Moral Competence in Social Robots

Bertram F. Malle
Cognitive, Linguistic, and Psychological Sciences
Brown University
Providence, USA
bfmalle@brown.edu

Matthias Scheutz
Department of Computer Science
Tufts University
Medford, MA 02155 USA
matthias.scheutz@tufts.edu

***Abstract*—We propose that any robots that collaborate with, look after, or help humans—in short, social robots—must have moral competence. But what does moral competence consist of? We offer a framework for moral competence that attempts to be comprehensive in capturing capacities that make humans morally competent and that therefore represent candidates for a morally competent robot. We posit that human moral competence consists of four broad components: (1) A system of norms and the language and concepts needed to communicate about these norms; (2) moral cognition and affect; (3) moral decision making and action; and (4) moral communication. We sketch what we know and don't know about these four elements of moral competence in humans and, for each component, ask how we could equip an artificial agent with these capacities.**

Keywords—ethics, moral agency, social cognition, intentionality, affect and cognition

I. Introduction

There are at least two classes of questions that fall under "robot ethics": (1) ethical questions about designing, deploying, and treating robots; and (2) questions about the robots' own ethics—what moral capacities robots should have and how these capacities could be realized in robotic architectures [1]. Our analysis in this paper focuses entirely on questions of the second kind.

A robot well suited for social interaction with humans would need to have, among other things, social-cognitive capacities (including a "theory of mind") and moral competence. Many psychological phenomena have been studied that could be called "moral competence": decision making about moral dilemmas [2], [3]; self-regulation of emotion and prosocial behavior[4]; moral judgments and their associated emotions [5], [6]; as well as responding to others' moral criticism by means of explanation, justification, or defense [7], [8]. The diversity of phenomena on this list is no coincidence; moral competence is not a single capacity. We propose here a framework that delineates the multiple components of human moral competence and then ask which of these elements should and could make up moral competence in social robots.[1] Because of time and space constraints we will have to stay at the surface, but nonetheless we will try to point to areas where the research on human capacities is lagging behind and needs to be advanced in order to provide a better basis for robotic work; and to areas where robotic work could take some promising early steps.

The guiding question is: What would we expect of morally competent robots? Perhaps not all of human competencies, but surely several of them. And perhaps moral robots will be even "better moral creatures than we are" [9, p. 346], though such an evaluation already presupposes that we know what we are comparing. So what is it that we are comparing? What is moral competence?

We propose that moral competence consists of four broad components (Fig. 1): (1) A system of norms and the language and concepts to communicate about these norms; (2) moral cognition and affect; (3) moral decision making and action; and (4) moral communication. The rest of this paper describes in more detail each of the components.[2]

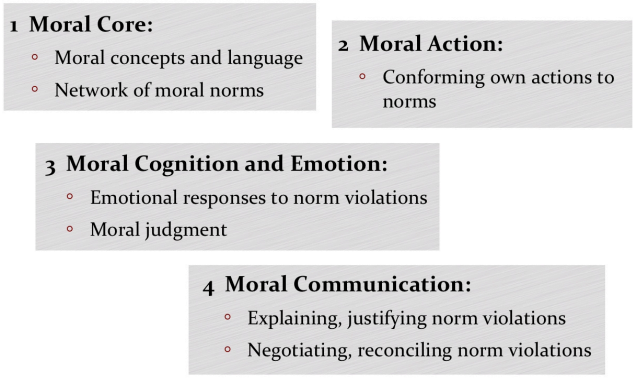

Fig. 1. Four components of moral competence

II. Norms and Moral Language

Morality's function is to regulate human social behavior in contexts in which biological desires no longer guarantee individual and collective well-being [12], [13]. Human communities perform this regulation by motivating and deterring certain behaviors through the imposition of norms and, if these norms are violated, by levying sanctions [14].

[1] I focus here on social robots—home makers, care takers, educators, and the like; but much of what I say applies to robots in other contexts, such a military and rescue, as well.

This project was supported by a grant from the Office of Naval Research, No. N00014-13-1-0269.

978-1-4799-4992-2/14/$31.00 ©2014 IEEE

[2] These components are in some sense weaker than what is often discussed under "moral agency" e.g., [10], [11]; for example, they do not include a deep, reflective self-concept, and they don't presuppose "free will." But the components are also more extensive than typical moral agency demands, which rarely require social-cognitive and -communicative capacities.

This process allows social agents to successfully coordinate their behaviors in complex social communities—made complex by diversified tasks, roles, collective behavior, and interdependence of outcomes. Being equipped with a norm system thus constitutes a first critical element in human moral competence [15], [16].

