Scheutz, M., & Malle, B. F. (forthcoming). Moral robots. In K. Rommelfanger and S. Johnson (eds.), *Routledge Handbook of Neuroethics*. New York, NY: Routledge/Taylor & Francis.

Moral Robots

Matthias Scheutz and Bertram F. Malle

5-10 key words

ethics, roboethics, machine morality, moral psychology, artificial intelligence, robotics, human-robot interaction

5 recommended readings with brief descriptive sentence

Arkin, R.C., 2009. *Governing lethal behavior in autonomous robots*. Chapman & Hall/CRC.
One of the most detailed discussions and implementations of moral decision making for a robotic system.

Asimov, I., 1942. *Runaround*. Astounding Science Fiction.
The renowned science fiction writer's classic story introducing the three laws of robotics.

Lin, P., Abney, K., & Bekey, G. A. (Eds.). (2012). *Robot ethics: The ethical and social implications of robotics*. Cambridge, MA: MIT Press.
The most comprehensive collection to date of influential scholars working on questions of robot ethics.

Malle, B.F. and Scheutz, M. (2014), "Moral Competence in Social Robots", *Proceedings of IEEE International Symposium on Ethics in Engineering, Science, and Technology, Ethics'2014*, IEEE, Chicago, IL, pp. 30–35
Introduces a model of moral competence.

Wallach, W., & Allen, C., 2009. *Moral machines: Teaching robots right from wrong*. Oxford University Press, Oxford.
Extensive and influential analysis of potential moral capacities in robots.

Notes on Contributors

Matthias Scheutz, Tufts University, USA. Trained in philosophy, formal logic, and computer science in Vienna, Austria, and computer science and cognitive science at Indiana University Bloomington. Past co-chair of the IEEE Technical Committee on Robot Ethics and of the 2009 ACM/IEEE International Conference on Human-Robot Interaction.

Bertram F. Malle, Brown University, USA. Trained in psychology, philosophy, and linguistics in Austria and at Stanford University. Past president of the Society of Philosophy and Psychology, author of *How the Mind Explains Behavior* (MIT Press, 2004) and *Other minds* (with S. D. Hodges, eds., Guilford, 2005).

Introduction

Humans differ from other animals in many respects. One of the most distinguishing human features is *morality*—that is, the capacity to perceive actions as *moral* or *immoral* and respond to them in very particular ways, such as by praising or blaming actors, demanding justifications, or accepting an apology. Members of a community are expected to abide by the moral norms and values of the community, and they pass on the knowledge of these norms through observed practices and explicit instruction. In addition, modern societies have made those norms explicit through philosophical reflection and formalized laws, thereby offering ethical foundations for their members to live by. We will thus use the term "moral" (e.g., in "moral processing" or "moral competence") to refer to those aspects of the (currently human) cognitive system that are involved in the representation and processing of norms, values, and virtues. Morality, in this sense, is a natural phenomenon of social groups in their daily lives, a phenomenon that can be studied with empirical scientific methods. The scientific study of these phenomena is now typically called *moral psychology*. We reserve the term "ethical" for normative discussions and debates about abstract principles (e.g., the doctrine of double effect), theological origins of values, or the difference between meta-ethical theories (e.g., Kantian, utilitarian). Ethics is then more of a philosophical, normative (not empirical) discipline. Consequently, it is possible for an agent to engage in moral decisions making (i.e., involving ordinary moral information processing) but to perform an act that is considered unethical within some normative (theological or philosophical) system; conversely, it is possible for an agent to act in conformity with ethical principles even if the decision to act was not guided by the person's moral processing of those abstract principles but, say, by imitation. This distinction helps distinguish two sets of questions that arise when considering the behavior of non-human agents, such as robots. One set of

scholarly questions concerns the robot's functional capacities and computational processes that mimic or adapt to human moral processing. This endeavor falls squarely within cognitive science, integrating, in particular, behavioral research and theorizing with computational modeling and engineering. Another set of scholarly questions concerns the ethical standards that robots should adhere to, the abstract principles (if any) the robots should implement, the ethical value of building morally competent robots in the first place, and so on. This endeavor falls into the domain of ethics as a normative discipline, typically conducted by philosophers or theologians.

