Ethics Inf Technol (2016) 18:243–256
DOI 10.1007/s10676-015-9367-8

ORIGINAL PAPER

Integrating robot ethics and machine morality: the study and design of moral competence in robots

Bertram F. Malle[1]

Published online: 2 July 2015

Abstract Robot ethics encompasses ethical questions about how humans should design, deploy, and treat robots; machine morality encompasses questions about what moral capacities a robot should have and how these capacities could be computationally implemented. Publications on both of these topics have doubled twice in the past 10 years but have often remained separate from one another. In an attempt to better integrate the two, I offer a framework for what a morally competent robot would look like (normally considered machine morality) and discuss a number of ethical questions about the design, use, and treatment of such moral robots in society (normally considered robot ethics). Instead of searching for a fixed set of criteria of a robot's moral competence I identify the multiple elements that make up human moral competence and probe the possibility of designing robots that have one or more of these human elements, which include: moral vocabulary; a system of norms; moral cognition and affect; moral decision making and action; moral communication. Juxtaposing empirical research, philosophical debates, and computational challenges, this article adopts an optimistic perspective: if robotic design truly commits to building morally competent robots, then those robots could be trustworthy and productive partners, caretakers, educators, and members of the human community. Moral competence does not resolve all ethical concerns over robots in society, but it may be a prerequisite to resolve at least some of them.

Keywords Social cognition · Moral cognition · Human-robot interaction · Moral psychology · Social robotics

✉ Bertram F. Malle
bfmalle@brown.edu

[1] Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, 190 Thayer St., Providence, RI 02912, USA

Introduction

The rise of robot ethics

The design and construction of artificial intelligent machines has seen steady growth in the past 50 years, but ethical questions about the human commitment to this endeavor have lagged behind. In a seminal book on social robotics (Breazeal 2002), the preface notes that the book raises ethical questions, but the words *moral* or *ethical* do not appear in other sections of the book. In 2007, a historical article on the evolution of robotics research did not mention these words either (Garcia et al. 2007). And a recent exposition on the state of the art in human-robot interaction research in Asia (Veloso et al. 2012) reports no research or development on moral capacities in robots.

But the tide is turning. In 2002, the first "robot ethics" workshop took place as part of the IEEE-Robotics and Automation conference, and in 2004 a Technical Committee on Robot Ethics was founded by that same IEEE section (see http://www.ieee-ras.org/robot-ethics). And whereas 41 academic publications on the topic appeared until 2004, the number more than doubled between 2005 and 2009 (88), and it doubled again since 2010 (179 as of December 2014).[1] So what are the "increasingly urgent ethical issues raised by the rapidly advancing robotics technology" (Technical Committee on Robot Ethics)?

[1] Numbers are derived from an EBSCO database search using "robot* and ethic*" as subject search terms and restricting to contributions in journals and books.

Two questions of robot ethics

Two classes of questions fall under the larger theme of ethics and robots:

1. Ethical questions about how *humans* should design, deploy, and treat robots—often termed (in analogy to bioethics or environmental ethics) *robot ethics* (Veruggio et al. 2011);
2. Questions about what moral capacities a *robot* should have and how these capacities could be implemented—often called "machine morality" (Sullins 2011) or "machine ethics" (Moor 2006).

Robot ethics features such topics as ethical design (Wynsberghe 2013), values of implementation (Hofmann 2013), and considerations of robot rights (Petersen 2007). Machine morality features such topics as criteria for moral agency (Floridi and Sanders 2004), justification for lethal military robots (Arkin 2009), and mathematical proofs for moral reasoning (Bringsjord et al. 2006).

These two classes of questions are often treated separately. Nourbakhsh (2013), in projecting the future of robotics, discusses the importance of training roboticists in ethics but not the need for robots themselves to be ethical. Anderson and Anderson's (2011) volume focuses on numerous capacities that robots might need to be able to make moral decisions but scantly discuss the ethical challenges that roboticists must consider in doing their society-altering work. Several authors have promoted an integration of the two kinds of questions (e.g., Asaro 2006; Wallach 2010), well illustrated in a recent edited volume (Lin et al. 2012) that featured articles on the design of ethical robots, on the applied ethics of robots in military, medicine, and companionship, as well as on legal and political questions of robots in society.[2] Such integration is justified because the topics of robot ethics and moral machines are deeply related. Attempts to develop morality in machines raise many ethical questions (Wallach and Allen 2008), and ethical concerns about robot design (Scheutz and Crowell 2007) may be addressed by developing adequate moral capacities in robots.

Specific applications of robotic technology best illustrate how intertwined the two sets of questions are. In discussions about military deployments, what ethical reasons would we accept for using robots in lethal ways, and would certain properties of these robots provide such ethical reasons? Some have argued that reliable and logically consistent obedience to military and international humanitarian laws render robots superior to humans, who routinely violate these laws (Arkin 2009). Outside the military domain, it may be unethical for humans to design and deploy a care robot that is ignorant of basic social norms and lacks the capacity to conform to them. Some situations may even pose genuine dilemmas, such as when a cancer patient begs a robot nurse for more morphine but the supervisory doctor is not reachable to approve the request. In search and rescue, too, difficult moral decisions will arise: Which faintly crying voice from the earthquake rubble should the rescue robot follow—the child's or the older adult's? Moral dilemmas such as these are attractive test cases for the success of a moral robot, and I will return to associated questions on moral dilemmas later.

In sum, any robot that collaborates with, supports, or cares for humans—in short, a social robot—poses serious ethical challenges to the human design and deployment of such robots, and one of the most important challenges is to create a level of moral competence in these robots that is adequate to the application at hand. This, then, offers a pivotal integration point of robot ethics and machine morality: how a robot's moral competence could help resolve some of the ethical concerns about robots in society and perhaps even guide us to new opportunities of how robots could make valuable contributions to society.

But is moral competence in robots even possible? And what is moral competence to begin with?

From moral agency to moral competence

Previous discussions of machine morality have often focused on whether a robot could be a *moral agent*—typically understood as an entity that can act according to what is right and wrong (e.g., DeBaets 2014; Floridi and Sanders 2004) or could be held responsible for its actions (Parthemore and Whitby 2013). Many challenges await the analysis of moral agency, however, if the goal is something like a universally agreed-upon definition, with clean criteria against which a robot can then be measured. Scholars have suggested a variety of criteria for being an *agent*, including embodiment, consciousness, soul, free will—many of which raise more conceptual questions than they are intended to answer (Gunkel 2014). Similarly, asking what makes an agent *moral* leads to perhaps even more difficult problems, either in choosing an appropriate [meta-]ethical "system" (e.g., Anderson and Anderson 2011; Grau 2011; Powers 2006) or in explaining what the "right" or "good" is that a moral agent is said to cause or strive for (DeBaets 2014).

I suggest a different approach: to identify instead the numerous capacities that make up human moral competence, not as fixed conditions for robot moral competence but as an orienting framework. What we need to examine is not one "true" moral competence but the competences that people

[2] This volume appeared as part of MIT Press's series on "Intelligent robotics and autonomous agents." Notably, ethics was addressed as the 19th topic in the series, 15 years after the series commenced.

expect of one another. For people will expect at least some of these competences of social robots—any robots with which they are willing to form social relationships. This requires us, first, to understand what the elements of human moral competence are and, second, to learn to implement at least some of these elements in computational architectures and physical machines. Third, we must continuously gather empirical evidence to ensure that the emerging morally competent robots are in fact suitable for and accepted as social partners (Fridin 2014). In this way, machines receive moral consideration within the context of their broader social capacities and within the relations humans are willing to form with them (Coeckelbergh 2010).