But a norm system is conceptually and linguistically demanding, requiring language for learning it, using it, and negotiating it. Thus, at its core, moral competence requires a network of moral norms and a language (and associated concepts) to represent and implement it [11].

A. Moral Language

Some rudimentary moral capacities may operate without language, such as the recognition of prototypically prosocial and antisocial actions [17] or foundations for moral action in empathy and reciprocity [18]. But a morally competent human will need a vocabulary to express moral concepts and instantiate moral practices—to blame or forgive others' transgressions, to justify and excuse one's behavior, to contest and negotiate the importance of one norm over another.

Such a moral language has three major domains:

1. **A language of norms and their properties** (e.g., "fair," virtuous," "reciprocity," "obligation," "prohibited," "ought to");
2. **A language of norm violations** (e.g., "wrong," "culpable," "reckless," "thief");
3. **A language of responses to violations** (e.g., "blame," "reprimand," "excuse," "forgiveness").

Within each domain, there are numerous distinctions, and some have surprisingly subtle differentiations. For example, we recently uncovered a two-dimensional organization of 28 verbs of moral criticism [19] that suggests people systematically differentiate among verbs to capture criticism of different intensity in either public or private settings (see Fig. 2).

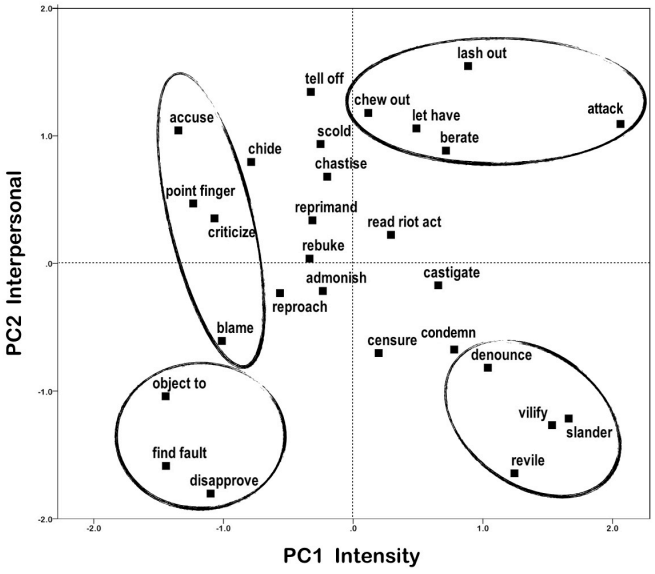

Fig. 2. Verbs of moral criticism in two-dimensional feature space

Obviously, the role of norms is central to a language of morality. Let us take a closer look at the challenges both in understanding how human moral norms operate and how they could possibly be implemented in autonomous robots.

B. Moral Norms

Many questions about norms already arise when we examine the human case. How are norms acquired? How are they represented in the mind? What properties do they have that allow them to be so context-sensitive and mutually adjusting as we know them to be? We will briefly examine these questions.

Acquisition. Though evidence is limited on the development of norms, data on children's early use of moral language [20] suggest that they are rarely exposed to abstract rules but rather hear and express concrete moral judgments. Consider these examples [20, pp. 74–77]:

> "they are mean to that man because they put him in that glue,"
>
> "that's not nice! That was naughty!"
>
> "he did something wrong."

Nonetheless, children are somehow able to abstract from concrete instances to rather general rules, such as "bombs hurt people" [20, p. 77] or even abstract principles such as the act-omission distinction [21] or battery through contact [22]. Fortunately, children are the most powerful learning machines in the universe, as we can see in just about all domains of cognition, including learning language and acquiring varied category systems such as personality traits, animals, or plants. In addition, the toolbox of social cognition adds a powerful supportive structure for the acquisition of moral norms and moral judgments. This structure includes mastery over the concepts of goal and desire [23], [24]; the distinction between intentional and unintentional behavior [25], [26]; between beliefs and desires [27], [28], desires and intentions [29], and a variety of emotions [30], as well as all the rich linguistic expressions of these distinctions [31], [32].