The main reason for raising the question about the ethical behavior of robots is the rapid progress in the development of autonomous social robots that are specifically created to be deployed in sensitive human environments: from elder and health care settings to law enforcement and military contexts. Clearly, such tasks and environments are very different from the traditional factory settings (of welding robots, say). Hence, these new social robots will require higher degrees of autonomy and decision-making than any previously developed machine, given that they will face a much more complex, open world. They might be required to acquire new knowledge on the fly to accomplish a never before encountered task. Moreover, they will likely face humans who are not specifically trained to interact with them, and robots thus need to be responsive to instructions by novices and feel "natural" to humans even in unstructured interactions (Scheutz et al., 2006). At these levels of autonomy and flexibility in near-future robots, there will be countless ways in which robots might make mistakes, violate a user's expectations and moral norms, or threaten the user's physical or psychological safety. These social robots must therefore also be moral robots.

Thus, for autonomous social robots deployed in human societies, three key questions arise: (1) What moral expectations do humans have for social robots? (2) What moral competence can and should such robots realize? (3) What should be the moral standing of these machines (if any)?

The first question, about moral expectations, follows from the well established fact that autonomous social robots, especially those with natural language abilities, are treated in many ways like humans, regardless of whether such treatment was intended or anticipated by the robot designers. In fact, there is mounting evidence that humans have very clear expectations of robot capacities based on the robot's appearance and people's perceptions of the robot behaviors. We will review some of this work in Section 3.

The second question, about moral competence, arises from the need to endow robots with sufficient capacities to operate safely in human societies. Answers to this question would ideally build on answers to the first question and provide mechanisms for robots to process social and moral norms in ways that humans expect. In addition, the design of the robots' control systems should ensure that robots behave ethically according to the norms of the society in which they are deployed.

The third question, about moral standing, is a consequence of allowing robots to make decisions on their own without human supervision, as will become necessary in cases where human supervisors are not reachable or could not react quickly enough. In such instances of autonomous decision-making and action, we will have to decide whether the robot is "responsible" for its actions, especially when its behavior caused damage to property or harm to humans. Questions about fault, accountability, but also about the robot's rights for due process and protection against harm, will have to be answered.

In this chapter we will focus on the first two questions, discussing both human expectations as well as computational architectural mechanisms that will allow robots to live up to those expectations, while leaving a detailed discussion of the third question to legal experts and philosophers (Asaro, 2012; Coeckelbergh, 2010; Gunkel et al., 2012; Pagallo, 2011). However, the philosophical and legal discussions on moral standing of robots do raise the inevitable question of whether robots could ever be moral. For if robots are not the kind of thing to which morality applies, requiring any future robot to be moral is meaningless. One might argue that robots are not conscious autonomous agents with free will and therefore cannot make decisions act on those decisions. A long tradition in philosophy has tried to understand what it means to be an agent with free will and moral responsibility. But discussions of free will have led to little consensus and often raise more questions than they are intended to answer, especially when applied to artificial agents (Gunkel, 2014). In the end, what counts for real social robots — machines in our contemporary world — is whether people treat those robots as targets of their moral sentiments. And it turns out that ascriptions of free will have little bearing on people's inclination to treat any agent as a moral being (Monroe et al., 2014). We will therefore not focus on the philosophical debate about agency and personhood. Rather, we will assume an operational behavioral definition of a "moral robot":

> **Definition**. A robot is "moral" if it has one or more relevant competences that people consider important for living in a moral community, such as detecting behaviors that violate moral norms or deciding to act in light of moral norms (Malle, 2015; Malle and Scheutz, 2014; Scheutz, 2014; Scheutz and Malle, 2014).

This definition will permit us to talk about a robot's morality in terms of various functional capacities that collectively give rise to morally significant behaviors in the contemporary world.

Moral competence, by this definition, neither requires nor implies an objective "moral agency" (Malle, 2015). Instead, framing the problem in terms of competences allows us to design and study the mechanisms in the robot's control architecture that make its behavior consistent with community expectations about moral judgment and decision making, moral communication, and ethically acceptable behavior. After all, it is the community of ordinary people that will interact with social, moral robots, and it is their expectations and standards that will make robots acceptable community members or not.