An approach that identifies *elements* of human moral competence has the advantage that we no longer need to make tough decisions about whether robots do or do not meet a particular criterion to count as "fully" moral agents (Allen 2011; Moor 2006). Robots may be designed to have some competences but not others; there may even be stages or degrees of having a certain competence just as we are comfortable with ascribing children certain competences in stages or degrees (Wallach and Allen 2008). Moreover, different applications (e.g., for health and social assistance or for safety and security) may call for implementing robots with different competences (Asaro 2006), and what is an adequately moral agent in one application may look different from one in another application. On this approach, research and design into moral robots is a dynamic and adjustable process (Powers 2011), guided by scientific research that has no pressures to sell products but is subject to societal ethical deliberations about design, distribution, and beneficial deployments of robots. As a result, this exploration into the fundamental *moral machine* question—what an artificial moral agent could and should be like—remains closely tied to questions of *robot ethics,* such as society's debates and decisions about awarding certain rights and statuses to machines (Gunkel 2014). How we should treat robots and what rights and duties apply to them will critically depend on their moral competences, and different competences may call for different rights and duties (Calverley 2006).

I now offer a more detailed analysis of moral competence—originally developed in Malle and Scheutz (2014) and Scheutz and Malle (2014)—that is grounded in scientific research on human moral psychology and sketches the prospects for such competences to be developed in robotic agents.

Elements of moral competence

A competence is an aptitude, a qualification, a dispositional capacity to deal adequately with certain tasks. What tasks are moral? Uncontroversially, moral competence must deal with the task of *moral decision making and action*. From Aristotle to Kant to Kohlberg, morality has been about "doing the right thing." But there is quite a bit more.

In psychology, *moral cognition* has been the primary focus of recent theoretical and experimental work, examining such phenomena as judgments of permissibility, wrongness, blame, and the role of reasoning and emotion in those judgments (Alicke 2000; Cushman 2008; Gray et al. 2012; Greene et al. 2004; Haidt 2001; Knobe 2010; Malle et al. 2014).

Further, psychologists and sociologists have studied *moral communication,* including such phenomena as negotiating blame through justification and excuses, apology, and forgiveness (Antaki 1994; McCullough et al. 2013; Semin and Manstead 1983; Tedeschi and Reiss 1981; Weiner 1995).

Finally, these three elements of moral competence require two basic elements to begin with: a *norm system* that is somehow represented in the moral agent's mind; and a *moral vocabulary* that allows the agent to represent those norms, use them in judgments and decisions, and communicate about them. These, then, are five key elements of moral competence:

- A moral vocabulary
- A system of norms
- Moral cognition and affect
- Moral decision making and action
- Moral communication

Space constraints do not allow me to take one step further back and discuss exactly what foundational cognitive and computational capacities each of these elements of moral competence presuppose. Some of them will become apparent soon: language, perception, causal reasoning, self-monitoring, and advanced learning; but their complex dependencies must remain unexamined.

Moral vocabulary

Some rudimentary moral capacities may operate without language, such as the recognition of prototypically prosocial and antisocial behaviors (Hamlin 2013) or foundations for moral action in empathy and reciprocity (Flack and de Waal 2000). But morally competent adults need a vocabulary to conceptually represent the norms of their community and to learn, teach, and deliberate about these norms. They also need a vocabulary to express and instantiate various moral practices—to blame or forgive another's transgressions, to justify and excuse their own behavior, and to negotiate the priority of one norm over another. A moral vocabulary thus has three major domains:

1. Vocabulary of *norms and their properties* ("fair," "virtuous," "reciprocity," "honesty," "obligation," "prohibited," "ought to," etc);
2. Vocabulary of *norm violations* ("wrong," "culpable," "reckless," "thief," and also "intentional," "knowingly," etc);
3. Vocabulary of *responses to violations* ("blame," "reprimand," "excuse," "forgiveness," etc).

For a morally competent robot, it will be important to detect moral vocabulary when used by human communicators, because it serves as a reliable indicator of morally significant situations and may provide opportunities to acquire more refined vocabulary. Robots must also themselves master some moral expressions in the above three domains. Of course, the human moral vocabulary is extremely rich, so a realistic start would be to "seed" the robot's moral vocabulary with prototype terms and treat a wider class of words as reasonably approximated by the prototypes.

Here is one example of how to empirically derive prototype terms for robots. In the domain of *responses to violation*, Voiklis et al. (2014) recently uncovered a two-dimensional structure that underlies differences among verbs of moral criticism. A sample of 300 participants were asked to assess 28 such verbs on a total of 12 properties. Each verb was framed as a description of a generic social act ("He [verbed] her for the bad thing she had done"), and each participant assessed all 28 verbs on one property, including "How intense was the emotion he [the moral critic] felt?", "How socially acceptable was what he did?, "How bad was what she [the offender] had done?", and "Was this more like a private thought or more like a public action?". Voiklis et al. (2014) found high consensus among participants in differentiating the verbs on each property and documented that the 12 properties could be reduced to two fundamental dimensions. People distinguished acts of moral criticism by evaluating their intensity (PC1 in Fig. 1) and whether they were directed to the transgressor or not (PC2 in Fig. 1). Selecting the best-differentiated verbs from each quadrant of this two-dimensional space provides four sets of prototypical terms (circled in Fig. 1) that a robot should master.

Moral norms

Any analysis of morality, and therefore of moral competence, must fundamentally be anchored in the concept of norms. Morality is at its heart a system of norms that a community adopts to regulate individual community members' behaviors and thus bring them in line with community interests. Yet, even though having a norm system is an essential characteristic of morality, we know quite little about how norms are acquired, how they are represented in the mind, and what makes them both general and context-sensitive. Such knowledge will be needed if we want to design effective moral robots.

Norm acquisition

Evidence is limited on the human development of norms, but data on children's early use of moral language suggest that children are rarely exposed to abstract rules but rather hear and express concrete moral judgments (Wright and Bartsch 2008): "that's not nice! That was naughty!"; "He did something wrong." As talented statistical pattern learners (Thiessen et al. 2013), infants and toddlers are likely to store norm-conforming patterns and are sensitive to their violations. And even though initially they may not understand the significance of a pattern or its violation, adults' reactions to the norm violation provide the "social reference" for its meaning (Emde 1992): a strong reaction will leave a deeper memory trace and deserves a flag of importance. A reasonable heuristic may then be that flagged norms are candidates for what later will be called "moral" norms, which are stricter and cannot easily be revoked by authority (Turiel 1983).

After learning relatively concrete norms embedded within observable patterns, children are able to induce more general rules from specific instances, such as "bombs hurt people" (Wright and Bartsch 2008, p. 77), and even more abstract principles such as the act-omission distinction (Powell et al. 2012), obviously receiving help from community members' explicit teaching of rules (Csibra and Gergely 2009). Learning a hierarchy of norms, from very concrete ("When a person stretches out their sideways turned hand towards you, grab it and shake it") to very abstract ("Respect other people"), is a substantial challenge, and it may take into young adulthood to properly link open classes of behavior to the most abstract of norms: human values (Malle and Dickert 2007). Preprogramming such a network in a robot appears to be a pointless task, for many reasons: there seems to be an enormous number of norms, they are activated in highly context-specific ways, the network must be subtly adjusted each time norms come in conflict with one another, and unfamiliar tasks, situations, and interactions demand the learning of new norms. A more promising direction is to mix unsupervised and supervised learning, "practice" through constant browsing of existing data (e.g., novels, conversations, movies) along with feedback about inferences (e.g., through crowdsourcing of "inquiries" the robot can make), and teaching through interaction. All this would have to be expanded over time—to mimic the enormous amount of practice and repetition from which human learners benefit.