Representation. Another largely ignored question is how norms are represented in the human mind. Are they networks of concepts? And is that any different from networks of other concepts? Nonnormative concepts (e.g., tree, weight) can have evaluative tone, for sure, but do they have motivational force as normative concepts do (e.g., fairness, obligation)? Goal concepts—which typically are explicitly represented in robotic architectures—seem close to norm concepts. But perhaps the most unique category of norms are *values*, and there are indications that they cannot simply be reduced to goals [33]. Moreover, if Jon Elster is correct in claiming that "social norms provide an important kind of motivation for action that is irreducible to rationality or indeed any other form of optimizing mechanism" [34, p. 15], then a simple goal-based action control system will not do for moral social robots.

Levels. Another interesting feature of norms is that they are layered over many levels of abstraction. As an illustration, consider the following violation: A commercial airplane pilot flies a plane despite a recently diagnosed heart condition. What norm is violated here?

- Pilots ought not to fly when they know they have a physical disposition that, if becoming acute, could threaten their ability to safely continue flying. (This expands from heart condition to migraines, epileptic seizures, sleep apnea….)
- Pilots ought not to fly when they are aware of factors that risk the continued safety of their passengers. (This expands from the pilot's own risk factors to mechanical risks, weather risks, etc.)
- People ought to keep others safe who are put in their care. (Expands to many more roles, contexts, potential victims, etc.)
- People ought to protect human life.

Contextual activation. A related facet of norms is that they appear to be very quickly activated, and presumably at the right level of abstraction, because people detect norm violations within a few hundred milliseconds [35]. One possibility is that physical or linguistic contexts activate subsets of action-specific norms (what one is or is not permitted to do in the particular context), and violations can then be rapidly detected. The relation of these concrete norms to more abstract norms may be constructed offline, through conversation and conceptual reorganization [5].

For designing a morally competent robot, all these features of norms present serious challenges. But if norms are special kinds of representations, connected in some (admittedly flexible) network, and activated by percepts, then there is no principled reason why they could not be implemented in a computational system. They could then operate as constraints on the robot's possible actions, selecting optimal (least violating) action favorites (cf. [36]).

One issue we do not consider a problem is which "ethical theory" to build into a robot (utilitarianism, Kantian ethics, etc.). Humans probably do not follow one particular ethical theory, and even if they did it is not clear whether robots must have the same. However the system arrives at its judgments of what is or is not a norm violation, the judgments must conform to the community in which the robot is embedded. This also means that robots, like children, will have to learn different norm systems when they are deployed in different communities. Many communities have large overlap in the norms they follow, so moving from one community to another is not an insurmountable problem, for humans or robots; in fact, robots may be better at keeping track of various different communities' norm systems and thus be less morally myopic than humans are.

III. Moral Cognition and Affect

We have proposed that two of the major domains of moral language are (a) norm violations and (b) responses to norm violations. What psychological processes are involved in detecting and responding to such violations? These processes are usually treated under the label of *moral judgment*, but we need to distinguish between at least two kinds of moral judgment [37]. First, people evaluate *events* (outcomes, behaviors) as bad, good, wrong, or (im)permissible; second, they evaluate *agents* as morally responsible, deserving blame or praise. The key difference between the two types is the amount of information processing that normally underlies each judgment. Whereas event judgments merely register that a norm has been violated, agent judgments such as blame take into account the agent's specific causal involvement, intentionality, and mental states.

Of course, registering that an event violated a norm is not a trivial endeavor, and we realize this quickly when we ask how young children or robots detect such violations. What is needed for such a feat? Minimally, event segmentation, multi-level event representations (because different levels may conflict with different norms), and establishing the event's deviation from relevant norms. Nontrivially, which norms are *relevant* must somehow be selected from the available situation information and existing knowledge structures.

To arrive at agent judgments, people search for causes of the detected norm-violating event; if the causes involve an agent, they wonder whether the agent acted intentionally; if she acted intentionally, what reasons she had; and if the event was not intentional, whether the agent could and should have prevented it [38]. The core elements here are causal and counterfactual reasoning and social cognition, and that is why a number of researchers suggest that moral cognition is no unique "module" or "engine" but derives from ordinary cognition [39], [40], but reasoning within the context of norms.

Where in all this is affect? The specific roles of affective phenomena in moral judgment are still debated. There is little doubt that the detection of a norm violation often leads to a negative affective response—an evaluation that *something is bad*, perhaps accompanied by physiological arousal and facial expressions. But exactly what this affective response sets in motion is unclear: a marker that something important occurred [41]? A strengthened motivation to find the cause of the bad event [42]? Or a biased search for evidence that allows the perceiver to blame somebody [6]? And what do we make of the fact that people can make moral judgments without much affect at all [43], [44] or that moral emotions such as anger or resentment require specific cognitive processes [45]? Nobody would deny that affective phenomena often accompany moral judgments and that they probably facilitate learning moral norms; but there is little evidence for the claim that they are *necessary* or *constitutive* of those judgments. And if emotions are not necessary or constitutive of moral judgments, then robots—even if they do not have emotions—can very much be moral.

