.

Sobel, D. M., & Kirkham, N. Z. (2006). Blickets and babies: The development of causal reasoning in toddlers and infants. *Developmental Psychology, 42*, 1103–1115.

Sobel, D. M., & Kirkham, N. Z. (2007). Bayes' nets and babies: Infants' developing statistical reasoning and their representation of causal knowledge. *Developmental Science, 10*, 298–306.

Sripada, C. S. (2012). Mental state attributions and the side-effect effect. *Journal of Experimental Social Psychology, 48*, 232–238.

Stauffer, W. R., Lak, A., & Schultz (2014). Dopamine reward prediction error responses reflect marginal utility. *Current Biology, 24*, 2491–2500.

Stowell, J. R., & Nelson, J. M. (2007). Benefits of electronic audience response systems on student participation, learning, and emotion. *Teaching of Psychology, 34*, 253–258.

Sutton, R. S., & Barto, A. (1999). Reinforcement learning. *Journal of Cognitive Neuroscience, 11*, 126–134.

Tanji, J., Shima, K., & Mushiake, H. (2007). Concept-based behavioral planning and the lateral prefrontal cortex. *Trends in Cognitive Sciences, 11*, 528–534.

Teglas, E., Vul, E., Girotto, V., Gonzalez, M., Tenenbaum, J. B., & Bonatti, L. L. (2011). Pure reasoning in 12-month-old infants as probabilistic inference. *Science, 332*, 1054–1059.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science, 331*, 1279–1285.

Thirioux, B., Mercier, M. R., Blanke, O., & Berthoz, A. (2014). The cognitive and neural time-course of empathy and sympathy: An electrical neuroimaging study of self-other interaction. *Neuroscience, 267*, 286–306.

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *Monist, 59*, 205–217.

Tobler, P. N., O'Doherty, J. P., Dolan, R. J., & Schultz, W. (2006). Reward value coding distinct from attitude-related uncertainty coding in human reward systems. *Journal of Neurophysiology, 97*, 1621–1632.

Tolman, E. C. (1948). Cognitive maps in rats and men. *Psychological Review, 55*, 189–208.

Turiel, E. (2002). *The culture of morality: Social development, context, and conflict*. Cambridge, UK: Cambridge University Press.

Tybur, J. M., Kursban, R., Lieberman, D., & DeScioli, P. (2013). Disgust: Evolved function and structure. *Psychological Review, 120*, 65–84.

Uhlmann, E. L., Zhu, L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition, 126*, 326–334.

Vaish, A., Carpenter, M., & Tomasello, M. (2009). Sympathy through affective perspective taking and its relation to prosocial behavior in toddlers. *Developmental Psychology, 45*, 534–543.

Vaish, A., Missana, M., & Tomasello, M. (2011). Three-year-old children intervene in third-party moral transgressions. *British Journal of Developmental Psychology, 29*, 124–130.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science, 311*, 1301–1303.

Weich, K., Kahane, G., Shackel, N., Farias, M., Savulescu, J., & Tracey, I. (2013). Cold or calculating? Reduced activity in subgenual cingulate cortex reflects decreased emotional aversion to harming in counterintuitive utilitarian judgment. *Cognition, 126*, 364–372.

Wellman, H. (2014). *Making minds: How theory of mind develops*. Oxford, UK: Oxford University Press.

Westgate, E. C., Riskind, R. G., & Nosek, B. A. (2015). Implicit preferences for straight people over lesbian and gay men weakened from 2006 to 2013. *Collabra, 1*, 1–10.

Williamson, R. A., Meltzoff, A. N., & Markham, E. M. (2008). Prior experiences and perceived efficacy influence 3-year-olds' imitation. *Developmental Psychology, 44*, 275–285.

Woodward, A. L., Sommerville, J. A., Gerson, S., Henderson, A. M. E., & Buresh, J. (2009). The emergence of intention attribution in infancy. *Psychology of Learning and Motivation, 51*, 187–222.

Yarrow, K., Brown, P., & Krakauer, J. W. (2009). Inside the brain of an elite athlete: The neural processes that support high achievement in sports. *Nature Reviews Neuroscience, 10*, 585–596.

Yu, L., Weilin, S., Yu, Z., Ting-yong, F., Hao, H., & Hong, L. (2013). Core disgust and moral disgust are related to distinct spatiotemporal patterns of neural processing: An event-related potential study. *Biological Psychology, 94*, 242–248.

Zahn-Waxler, C., Radke-Yarrow, M., Wagner, M., & Chapman, M. (1992). Development of concern for others. *Developmental Psychology, 28*, 1038–1047.

Zmyj, N., Buttelmann, D., Carpenter, M., & Daum, M. M. (2010). The reliability of a model influences 14-month-olds' imitation. *Journal of Experimental Child Psychology, 106*, 208–220.