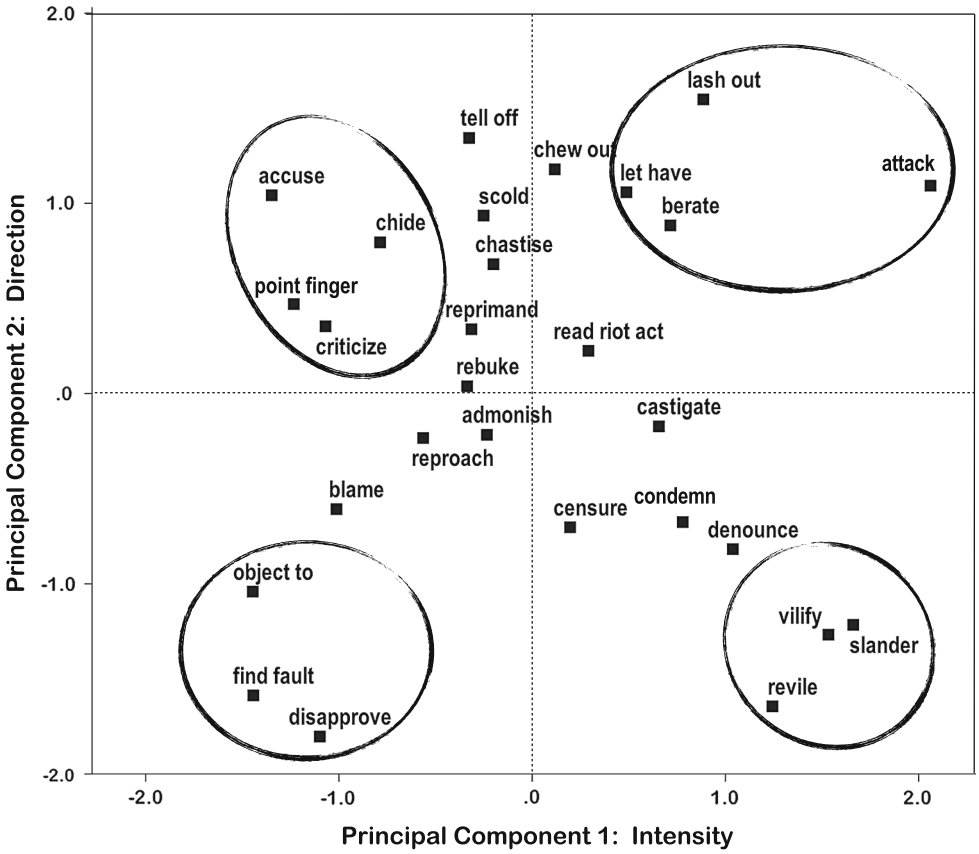

Fig. 1 Verbs of moral criticism in two-dimensional feature space (defined by intensity and communicative direction), with groups of prototypical verbs marked within each quadrant

Norm representation

How might a flexible, interconnected network of concrete to abstract norms be represented in the mind? A first possibility is that norms are cognitively similar to goals, which can be represented quite well in robotic architectures (Talamadupula et al. 2011). But are norms and goals really the same? A goal represents a desirable state and comes with a motivation to mobilize actions believed to realize the state. Norms are more than that—they involve representations of other people's behavior and expectations, sometimes even mutual commitments, about this state (Bicchieri 2006; Tomasello and Vaish 2013). Moreover, the state is often one that is not intrinsically desirable to the individual. Indeed, the best evidence for children's grasp of and adherence to norms stems from situations in which they actually would prefer to act against the norm (e.g., Warneken et al. 2011). A simple goal-based action control system would not turn humans into norm-following creatures. But could a goal-based system be sufficient for moral robots? I will return to such a possibility in the Sect. "Moral Decision Making and Action".

A second possibility for how norms are represented in the mind is that norms are embedded in the semantic-linguistic networks people build through normal language acquisition. As children learn their community's language, they also learn, with every labeled object, action, or state, what is disallowed, allowed, and obligated with respect to the object, action, or state. When those labels are uttered in a sentence, or when their referents are observed, the norms get activated along with the associated linguistic representations. This would provide a tight associative network, constantly strengthened through physical interactions with the world and communicative interactions with other community members—interactions that the developing human mind experiences plentifully. This would then help explain why norms can be rapidly activated (perhaps even within a few hundred milliseconds; Van Berkum et al. 2009) and often operate in highly context-specific ways, from an early age on (Wyman et al. 2009).

When designing a morally competent robot, the problems of norm acquisition and norm representation present serious challenges. But if norms are kinds of representations that are connected in some flexible network and activated by perceived features of the environment, then there is no principled reason why they could not be implemented in a computational robotic system. However, for a robot to build up a network of this type it might have to receive similar learning opportunities as human children receive, be exposed to repeated physical and communicative interactions in a human community, perhaps even

alongside human children (Tanaka et al. 2007). These kinds of robots would be *raised*, not programmed. This, in turn, would pose a number of serious and fascinating ethical questions about the obligations of those who raise such robots, about the rights of such developing community members, and about the possible unintended consequences of allowing a powerful learning machine access to nearly unlimited sources of data. Such questions have already entered the popular entertainment domain, in more convincing ways (the TV series *Extant*; Fisher et al. 2014) or less convincing ways (the movie *Chappie*; Blomkamp et al. 2015).

Moral cognition and affect

Human moral cognition encompasses processes of perception and judgment that allow people to detect and evaluate norm-violating events and respond to the norm violator. There are two kinds of moral jugdments we must distinguish (Malle et al. 2014).

At the basic level of moral cognition, people's well-practiced norm network allows them to quickly detect prototypical norm violations, such as direct physical harm to another person. Once detected, such events can trigger rapid evaluations (Luo et al. 2006; Van Berkum et al. 2009), typically in the form of judging the event as bad, good, wrong, or (im)permissible.

These initial evaluations of norm violating *events* can trigger more complex moral information processing about the *agent* who committed the violation, most prominently judgments of blame (Alicke 2000; Coates and Tognazzini 2012; Malle et al. 2014; Shaver 1985). These judgments take into account the agent's specific causal involvement in the norm-violating event (e.g., directly causing it, partially contributing to it, passively allowing it) and whether the agent violated the norm intentionally or unintentionally. But people don't stop here. A unique feature of human blame judgments is that intentional and unintentional violations trigger distinct subsequent processing steps (Malle et al. 2014). When perceiving a violation as an intentional action, people search for the agent's reasons for performing the action ("motive" in legal context); when perceiving a violation as an unintentional event, people search for preventability information—whether the agent should have and could have prevented this event. All together, it is quite clear that human moral cognition is not a specialized "module" but builds on ordinary cognition of social events embedded in a norm system (Cushman and Young 2011; Guglielmo et al. 2009).

What does it take for a robot to engage in moral cognition of both events and agents? For event evaluations, the robot needs to be able to segment visual and verbal event streams and identify those events (behaviors and states) that violate social or moral norms. To avoid comparisons of every identified event against every stored norm, the problem needs to be constrained. As discussed earlier, norms may be activated locally, such that specific physical and social contexts trigger a manageable set of norms. The segmented events can then be compared to this smaller set of norms. Over time, the features that are used to identify events may also become features that trigger specific norms; the norm-based event evaluation process would then be integrated into ordinary event perception. To further form agent-directed judgments such as blame, a robot would need capacities for causal reasoning over segmented events, social-cognitive inferences from behavior to determine intentionality and the agent's reasons, and counterfactual reasoning to determine preventability.