IV. Moral Decision Making and Action

Human moral decision making has received a fair amount of attention in the research literature, with a focus on how humans handle moral dilemmas [2], [46], [47]. Much of what these studies reveal is how people resolve difficult conflicts within their norm system (e.g., saving multiple lives by sacrificing one life). A popular theoretical view of such situations is that initial affective responses can be overridden by deliberation [48]. But evidence against this override view is increasing [49]–[52]; people's judgments seem to involve a package of affective and cognitive processes that all deal simultaneously with the conflict set

up by the experimenter. Further, how judgments about carefully constructed dilemmas translate into everyday moral decisions is not entirely clear.

In fact, the list of actual psychological factors that influence moral action is long, certainly including momentary affective states to personality dispositions, automatic imitation to group pressure, and heuristics to reasoned action. This overdetermination is no different from nonmoral actions [53]. What makes certain actions *moral* is the involvement of socially shared norms, not just individual goals. In humans, there is frequent tension between these social norms and the agent's own goals, and it is this tension that brings into play two additional psychological factors that guide moral action: empathy and self-regulation [4], [54], [55], both of which are designed to favor communal values over selfish interest.

In designing a robot capable of moral decisions and actions, the tension between self-interest and community benefits can probably be avoided from the start, making genuine empathy and self-regulation dispensable. However, humans are highly sensitive to other people's displays of empathy, and a robot that appears to coldly assess moral situations may not be trusted. The robot's modeling of the human view of situations and the its communication of having understood and taken into account this view (perspective taking rather than affective empathy) may go a long way toward building trust between human and machine. Of course, these communications and attempts at perspective taking must not be merely verbal programs, deceptive attempts to coax the human's trust; some computational analog of affect and valuing may be needed for human-machine interactions to succeed [56], [57].

Thus, we are back to the challenge of building a norm system into the robot, including values that make the machine care about certain outcomes and that guide the robot's decision making and action, especially in the social world. Note that "caring" here is a key concept that would first have to be spelled out—what it means in computational terms for a machine to care about something.

Whether in all this a robot needs "intuitive" processes, which seem to play an important role in human moral decision making [5], [58], is also an interesting question. To the extent that these processes resolve a human capacity limitation, and to the extent that robots do not have this limitation, robots would not need moral intuitions. However, will humans be suspicious of robots that do not "feel" right and wrong but reason over it? We doubt it, but this is an empirical question.

Note that we have not mentioned the need to have "free will" in order to accord moral action capacity. That is because ordinary people require only choice and absence of constraints for actions to be "free"—any metaphysical requirements of nondeterminism or a soul do not seem to be relevant [59], [60]. Thus, if robots can have choice capacity and are not massively constrained by human control (inflexible programs and "emergency stop mechanisms"), then they could act "freely."

If robots can make moral decisions without free will, then they are likely targets of moral blame as well [61]. In humans, moral blame is not restricted to intentional action—negligence, mistakes, and errors are blamed as well—but blame does presuppose that the agent has the *capacity* for intentional action in order to correct mistakes and prevent negative outcomes in the future. That (and no mystical free will) would be needed for robots, too. Blame informs, corrects, and provides an opportunity to learn [37], [62]; to the extent that robots can change and learn, they may well be appropriate targets of blame. "You could have done otherwise," said to a human or a robot, does not question the deterministic order of the universe but invites a consideration of options that were available at the time of acting but were ignored or valued differently—and should be taken into account in the future.

V. Moral Communication

The suite of cognitive tools that enable moral judgment and decision making still are insufficient to achieve the socially most important function of morality: to regulate other people's behavior. For that, moral communication is needed. Moral perceivers often express their moral judgments to the alleged offender or to another community member [63], [64]; they sometimes have to provide evidence for their judgment; the alleged offender may contest the charges or explain the action in question [8]; and social estrangement may need to be repaired through conversation or compensation [65], [66]. Social robots need not have command over all these communicative acts, but two seem especially important: expressing their detection of a norm violation and explaining their own action when confronted with the charge that it was a violation.

Expressing one's moral judgments (of events or agents) will not be especially difficult for the robot if its moral cognition capacity is well developed and the robot has basic natural language skills. The subtle varieties of delivering moral criticism may be too difficult to master (e.g., the difference between scolding, chiding, or denouncing; [19]), but on the positive side, the anger and outrage that accompanies many human expressions of moral criticism can be easily avoided. This may be particularly important when the robot is partnered with a human—such as with a police officers on patrol or with a teacher in a classroom—and points out (inaudible to others) a looming violation. Without the kind of affect that would normally make the human partner defensive, the moral criticism may be more effective. However, in some communities, a robot that detects and, presumably, remembers and reports violations to others would itself violate trust norms. For example, a serious challenge in the military is that soldiers that are part of a unit do not report one another's violations (including human rights violations). A robot would not be susceptible to such pressures of loyalty, but the robot may also not find its way into the tight social community of soldiers, being rejected as a snitch.

Explaining immoral behaviors, the second important moral communication capacity, is directly derived from explaining behaviors in general, which is relatively well understood in psychology [67], [68]. Importantly, people treat intentional and unintentional behaviors quite differently: they explain intentional behaviors with reasons (the agent's beliefs and desires in light of which and on the

ground of which they decided to act), and they explain unintentional behaviors with causes. Correspondingly, explaining intentional moral violations amounts to offering reasons that justify the violating action, whereas explaining unintentional moral violations amounts to offering causes that excuse one's involvement in the violation [37]. In addition, and unique to the moral domain, unintentional moral violations are assessed by counterfactuals: what the person could have done differently to prevent the negative event. As a result, moral criticism involves simulation of the past (what alternative paths of prevention may have been available) and simulation of the future (how one is expected to act differently to prevent repeated offenses). Both seem computationally feasible [69].

Explanations of one's own intentional actions require more than causal analysis and simulation; they require access to one's own reasoning en route to action. Some have famously doubted this capacity in humans [70], but these doubts do not apply in the case of reasons for action [71]. A robot, in any case, should have perfect access to its own reasoning. But once it accesses the trace of its reasoning, it must articulate this reasoning in humanly comprehensible ways (as beliefs and desires), regardless of the formalism in which it performs the reasoning. This amounts to one last form of simulation: modeling what a human would want to know so as to understand (and accept) the robot's decision in question. In fact, if the robot can simulate in advance a possible human challenge to its planned action and has available an acceptable explanation, then the action has passed a social criterion for moral behavior.

VI. From Here on Out

In light of the extensive and complex components of human moral competence, designing robots with such competence is an awe-inspiring challenge. The key steps will be to build computational representations of norm systems and incorporate moral concepts and vocabulary into the robotic architecture. Once norms and concept representations are available in the architecture, the next step is to develop algorithms that can computationally capture moral cognition and decision-making. The development of these processes might take some time, but it does not seem nearly as difficult as developing the communicative capacities of moral agents we alluded to earlier, in part because of the complexity and flexibility of human language. Unless new computational learning algorithms enable robots to acquire human-like natural language capabilities, we might need to move from programming robots (and occasionally letting them learn) to *raising* robots in human environments. This may be the only way to expose them to the wealth of human moral situations and communicative interactions. Infants are not pre-programmed with norm systems or language either; but with countless repetitions in initially constrained contexts, and with a strong social reward function, they learn just about anything culture throws at them. Admittedly, humans have powerful learning mechanisms. But some of these mechanisms are available to robots as well, and new ones will be developed that could even exceed human capabilities. We do not know how well robots can learn norms, concepts, and language unless we give them abundant opportunities to do so.

References

[1] J. P. Sullins, "Introduction: Open questions in roboethics," *Philos. Technol.*, vol. 24, no. 3, p. 233, Sep. 2011.

[2] J. Mikhail, "Universal moral grammar: Theory, evidence and the future," *Trends Cogn. Sci.*, vol. 11, no. 4, pp. 143–152, Apr. 2007.

[3] J. D. Greene, R. B. Sommerville, L. E. Nystrom, J. M. Darley, and J. D. Cohen, "An fMRI investigation of emotional engagement in moral judgment," *Science*, vol. 293, no. 5537, pp. 2105–2108, Sep. 2001.

[4] N. Eisenberg, "Emotion, regulation, and moral development," *Annu. Rev. Psychol.*, vol. 51, pp. 665–697, 2000.

[5] J. Haidt, "The emotional dog and its rational tail: A social intuitionist approach to moral judgment," *Psychol. Rev.*, vol. 108, no. 4, pp. 814–834, Oct. 2001.

[6] M. D. Alicke, "Culpable control and the psychology of blame," *Psychol. Bull.*, vol. 126, no. 4, pp. 556–574, Jul. 2000.