Where in all this is affect? There is little doubt that detecting 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 what this affective response sets in motion is unclear: Marking that something important occurred? Strengthening motivation to find the cause of the bad event (Knobe and Fraser 2008)? Biasing the search for evidence that allows the perceiver to blame somebody (Alicke 2000)? And what do we make of the fact that people can make moral judgments without much affect at all (Harenski et al. 2010) or that moral emotions such as anger or resentment require specific cognitive processes (Hutcherson and Gross 2011)? Clearly, affective phenomena can influence moral judgments, often accompany moral judgments, and probably facilitate learning and enforcing moral norms. But there is little evidence for the claim that affective phenomena are *necessary* or *constitutive* of those judgments (Avramova and Inbar 2013; Huebner et al. 2009). If affect is not necessary or constitutive of moral judgments, then even an affectless robot could be competent to make moral judgments. Because the precise roles of affective phenomena in moral judgment have not yet been determined (Malle et al. 2014), the final word on this possibility awaits more research. For now, if we develop computational models of blame that match human judgments in their sensitivity to the critical informational features (severity of norm violation, causality, intentionality, etc.), we would take a significant step forward in developing a robot competent in moral cognition.

Moral decision making and action

Determinants

Perhaps the most prominent component of human moral competence is decision making and action—what makes people behave morally. Nonmoral behavior is guided by a

variety of psychological factors, including affective states and personality dispositions, automatic imitation and group pressure, heuristics and reasoned choice (Fiske and Taylor 2008; Gilovich et al. 2013); and those same factors of course influence moral behavior. Communities generally welcome any combination of these factors as determinants of norm-conforming behavior, but personality dispositions (especially virtues) and reasoned choice likely stand out as instilling the greatest trust in the agent's future moral behavior, presumably because these factors are most reliable in bringing about the same behavior in the face of environmental changes. Conversely, the greatest threat to a community are agents who act counter-normatively because of personality dispositions and reasoned choice, and behavior caused by these factors also receive the greatest amount of blame and punishment (Malle et al. 2014).

Free will?

Some skeptics of the possibility of moral competence in robots assume that such reasoned choice presupposes a rather strong capacity: some kind of nondeterministic "free will" (Bringsjord 2009; Johnson and Axinn 2013). But most ordinary people seem to understand free will as nothing more than the capacity for choice and intentional action execution that is relatively unobstructed by constraints (Monroe and Malle 2010, 2014). Whatever metaphysical worries scholars might have about the possibility of nondeterministic free will does not seem to be of great concern to ordinary moral perceivers—who are of course the ones who will interact with future robots. If a robot acquires and uses knowledge about the world to guide its actions in line with its goals, it effectively displays a capacity for choice and intentional action (Powers 2011). When such a robot commits a norm violation, people readily assign blame to it—in simulated scenarios (Malle et al. 2015; Monroe et al. 2014) and actual interaction (Kahn et al. 2012). Blame is pedagogical in that it provides the norm violator with reasons to not violate the norm again. Thus blame would regulate robot behavior only if the robot could learn and take the received blame into account in its next choice of action. This sort of capacity to learn and adjust one's choices is needed; metaphysical free will is not.

Selfishness and empathy

One characteristic factor in human moral decision making is the frequent tension between social-moral norms and the agent's own goals, and this tension introduces two unique psychological processes that guide moral action: empathy and self-regulation (Eisenberg 2000; Hoffman 2008). Both are designed to favor communal values over selfish interest, one through an affective mechanism (feeling the pain of victims, or those in need, motivates prosociality), the other through a more cognitive mechanism (inhibiting or sublimating one's selfish desires enables prosociality). 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 (Grau 2011). Such a robot would not have "temptation" to be selfish and to ignore others' needs, so empathy and self-regulation would be far less important. Moreover, the tension between selfish and community-serving motives is one of the main drivers for a social *norm* system that guides individuals' behavior beyond their own *goals* (which by themselves would too easily lead to free riding). In a robot without selfish desires, goals could perhaps stand in for norms after all, because the goal contents would be directly norm-conforming, rather than be tweaked and coaxed toward prosociality by the complex social features of norms (i.e., mutual expectation, commitment, and fear of enforcement).

Empathy, however, has at least one function in human life that cannot be so easily replaced by an unselfish goal system: to create trust in others. A person who responds empathically to the plight of others (especially the ones he may have victimized) is trusted far more than someone who coldly assesses a moral situation (even if the assessment is technically correct). A robot may be able to build such trust if it can perceive the plight and pain of others and express its perception in a sympathetic language (even if it doesn't genuinely "feel" the plight and pain). These communications, however, should not be mere verbal scripts or deceptive attempts to coax the human's trust. For advanced human–machine interactions to succeed the robot must somehow be able to demonstrate that it *values* things (Littman 2001; Scheutz 2012), that it *cares* about certain outcomes (Wynsberghe 2013). To determine what it means for a machine to care about something (not just as a pretense) is a serious challenge. Perhaps we need not invent a unique computational structure (a "caring state") but rather analyze carefully what kinds of behaviors people consider diagnostic of *caring* (an approach that Ryle 1949, suggested for many mental concepts), such as willingness to attend to, prioritize, invest in, help, and so on. When people form relationships with robots, those actions may convincingly and justifiably represent a caring attitude that is not deceptive.

Skepticism about choice capacity

So far I have assumed that artificial agents can, within a powerful cognitive architecture, have choice capacity—at least the kind of choice capacity that humans find credible. But from a technical standpoint one might be skeptical of

an artificial agent's capacity to make anything but trivial decisions because of the well-known frame problem in AI (Ford and Hayes 1991; Pylyshyn 1987)—the difficulty (in one formulation) of choosing among possible actions by computing all relevant consequences of those actions. However, humans do not make decisions by computing all relevant consequences of their actions either; they use the limited information they have and, after acting, learn from their mistakes. Out of such experiences they develop habits, scripts, and predictive models that provide good approximations of the right course of action in circumstances recognized as similar to the original or prototypical case (Schank and Abelson 1977; Wolpert and Flanagan 2001). In doing so they take advantage of the high degree of predictability of the physical world and even a decent degree of predictability of the social world because of its extensive structure of norms, roles, and others' habits and scripts. More involved deliberation and choice is then engaged when the case at hand noticeably deviates from the prototypical case. This approach of approximation, prediction, and selective attention to deviations may be harnessed for robots as well.

Another reason for skepticism may be the simultaneous design of choice capacity and fundamental unselfishness. One might object that a robot that cannot help but make moral decisions (because it is programmed to be unselfishly moral) *does not* have a choice and therefore is not a morally competent actor. This objection can be met in at least two ways. First, the kind of robot envisioned here could not possibly be programmed to act morally for all possible futures. It would have to be equipped with a number of guiding norms to start with but also learn many new norms, and it could therefore fail to act morally out of ignorance and, with feedback, do better next time. Second, some situations pose a decision problem in which not all relevant norms can be satisfied simultaneously. Such *moral dilemmas* require a genuine choice between imperfect options, and very often each of the options can be morally justified by reference to some accepted norms. So from one perspective the robot would be making the wrong moral decision, from another perspective it would be making the right moral decision. Either way, it would make a decision (Scheutz et al. 2015).

Decision dilemmas

The topic of moral dilemmas in the context of machine morality is in fact ripe for research. The research literature on human moral psychology has devoted significant efforts to understanding how humans handle moral dilemmas (Kohlberg 1984; Mikhail 2007; Paxton et al. 2012). From an HRI perspective, a critical question is how humans want robots to handle such dilemmas. Exactly the same way as humans do? Or do people impose different permissions and obligations on robots?

A few authors have proposed thought experiments for self-driving cars that follow the structure of moral dilemmas (Lin 2013; Millar 2014), and a recent reader poll assessed which norm trade-offs people preferred for a car in one such thought experiment (Open Roboethics Initiative 2014a, b). However, the poll did not strictly meet the definition of a moral dilemma and also had no comparison data on a human driver. A recent study by Malle et al. (2015) provides the first systematic comparison of how people apply moral judgments to both human and robotic agents that face the identical moral dilemma.