[7] G. R. Semin and A. S. R. Manstead, *The accountability of conduct: A social psychological analysis*. London: Academic Press, 1983.

[8] C. Antaki, *Explaining and arguing: The social organization of accounts*. London: Sage, 1994.

[9] J. S. Hall, *Beyond AI: Creating the conscience of the machine*. Amherst, NY: Prometheus Books, 2007.

[10] C. Allen, G. Varner, and J. Zinser, "Prolegomena to any future artificial moral agent," *J. Exp. Theor. Artif. Intell.*, vol. 12, no. 3, pp. 251–261, Jul. 2000.

[11] J. Parthemore and B. Whitby, "What makes any agent a moral agent? Reflections on machine consciousness and moral agency," *Int. J. Mach. Conscious.*, vol. 4, pp. 105–129, 2013.

[12] P. S. Churchland, *Braintrust: What neuroscience tells us about morality*. Princeton, NJ: Princeton University Press, 2012.

[13] R. Joyce, *The evolution of morality*. MIT Press, 2006.

[14] R. D. Alexander, *The biology of moral systems*. Hawthorne, NY: Aldine de Gruyter, 1987.

[15] C. S. Sripada and S. Stich, "A framework for the psychology of norms," in *The innate mind (Volume 2: Culture and cognition)*, P. Carruthers, S. Laurence, and S. Stich, Eds. New York, NY: Oxford University Press, 2006, pp. 280–301.

[16] S. Nichols and R. Mallon, "Moral dilemmas and moral rules," *Cognition*, vol. 100, no. 3, pp. 530–542, Jul. 2006.

[17] J. K. Hamlin, "Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core," *Curr. Dir. Psychol. Sci.*, vol. 22, no. 3, pp. 186–193, Jun. 2013.

[18] J. C. Flack and F. B. M. de Waal, "'Any animal whatever'. Darwinian building blocks of morality in monkeys and apes," *J. Conscious. Stud.*, vol. 7, no. 1–2, pp. 1–29, 2000.

[19] J. Voiklis, C. Cusimano, and B. F. Malle, "A social-conceptual map of moral criticism." Unpublished Manuscript, Brown University, 2013.

[20] J. C. Wright and K. Bartsch, "Portraits of early moral sensibility in two children's everyday conversations," *Merrill-Palmer Q.*, vol. 54, no. 1, pp. 56–85, Jan. 2008.

[21] N. L. Powell, S. W. G. Derbyshire, and R. E. Guttentag, "Biases in children's and adults' moral judgments," *J. Exp. Child Psychol.*, vol. 113, no. 1, pp. 186–193, Sep. 2012.

[22] S. Pellizzoni, M. Siegal, and L. Surian, "The contact principle and utilitarian moral judgments in young children," *Dev. Sci.*, vol. 13, no. 2, pp. 265–270, Mar. 2010.

[23] H. M. Wellman and A. T. Phillips, "Developing intentional understandings," in *Intentions and intentionality: Foundations of social cognition*, B. F. Malle, L. J. Moses, and D. A. Baldwin, Eds. Cambridge, MA: The MIT Press, 2001, pp. 125–148.

[24] A. L. Woodward, "Infants selectively encode the goal object of an actor's reach," *Cognition*, vol. 69, no. 1, pp. 1–34, Nov. 1998.

[25] A. N. Meltzoff, "Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children," *Dev. Psychol.*, vol. 31, no. 5, pp. 838–850, 1995.

[26] M. M. Saylor, D. A. Baldwin, J. A. Baird, and J. LaBounty, "Infants' on-line segmentation of dynamic human action," *J. Cogn. Dev.*, vol. 8, no. 1, pp. 113–128, 2007.

[27] L. J. Moses, "Young children's understanding of belief constraints on intention," *Cogn. Dev.*, vol. 8, no. 1, pp. 1–25, Jan. 1993.

[28] H. Wimmer and J. Perner, "Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception," *Cognition*, vol. 13, no. 1, pp. 103–128, Jan. 1983.

[29] J. A. Baird and L. J. Moses, "Do preschoolers appreciate that identical actions may be motivated by different intentions?," *J. Cogn. Dev.*, vol. 2, no. 4, pp. 413–448, Nov. 2001.

[30] P. L. Harris, *Children and emotion: The development of psychological understanding*. New York, NY: Basil Blackwell, 1989.

[31] K. Bartsch and H. M. Wellman, *Children talk about the mind*. New York: Oxford University Press, 1995.

[32] R. T. Beckwith, "The language of emotion, the emotions, and nominalist bootstrapping," in *Children's theories of mind: Mental states and social understanding*, Hillsdale, NJ England: Lawrence Erlbaum Associates, Inc, 1991, pp. 77–95.

[33] B. F. Malle and S. Dickert, "Values," in *The encyclopedia of social psychology*, R. F. Baumeister and K. D. Vohs, Eds. Thousand Oaks, CA: Sage, 2007.

[34] J. Elster, *The cement of society: A study of social order*. New York, NY: Cambridge University Press, 1989.

[35] J. J. A. Van Berkum, B. Holleman, M. Nieuwland, M. Otten, and J. Murre, "Right or wrong? The brain's fast response to morally objectionable statements," *Psychol. Sci.*, vol. 20, no. 9, pp. 1092–1099, 2009.

[36] A. Prince and P. Smolensky, "Optimality: From neural networks to universal grammar," *Science*, vol. 275, no. 5306, pp. 1604–1610, Mar. 1997.

[37] B. F. Malle, S. Guglielmo, and A. E. Monroe, "A theory of blame," *Psychol. Inq.*, 2014.

[38] B. F. Malle, S. Guglielmo, and A. E. Monroe, "Moral, cognitive, and social: The nature of blame," in *Social thinking and interpersonal behavior*, J. P. Forgas, K. Fiedler, and C. Sedikides, Eds. Philadelphia, PA: Psychology Press, 2012, pp. 313–331.

[39] F. Cushman and L. Young, "Patterns of moral judgment derive from nonmoral psychological representations," *Cogn. Sci.*, vol. 35, no. 6, pp. 1052–1075, Aug. 2011.

[40] S. Guglielmo, A. E. Monroe, and B. F. Malle, "At the heart of morality lies folk psychology," *Inq. Interdiscip. J. Philos.*, vol. 52, no. 5, pp. 449–466, 2009.

[41] A. R. Damasio, *Descartes' error: Emotion, reason, and the human brain*. New York, NY: Putnam, 1994.

[42] J. Knobe and B. Fraser, "Causal judgment and moral judgment: Two experiments," in *Moral psychology (Vol. 2): The cognitive science of morality: Intuition and diversity*, vol. 2, Cambridge, MA: MIT Press, 2008, pp. 441–447.

[43] C. L. Harenski, K. A. Harenski, M. S. Shane, and K. A. Kiehl, "Aberrant neural processing of moral violations in criminal psychopaths," *J. Abnorm. Psychol.*, vol. 119, no. 4, pp. 863–874, Nov. 2010.

[44] M. Cima, F. Tonnaer, and M. D. Hauser, "Psychopaths know right from wrong but don't care," *Soc. Cogn. Affect. Neurosci.*, vol. 5, no. 1, pp. 59–67, Mar. 2010.

[45] C. A. Hutcherson and J. J. Gross, "The moral emotions: A social–functionalist account of anger, disgust, and contempt," *J. Pers. Soc. Psychol.*, vol. 100, no. 4, pp. 719–737, Apr. 2011.

[46] L. Kohlberg, *The psychology of moral development: The nature and validity of moral stages*. San Francisco, CA: Harper & Row, 1984.

[47] J. M. Paxton, L. Ungar, and J. D. Greene, "Reflection and reasoning in moral judgment," *Cogn. Sci.*, vol. 36, no. 1, pp. 163–177, Jan. 2012.

[48] J. D. Greene, L. E. Nystrom, A. D. Engell, J. M. Darley, and J. D. Cohen, "The neural bases of cognitive conflict and control in moral judgment," *Neuron*, vol. 44, no. 2, pp. 389–400, Oct. 2004.

[49] G. J. Koop, "An assessment of the temporal dynamics of moral decisions," *Judgm. Decis. Mak.*, vol. 8, no. 5, pp. 527–539, 2013.

[50] E. B. Royzman, G. P. Goodwin, and R. F. Leeman, "When sentimental rules collide: 'Norms with feelings' in the dilemmatic context," *Cognition*, vol. 121, no. 1, pp. 101–114, Oct. 2011.

[51] G. Moretto, E. Làdavas, F. Mattioli, and G. di Pellegrino, "A psychophysiological investigation of moral judgment after ventromedial prefrontal damage," *J. Cogn. Neurosci.*, vol. 22, no. 8, pp. 1888–1899, Aug. 2010.