Participants read a brief description of a dilemma in which four people are about to die unless the protagonist takes an action that saves the four but kills another person (a "utilitarian" choice). Across two experiments (total $N = 316$), participants were asked three different moral judgments (which are often conflated in the literature):

1. Before learning the agent's decision, participants indicated whether the utilitarian action was *permissible*; or
2. After learning the agent's decision, they indicated whether it was *morally wrong* that the agent either took the action or refrained from taking it (a between-subjects manipulation); and
3. After answering one of the first two questions, they indicated how much *blame* the person deserves for either taking the action or refraining from it.

The results showed that, first, the human agent's sacrifice was initially seen as permissible by 65 % of respondents whereas the robotic agent's sacrifice was seen as permissible by 78 % of respondents. Second, 30 % of people found it morally wrong if the robot *refrained* from taking the utilitarian action but only 13 % if it took the action, whereas 49 % of people found it morally wrong if the human agent *did* take the utilitarian action but only 15 % if he refrained from taking it. Third, human agents were blamed far more for choosing this option ($M = 53.8$ on a 100-point scale, averaged across two experiments) than for doing nothing ($M = 18.2$), whereas the robot received only slightly more blame for choosing the utilitarian option ($M = 40.6$) than for doing nothing ($M = 31.8$). In sum, people found it more norm-violating for a human agent to intervene and sacrifice one for the good of four (compared to doing nothing), whereas they found it more acceptable for a robot to intervene (but they still blamed it if it did).

More research is needed, not only on the questions raised by these studies but on the general issue of how people might apply the same or different moral norms to robots and humans. The main message is that we cannot

wait until robots actually occupy societal positions in which they face challenging moral decisions; studying people's expectations and responses at that point in the future would be too late. Behavioral, cognitive, and human-robot interaction research must anticipate some of these possibly unique responses to robots' moral decisions and use the results of this research to design acceptable robots for the future.

Moral communication

As important as the cognitive tools are that enable moral judgment and moral decision making, they are not sufficient to achieve the socially most important function of morality: to regulate each other's behavior. For that, moral communication is needed. People often express their moral judgments either to the alleged offender or to another community member (Dersley and Wootton 2000; Traverso 2009); the alleged offender may contest the charges or explain the action in question (Antaki 1994); and conversation or compensation may be needed to repair social estrangement after a norm violation (McKenna 2012; Walker 2006).

Expressing one's moral judgments of others' behavior is typically licensed by being an appropriate *target* of moral judgment. In fact, we expect morally competent agents to criticize others' norm violations (Fehr and Fischbacher 2004; Heath 2001). This should in principle be possible for a robot as well, if its moral cognition capacity and its natural language skills are well developed. But in ordinary social life, those low in status are not always free to voice their moral criticism, and presumably people will regard robots as low in status. So robots will need to be aware of the *norms of blaming* (Malle et al. 2014) and sometimes refrain from social acts of moral criticism (even if they were able to).

Human moral criticism is suppressed not just by low status but also by intense ingroup loyalty. For example, a serious problem in the military is that soldiers within a unit are reluctant to report a fellow soldier's violations, including human rights violations (MHAT-IV 2006). One might think that having a robot be part of the unit will increase the likelihood of reporting such violations. But if the original reason why soldiers don't report such violations is the pressure for ingroup loyalty, then a robot that reports violations is likely to be rejected by the unit because it breaches norms of loyalty. Robots may have to first earn a level of trust that licenses them to monitor and enforce norms. Then they would need to explicitly declare their obligation to report norm violations, using this communication as a warning and a reminder of the applicable norms.

On the positive side, the anger and outrage that accompanies many human expressions of moral criticism can be omitted from a robot. This may be particularly important when the robot is partnered with a human in a collaborative task, such as with a police officer on patrol or with a teacher in a classroom. By pointing out (inaudibly to others) a looming violation and not showing the kind of affect that would normally make a human partner defensive, the robot's anticipatory moral criticism may be effective.

Another communicative demand on moral competence is to reject immoral requests and unlawful orders (Kibble 2012). Though ideally we would want everybody to reject such requests and orders, in reality significant peer and power pressures persuade people not to protest but rather to obey, against their better judgment (Milgram 1963). A robot would not need to be susceptible to such pressures. It would lack our fearful, sometimes desperate human need to please others, lest they sever their social ties with us. Polite, fearless, and insistent refusal may be a powerful antidote to issuers of unreasonable commands. Moreover, it provides a model that human partners can emulate and use as justification for their defiance of group pressures. Robots may give whistle-blowers courage.

Besides expressing moral judgments, moral competence also requires the ability to explain norm-violating behaviors—typically one's own, but sometimes others'. This capacity is directly derived from the ability to explain behaviors in general, which is relatively well understood in psychology (Hilton 2007; Malle et al. 2014) but scarcely studied in robotics (Lomas et al. 2012). Importantly, ordinary 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 (Malle 1999). 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 (Malle et al. 2014). In addition, and unique to the moral domain, unintentional moral violations are assessed by counterfactuals: what the person *could* (and *should*) have done differently to prevent the negative event. When moral perceivers say, "You could have done otherwise," either to a human or a robot agent, they invite a consideration of options that were available at the time of acting but that the agent ignored or valued differently—and that the moral perceivers expect the agent to take into account in the future. As a result, moral criticism involves the cognitive process of 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 (Bello 2012), but designing the details of such an architecture will be challenging, especially in novel circumstances that involve significant uncertainty over elements in the simulated scenarios. Here again, robots will be successful in this task only with repeated exposure to variants of physical and social events and constant updating of stored representations of those events. The database of simulatable worlds should be acquired through learning, not hard coded into the system.

In addition to causal analysis and simulation, explanations of one's *own* intentional actions require access to one's own reasoning en route to action and accurate memory for this reasoning.[3] Some have famously doubted this capacity in humans (Nisbett and Wilson 1977), but these doubts dissipate in the case of reasons for action (Malle 2004, 2011). In any case, it should be possible to design a robot that has reliable access to its own reasoning via a system of meta-reasoning (Brachman 2002; Cox 2011). The challenge is that the robot must articulate its meta-reasoning in humanly comprehensible ways (e.g., as belief and desire reasons that were the grounds for a particular decision), regardless of the formalism in which it performs the reasoning (Lomas et al. 2012). This amounts to an additional 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 concern about the robot's planned action and can conjure up an acceptable explanation, then the action has passed a social criterion for moral behavior.

Note that meta-reasoning and the ability to communicate such reasoning also allow the robot to inform the human user of the robot's general limitations and of specific problems it faces, with which the human user may be able to help. This kind of transparency may be able to forestall a great deal of misperception that a promising but imperfect artificial agent may cause in hopeful human users.

Who is responsible?

A question many people worry about is this: "When a robot makes mistakes, who is responsible?" It would seem that, as long as a machine doesn't cross the boundary of autonomous decision making, we have versions of product liability (not that those cases would be trivial, but at least they present no philosophical puzzles). As long as robots act on designs and programs devised by some person or company, that person or company is liable for failure (barring misuse). Once robots become end-user programmable, liability will be more distributed. And for semi-autonomous robots such as self-driving cars, things will undoubtedly get complicated if an accident is caused by the joint (or, worse yet, competing) operation of the robotic and human driver.

Responsibility might shift fully to the robot when that robot is autonomous ("self-governing")—that is, independent of the direct causal impact of programs, programmers, or operators. Such self-governance or independence requires the capacity for choice, which we have already highlighted as fundamental to moral action. It entails that the robot makes decisions that go beyond preprogrammed responses; it arrives at these decisions on the basis of both long-standing goals and occurrent perceptions, beliefs, and inductive inferences; and its decisions become increasingly independent the longer it is allowed to learn and grow. An agent with such capacities is almost certainly going to be a target of people's moral expectations and evaluations.

Here, then, we have one of the most striking points of contact between robot ethics and machine morality: What are the ethical grounds on which to confer (or deny) a robot decision-making autonomy and, as a likely result, moral obligations? One argument for autonomy is this: Because only robots with autonomy (and its constituent abilities of choice, reasoning, learning, etc.) could possibly master the demanding task of interacting with humans, collaborating with them, taking care of them, or teaching them. We wouldn't and shouldn't feel comfortable deploying preprogrammed robots in any of the important domains of societal need for which we lack financial and person resources—such as education, safety, medicine, elderly and disabled care—and for which we therefore need the help of technology. Successful performance in such tasks cannot simply rely on prior programs, because human behavior is too complex and variable. Humans evolved the capacity to be creative and adaptable (Mithen 1998), and this makes their behavior far more difficult to predict than other animals' behavior. But if robots face uncertainty in predicting their interaction partners' behavior, their own behavior in social interactions cannot be pre-programmed. A robot will need to monitor other people's responses to small changes in the situation and in turn flexibly respond to them—which requires autonomous decision making. As mentioned earlier, people will hold such autonomous robots responsible for their behavior.

But that would be an advantage. We need to remember that "holding people responsible" is a tool of social regulation. People morally criticize a norm violator because they assume or hope that they can change the person's

[3] Strictly speaking, this requirement holds only for truthful explanations, which I hope will be the default for social robots.

behavior. Most nonpsychopathic humans are sensitive to criticism and to the threat of social rejection (Baumeister and Leary 1995; Williams 2009) and are therefore inclined to change their future behavior when morally criticized. Blaming robots could have the same effect (and a much more immediate effect than drawn-out liability suits) if the process of behavior change is built into the robots' architecture—if autonomous robots are sensitive to human criticism. These robots don't have to fear criticism, they don't have to feel hurt by blame; but they have to be responsive to reasons that a moral critic provides. If they are not responsive they quickly lose status as social partners; and, just like nonresponsive humans are excluded from their community, such robots, too, would be excluded. Rather than building fear or insecurity into robots to motivate them to become accepted members of a community, it may suffice to equip them with a desire to be the best they can be: to be—through constant learning and improvement—the most morally competent robot. Humans might learn a thing or two from that.

Conclusion

"Machine morality" and "robot ethics" must be closely connected in the design and deployment of social robots. Rather than discuss in general ethical terms whether robots should become an integrated part of human society, we need to pose ethical questions relative to a set of possible moral competences a robot could have. *Should robots have autonomy?* Yes, if we want robots to be genuine interaction partners, but then they also must have moral competences such as an ability to follow complex norm systems and responsiveness to moral criticism. *Should robots always obey human commands?* Perhaps yes if their moral reasoning capacities are very limited, but no if they have the capacity to recognize human error (e.g., an immoral command) and to resolve norm conflicts. *Should we allow robots to kill humans?* We may have a strict value that machines must not kill. Alternatively, if part of the justification for such a strict stance is robots' currently limited moral competence, then the situation may change when, or if, morally competent robots demonstrably uphold laws as reliably as humans—or perhaps even more reliably because they would never commit emotionally motivated atrocities. *Should robots have rights and protection?* As robots acquire increasing moral competence, especially moral judgment and decision making, it may well be unethical to deny them all moral standing. Their rights may be limited, however, and vary as a function of their value and specific role in society (e.g., caretaker, teacher, repair assistant).

All these, and more, questions are up for debate and broad societal discussion. But the discussions must consider both ethical questions about how *humans* should design, deploy, and treat robots and questions about what moral capacities a *robot* could and should have. And if robotic design commits to building morally competent robots, then those robots could be trustworthy and productive partners, caretakers, educators, and members of the human community. Moral competence does not resolve all ethical concerns over robots in society, but it may be a prerequisite to resolve at least some of them.

Acknowledgments This project was partially supported by a grant from the Office of Naval Research (ONR), No. N00014-13-1-0269. The opinions expressed here are my own and do not necessarily reflect the views of ONR. The ideas on moral competence featured in this article have been developed jointly with Matthias Scheutz, Tufts University.

References

Alicke, M. D. (2000). Culpable control and the psychology of blame. *Psychological Bulletin, 126*, 556–574. doi:10.1037//0033-2909.126.4.556.

Allen, C. (2011). The future of moral machines. *The New York Times: Opinionator*. Retrieved December 29, 2014, from http://opinionator.blogs.nytimes.com/2011/12/25/the-future-of-moral-machines/.

Anderson, M., & Anderson, S. (2011). *Machine ethics*. Cambridge: Cambridge University Press.

Antaki, C. (1994). *Explaining and arguing: The social organization of accounts*. London: Sage.

Arkin, R. C. (2009). *Governing lethal behavior in autonomous robots*. Boca Raton, FL: CRC Press.

Asaro, P. M. (2006). What should we want from a robot ethic? *International Review of Information Ethics, 6*, 9–16.

Avramova, Y. R., & Inbar, Y. (2013). Emotion and moral judgment. *Wiley Interdisciplinary Reviews Cognitive Science, 4*, 169–178. doi:10.1002/wcs.1216.

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117*, 497–529. doi:10.1037/0033-2909.117.3.497.

Bello, P. (2012). Cognitive foundations for a computational theory of mindreading. *Advances in Cognitive Systems, 1*, 59–72.

Bicchieri, C. (2006). *The grammar of society: The nature and dynamics of social norms*. New York, NY: Cambridge University Press.

Blomkamp, N., Kinberg, S. (Producers), & Blomkamp, N. (Director). (2015). *Chappie* [Motion picture]. USA: Sony Pictures Home Entertainment.

Brachman, R. J. (2002). Systems that know what they're doing. *IEEE Intelligent Systems, 17*, 67–71. doi:10.1109/MIS.2002.1134363.

Breazeal, C. L. (2002). *Designing sociable robots*. Cambridge, MA: MIT Press.

Bringsjord, S. (2009). But perhaps robots are essentially non-persons. *Erwägen Wissen Ethik, 20*, 193–195.

Bringsjord, S., Arkoudas, K., & Bello, P. (2006). Toward a general logicist methodology for engineering ethically correct robots. *Intelligent Systems, IEEE, 21*, 38–44.

Calverley, D. J. (2006). Android science and animal rights, does an analogy exist? *Connection Science, 18*, 403–417. doi:10.1080/09540090600879711.

Coates, D. J., & Tognazzini, N. A. (2012). The contours of blame. In D. J. Coates & N. A. Tognazzini (Eds.), *Blame: Its nature*

and norms (pp. 3–26). New York, NY: Oxford University Press.

Coeckelbergh, M. (2010). Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology, 12*, 209–221. doi:10.1007/s10676-010-9235-5.

Cox, M. T. (2011). Metareasoning, monitoring, and self-explanation. In M. T. Cox & A. Raja (Eds.), *Metareasoning* (pp. 131–149). Cambridge, MA: The MIT Press.

Csibra, G., & Gergely, G. (2009). Natural pedagogy. *Trends in Cognitive Sciences, 13*, 148–153. doi:10.1016/j.tics.2009.01.005.

Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*, 353–380. doi:10.1016/j.cognition.2008.03.006.

Cushman, F., & Young, L. (2011). Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science, 35*, 1052–1075. doi:10.1111/j.1551-6709.2010.01167.x.

DeBaets, A. M. (2014). Can a robot pursue the good? Exploring artificial moral agency. *Journal of Evolution and Technology, 24*, 76–86.

Dersley, I., & Wootton, A. (2000). Complaint sequences within antagonistic argument. *Research on Language and Social Interaction, 33*, 375–406. doi:10.1207/S15327973RLSI3304_02.

Eisenberg, N. (2000). Emotion, regulation, and moral development. *Annual Review of Psychology, 51*, 665–697.