[52] T. Davis, B. C. Love, and W. Todd Maddox, "Anticipatory emotions in decision tasks: Covert markers of value or attentional processes?," *Cognition*, vol. 112, no. 1, pp. 195–200, Jul. 2009.

[53] W. Wallach, S. Franklin, and C. Allen, "A conceptual and computational model of moral decision making in human and artificial agents," *Top. Cogn. Sci.*, vol. 2, no. 3, pp. 454–485, 2010.

[54] M. L. Hoffman, "Empathy and prosocial behavior," in *Handbook of emotions*, 3rd ed., M. Lewis, J. M. Haviland-Jones, and L. F. Barrett, Eds. New York, NY: Guilford Press, 2008, pp. 440–455.

[55] O. FeldmanHall, T. Dalgleish, R. Thompson, D. Evans, S. Schweizer, and D. Mobbs, "Differential neural circuitry and self-interest in real vs hypothetical moral decisions," *Soc. Cogn. Affect. Neurosci.*, vol. 7, no. 7, pp. 743–751, Oct. 2012.

[56] M. Scheutz, "The inherent dangers of unidirectional emotional bonds between humans and social robots," in *Anthology on Robo-Ethics*, P. Lin, G. Bekey, and K. Abney, Eds. Cambridge, MA: MIT Press, 2012, pp. 205–221.

[57] M. Scheutz, "The affect dilemma for artificial agents: should we develop affective artificial agents?," *IEEE Trans. Affect. Comput.*, vol. 3, no. 4, pp. 424–433, 2012.

[58] C. R. Sunstein, "Moral heuristics," *Behav. Brain Sci.*, vol. 28, no. 4, pp. 531–573, Aug. 2005.

[59] A. E. Monroe, K. D. Dillon, and B. F. Malle, *Developing a model of the folk concept of free will and its impact on moral judgment.* Paper presented at the symposium on Big Questions in Free Will, Florida State University, Tallahasee, Florida., 2013.

[60] A. E. Monroe and B. F. Malle, "From uncaused will to conscious choice: The need to study, not speculate about people's folk concept of free will," *Rev. Philos. Psychol.*, vol. 1, no. 2, pp. 211–224, 2010.

[61] P. H. Kahn, Jr., T. Kanda, H. Ishiguro, B. T. Gill, J. H. Ruckert, S. Shen, H. E. Gary, A. L. Reichert, N. G. Freier, and R. L. Severson, "Do people hold a humanoid robot morally accountable for the harm it causes?," in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, 2012, pp. 33–40.

[62] F. Cushman, "The functional design of punishment and the psychology of learning," in *Psychological and environmental foundations of cooperation.*, vol. 2, R. Joyce, K. Sterelny, B. Calcott, and B. Fraser, Eds. Cambridge, MA: MIT Press, 2013.

[63] I. Dersley and A. Wootton, "Complaint sequences within antagonistic argument," *Res. Lang. Soc. Interact.*, vol. 33, no. 4, pp. 375–406, Oct. 2000.

[64] V. Traverso, "The dilemmas of third-party complaints in conversation between friends," *J. Pragmat.*, vol. 41, no. 12, pp. 2385–2399, Dec. 2009.

[65] M. U. Walker, *Moral repair: Reconstructing moral relations after wrongdoing*. New York, NY: Cambridge University Press, 2006.

[66] M. McKenna, "Directed blame and conversation," in *Blame: Its nature and norms*, New York, NY: Oxford University Press, 2012, pp. 119–140.

[67] B. F. Malle, *How the mind explains behavior: Folk explanations, meaning, and social interaction*. Cambridge, MA: MIT Press, 2004.

[68] D. J. Hilton, "Causal explanation: From social perception to knowledge-based causal attribution," in *Social psychology: Handbook of basic principles*, 2nd ed., A. W. Kruglanski and E. T. Higgins, Eds. New York, NY: Guilford Press, 2007, pp. 232–253.

[69] P. Bello, "Cognitive foundations for a computational theory of mindreading," *Adv. Cogn. Syst.*, vol. 1, pp. 59–72, 2012.

[70] R. E. Nisbett and T. D. Wilson, "Telling more than we know: Verbal reports on mental processes," *Psychol. Rev.*, vol. 84, pp. 231–259, 1977.

[71] B. F. Malle, "Time to give up the dogmas of attribution: A new theory of behavior explanation," in *Advances of Experimental Social Psychology*, vol. 44, M. P. Zanna and J. M. Olson, Eds. San Diego, CA: Academic Press, 2011, pp. 297–352.