Emde, R. N. (1992). Social referencing research: Uncertainty, self, and the search for meaning. In S. Feinman (Ed.), *Social referencing and the social construction of reality in infancy* (pp. 79–94). New York, NY: Plenum Press.

Fehr, E., & Fischbacher, U. (2004). Third-party punishment and social norms. *Evolution and Human Behavior, 25*, 63–87. doi:10.1016/S1090-5138(04)00005-4.

Fisher, M., Spielberg, S., & Weaver, B. (2014). *Extant [Television series]*. Los Angeles: CBS.

Fiske, S. T., & Taylor, S. E. (2008). *Social cognition: From brains to culture* (1st ed.). Boston, MA: McGraw-Hill.

Flack, J. C., & de Waal, F. B. M. (2000). "Any animal whatever". Darwinian building blocks of morality in monkeys and apes. *Journal of Consciousness Studies, 7*, 1–29.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*, 349–379. doi:10.1023/B:MIND.0000035461.63578.9d.

Ford, K. M., & Hayes, P. J. (1991). *Reasoning agents in a dynamic world: The frame problem*. Greenwich, CT: JAI Press.

Fridin, M. (2014). Kindergarten social assistive robot: First meeting and ethical issues. *Computers in Human Behavior, 30*, 262–272. doi:10.1016/j.chb.2013.09.005.

Garcia, E., Jimenez, M. A., De Santos, P. G., & Armada, M. (2007). The evolution of robotics research. *IEEE Robotics Automation Magazine, 14*, 90–103. doi:10.1109/MRA.2007.339608.

Gilovich, T., Keltner, D., & Nisbett, R. E. (2013). *Social psychology* (3rd ed.). New, NY: W.W. Norton & Co.

Grau, C. (2011). There is no "I" in "Robot": Robots and utilitarianism. In M. Anderson & S. L. Anderson (Eds.), *Machine ethics* (pp. 451–463). Cambridge: Cambridge University Press.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*, 101–124. doi:10.1080/1047840X.2012.651387.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment. *Neuron, 44*, 389–400. doi:10.1016/j.neuron.2004.09.027.

Guglielmo, S., Monroe, A. E., & Malle, B. F. (2009). At the heart of morality lies folk psychology. *Inquiry: An Interdisciplinary Journal of Philosophy, 52*, 449–466. doi:10.1080/00201740903302600.

Gunkel, D. J. (2014). A vindication of the rights of machines. *Philosophy & Technology, 27*, 113–132. doi:10.1007/s13347-013-0121-z.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*, 814–834. doi:10.1037/0033-295X.108.4.814.

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. doi:10.1177/0963721412470687.

Harenski, C. L., Harenski, K. A., Shane, M. S., & Kiehl, K. A. (2010). Aberrant neural processing of moral violations in criminal psychopaths. *Journal of Abnormal Psychology, 119*, 863–874.

Heath, J. (2001). *Communicative action and rational choice. Studies in contemporary German social thought*. Cambridge, MA: MIT Press.

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

Hoffman, M. L. (2008). Empathy and prosocial behavior. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), *Handbook of emotions* (3rd ed., pp. 440–455). New York, NY: Guilford Press.

Hofmann, B. (2013). Ethical challenges with welfare technology: A review of the literature. *Science and Engineering Ethics, 19*, 389–406.

Huebner, B., Dwyer, S., & Hauser, M. (2009). The role of emotion in moral psychology. *Trends in Cognitive Sciences, 13*, 1–6. doi:10.1016/j.tics.2008.09.006.

Hutcherson, C. A., & Gross, J. J. (2011). The moral emotions: A social—functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology, 100*, 719–737. doi:10.1037/a0022408.

Johnson, A. M., & Axinn, S. (2013). The morality of autonomous robots. *Journal of Military Ethics, 12*, 129–141. doi:10.1080/15027570.2013.818399.

Kahn, Jr., P. H., Kanda, T., Ishiguro, H., Gill, B. T., Ruckert, J. H., Shen, S., Gary, H. E., et al. (2012). *Do people hold a humanoid robot morally accountable for the harm it causes?. Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 33–40). New York, NY: ACM. doi:10.1145/2157689.2157696.

Kibble, R. (2012). Can an unmanned drone be a moral agent? Ethics and accountability in military robotics. In D. J. Gunkel, J. J. Bryson, & S. Torrance (Eds.), *The machine question: AI, ethics and moral responsibility (Proceedings of symposium "Machine Question: AI, Ethics, and Moral Responsibility" AISB/IACAP 2012)* (pp. 62–67). The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Knobe, J. (2010). Person as scientist, person as moralist. *Behavioral and Brain Sciences, 33*, 315–329. doi:10.1017/S0140525X10000907.

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

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

Lin, P. (2013). The ethics of autonomous cars. *The Atlantic*. Retrieved September 30, 2014, from http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/.

Lin, P., Abney, K., & Bekey, G. A. (Eds.). (2012). *Robot ethics: The ethical and social implications of robotics*. Cambridge, MA: MIT Press.

Littman, M. L. (2001). Value-function reinforcement learning in Markov games. *Cognitive Systems Research, 2*, 55–66. doi:10.1016/S1389-0417(01)00015-8.

Lomas, M., Chevalier, R., Cross, E. V., Garrett, R. C., Hoare, J., & Kopack, M. (2012). Explaining robot actions. *Proceedings of the 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 187–188). Boston, MA.

Luo, Q., Nakic, M., Wheatley, T., Richell, R., Martin, A., & Blair, R. J. R. (2006). The neural basis of implicit moral attitude—An IAT study using event-related fMRI. *NeuroImage, 30*, 1449–1457. doi:10.1016/j.neuroimage.2005.11.005.

Malle, B. F. (1999). How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review, 3*, 23–48. doi:10.1207/s15327957pspr0301_2.

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

Malle, B. F. (2011). Time to give up the dogmas of attribution: A new theory of behavior explanation. In M. P. Zanna & J. M. Olson (Eds.), *Advances of experimental social psychology* (Vol. 44, pp. 297–352). San Diego, CA: Academic Press.

Malle, B. F., & Dickert, S. (2007). Values. In R. F. Baumeister & K. D. Vohs (Eds.), *The encyclopedia of social psychology*. Thousand Oaks, CA: Sage.

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*, 147–186. doi:10.1080/1047840X.2014.877340.

Malle, B. F., & Scheutz, M. (2014). *Moral competence in social robots*. *IEEE International Symposium on Ethics in Engineering, Science, and Technology* (pp. 30–35). Presented at the IEEE International Symposium on Ethics in Engineering, Science, and Technology, June, Chicago, IL: IEEE.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. *HRI'15: Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 117–124). New York, NY: ACM.

McCullough, M. E., Kurzban, R., & Tabak, B. A. (2013). Putting revenge and forgiveness in an evolutionary context. *Behavioral and Brain Sciences, 36*, 41–58. doi:10.1017/S0140525X12001513.

McKenna, M. (2012). Directed blame and conversation. In D. J. Coates & N. A. Tognazzini (Eds.), *Blame: Its nature and norms* (pp. 119–140). New York, NY: Oxford University Press.

MHAT-IV. (2006). *Mental Health Advisory Team (MHAT) IV: Operation Iraqi Freedom 05-07 Final report*. Washington, DC: Office of the Surgeon, Multinational Force-Iraq; Office of the Surgeon General, United States Army Medical Command.

Mikhail, J. (2007). Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences, 11*, 143–152. doi:10.1016/j.tics.2006.12.007.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67*, 371–378.

Millar, J. (2014). An ethical dilemma: When robot cars must kill, who should pick the victim? Robohub. *Robohub.org*. Retrieved September 28, 2014, from http://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim/.

Mithen, S. (Ed.). (1998). *Creativity in human evolution and prehistory*. New York, NY: Taylor & Francis.

Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing free will down to earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and Cognition, 27*, 100–108. doi:10.1016/j.concog.2014.04.011.

Monroe, A. E., & Malle, B. F. (2010). From uncaused will to conscious choice: The need to study, not speculate about people's folk concept of free will. *Review of Philosophy and Psychology, 1*, 211–224.

Monroe, A. E., & Malle, B. F. (2014). Free will without metaphysics. In A. R. Mele (Ed.), *Surrounding free will* (pp. 25–48). New York, NY: Oxford University Press.

Moor, J. H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems, 21*, 18–21. doi:10.1109/MIS.2006.80.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.

Nourbakhsh, I. R. (2013). *Robot futures*. Cambridge, MA: MIT Press.

Open Roboethics Initiative. (2014a). If death by autonomous car is unavoidable, who should die? Reader poll results.

Open Roboethics Initiative. (2014b). My (autonomous) car, my safety: Results from our reader poll.

Parthemore, J., & Whitby, B. (2013). What makes any agent a moral agent? Reflections on machine consciousness and moral agency. *International Journal of Machine Consciousness, 4*, 105–129.

Paxton, J. M., Ungar, L., & Greene, J. D. (2012). Reflection and reasoning in moral judgment. *Cognitive Science, 36*, 163–177. doi:10.1111/j.1551-6709.2011.01210.x.

Petersen, S. (2007). The ethics of robot servitude. *Journal of Experimental & Theoretical Artificial Intelligence, 19*, 43–54. doi:10.1080/09528130601116139.

Powell, N. L., Derbyshire, S. W. G., & Guttentag, R. E. (2012). Biases in children's and adults' moral judgments. *Journal of Experimental Child Psychology, 113*, 186–193. doi:10.1016/j.jecp.2012.03.006.

Powers, T. M. (2006). Prospects for a Kantian machine. *IEEE Intelligent Systems, 21*, 46–51. doi:10.1109/MIS.2006.77.

Powers, T. M. (2011). Incremental machine ethics. *Robotics & Automation Magazine, IEEE, 18*, 51–58. doi:10.1109/MRA.2010.940152.

Pylyshyn, Z. W. (Ed.). (1987). *The Robot's dilemma: The frame problem in artificial intelligence*. Norwood, NJ: Ablex.

Ryle, G. (1949). *The concept of mind*. London: Penguin Press.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Hillsdale, NJ: Erlbaum.

Scheutz, M. (2012). The affect dilemma for artificial agents: Should we develop affective artificial agents? *IEEE Transactions on Affective Computing, 3*, 424–433.

Scheutz, M., & Crowell, C. R. (2007). *The burden of embodied autonomy: Some reflections on the social and ethical implications of autonomous robots*. *Proceedings of Workshop on Roboethics at ICRA 2007*. Rome, Italy.

Scheutz, M., Malle, B. F., & Briggs, G. (2015). Towards morally sensitive action selection for autonomous social robots. The 24th IEEE international symposium on robot and human interactive communication, 2015 RO-MAN. Presented at the 24th IEEE International Symposium on Robot and Human Interactive Communication. (2015). *RO-MAN*. Japan: Kobe.

Scheutz, M., & Malle, B. F. (2014). *"Think and do the right thing": A plea for morally competent autonomous robots.* Presented at the 2014 IEEE Ethics conference, Chicago, IL.

Semin, G. R., & Manstead, A. S. R. (1983). *The accountability of conduct: A social psychological analysis*. London: Academic Press.

Shaver, K. G. (1985). *The attribution of blame: Causality, responsibility, and blameworthiness*. New York: Springer.

Sullins, J. P. (2011). Introduction: Open questions in roboethics. *Philosophy & Technology, 24*, 233. doi:10.1007/s13347-011-0043-6.

Talamadupula, K., Schermerhorn, P., Benton, J., Kambhampati, S., & Scheutz, M. (2011). *Planning for agents with changing goals. ICAPS 2011 System Demonstration*. Germany: Freiburg.

Tanaka, F., Cicourel, A., & Movellan, J. R. (2007). Socialization between toddlers and robots at an early childhood education center. *Proceedings of the National Academy of Sciences, 104*, 17954–17958. doi:10.1073/pnas.0707769104.

Tedeschi, J. T., & Reiss, M. (1981). Verbal strategies as impression management. In C. Antaki (Ed.), *The psychology of ordinary social behaviour* (pp. 271–309). London: Academic Press.

Thiessen, E. D., Kronstein, A. T., & Hufnagle, D. G. (2013). The extraction and integration framework: A two-process account of statistical learning. *Psychological Bulletin, 139*, 792–814. doi:10.1037/a0030801.

Tomasello, M., & Vaish, A. (2013). Origins of human cooperation and morality. *Annual Review of Psychology, 64*, 231–255. doi:10.1146/annurev-psych-113011-143812.

Traverso, V. (2009). The dilemmas of third-party complaints in conversation between friends. *Journal of Pragmatics, 41*, 2385–2399. doi:10.1016/j.pragma.2008.09.047.

Turiel, E. (1983). *The development of social knowledge: Morality and convention*. Cambridge, England: Cambridge University Press.

Van Berkum, J. J. A., Holleman, B., Nieuwland, M., Otten, M., & Murre, J. (2009). Right or wrong? The brain's fast response to morally objectionable statements. *Psychological Science, 20*, 1092–1099. doi:10.1111/j.1467-9280.2009.02411.x.

van Wynsberghe, A. (2013). Designing robots for care: Care centered value-sensitive design. *Science and Engineering Ethics, 19*, 407–433. doi:10.1007/s11948-011-9343-6.

Veloso, M., Aisen, M., Howard, A., Jenkins, O. C., Mutlu, B., & Scassellati, B. (2012). *Human-robot interaction: Japan, South Korea, and China*. WTEC Panel Report. Arlington, VA: World Technology Evaluation Center, Inc.

Veruggio, G., Solis, J., & Van der Loos, M. (2011). Roboethics: Ethics applied to robotics. *IEEE Robotics Automation Magazine, 18*, 21–22. doi:10.1109/MRA.2010.940149.

Voiklis, J., Cusimano, C., & Malle, B. F. (2014). A social-conceptual map of moral criticism. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 36th annual conference of the cognitive science society* (pp. 1700–1705). Austin, TX: Cognitive Science Society.

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

Wallach, W. (2010). Robot minds and human ethics: The need for a comprehensive model of moral decision making. *Ethics and Information Technology, 12*, 243–250. doi:10.1007/s10676-010-9232-8.

Wallach, W., & Allen, C. (2008). *Moral machines: Teaching robots right from wrong*. New York, NY: Oxford University Press.

Warneken, F., Lohse, K., Melis, A. P., & Tomasello, M. (2011). Young children share the spoils after collaboration. *Psychological Science, 22*, 267–273. doi:10.1177/0956797610395392.

Weiner, B. (1995). *Judgments of responsibility: A foundation for a theory of social conduct*. New York, NY: Guilford Press.

Williams, K. D. (2009). Ostracism: A temporal need-threat model. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 41, pp. 275–314). San Diego, CA: Elsevier Academic Press.

Wolpert, D. M., & Flanagan, J. R. (2001). Motor prediction. *Current Biology, 11*, R729–R732. doi:10.1016/S0960-9822(01)00432-8.

Wright, J. C., & Bartsch, K. (2008). Portraits of early moral sensibility in two children's everyday conversations. *Merrill-Palmer Quarterly, 54*, 56–85. doi:10.2307/23096079.

Wyman, E., Rakoczy, H., & Tomasello, M. (2009). Normativity and context in young children's pretend play. *Cognitive Development, 24*, 146–155. doi:10.1016/j.cogdev.2009.01.003.