Why Moral Robots?

The topics of intelligent autonomous machines and human morality have frequently been paired in the science fiction literature, most notably in the opus of Isaac Asimov, who early on addressed the tensions and challenges resulting from machines operating in human societies. In his stories, he specifically conceived of "Three Laws of Robotics" (Asimov, 1942), which he envisioned to be ingrained in the robots' "positronic brains," allowing them to make ethically sound decisions and thus exhibit ethical behavior (usually, but not always, living up to human moral expectations). Specifically, the three laws set up a system of strictly prioritized principles that robots had to obey:

L1: A robot may not injure a human being or, through inaction, allow a human being to come to harm

L2: A robot must obey orders given it by human beings except where such orders would conflict with L1

L3: A robot must protect its own existence as long as such protection does not conflict with L2

While these three laws provided fertile ground for stories built on the implicit tensions among the laws—whom to save in a crowd when not all can be saved, how to evaluate the consequences of commands that are not clear, or how to determine to whom the predicate "human" should even apply—they are neither theoretically nor practically adequate for providing the foundation of moral competence in robots (Murphy and Woods, 2009). However, they do serve the purpose of pointing out the need to develop *some provisions* for ethical behavior in autonomous robots.

Of course, robot designers have been aware of the need to ensure the safe operation of robots all along, without the need to look to the science fiction literature for suggestions. Autonomous robotic systems that could have an impact on humans or property have precautionary built-in safety mechanisms that allow them to either completely avoid or to massively reduce the likelihood of any sort of harm. For example, self-parking cars will automatically stop if they sense an obstacle in their way, without the need for an explicitly represented ethical principle "do not drive into obstacles." Similarly, compliant robot arms intended to operate in human workspaces will yield when coming in contact with another object such as a human body part. Moreover, instructions to the car or the robot arm to continue a colliding trajectory will be automatically rejected, again under the same constraints. In all of these cases, the robots' actions or rejections of actions are not *explicitly* defined in terms of (moral) rules or principles, but rather *implicitly* in the algorithms that control the robot's behaviors.

The "implicit-explicit" distinction of safety principles can be generalized to a distinction between implicit and explicit ethical agents based on a taxonomy introduced by (Moor, 2006). Implicit ethical agents are agents that "have ethical considerations built into (i.e., implicit in) their design. Typically, these are safety or security considerations" (Moor, 2013). By contrast,

explicit ethical agents are agents that "can identify and process ethical information about a variety of situations and make sensitive determinations about what should be done. When ethical principles are in conflict, these robots can work out reasonable resolutions." (ibid.) And Moor continues: "Explicit ethical agents are the kind of agents that can be thought of as acting from ethics, not merely *according* to ethics"—as is the case with implicit ethical agents.

Moor also introduced two additional categories: *ethical impact agents*—that is, agents whose behavior can have ethical consequences—and *full ethical agents* (referring to normal adult humans with features such as consciousness, intentionality, and free will). Assuming that autonomous social robots will at least be ethical impact agents, the question then is how one would go about developing algorithms that will turn them into either implicit or explicit ethical agents. Mechanisms producing implicit ethical agents might be sufficient for a variety of tasks and domains (e.g., where most or all demands and risks are known before task execution, such as in the case of a robotic vacuum cleaner). However, mechanisms producing explicit ethical agents will be required for robots deployed in more open-ended tasks and environments, such as for household robots that have to learn the customs of a particular home and adapt to its changing situations.

We will start by looking at the empirical evidence for human expectations about moral robots and then consider ways to implement them in robotic control systems.

Human Moral Expectations About Autonomous Social Robots

It has long been known that humans have a natural propensity to view moving objects as "agents with intentions", even if those objects do not resemble any known life-form at all. Early studies by (Heider and Simmel, 1944) showed that human observers "see" mental states such as emotions and intentions even in circles and triangles moving around in a cartoon-like scene.

Moreover, humans from infancy on can easily be brought to make judgments about whether the intentions of those agents are benevolent or malicious, and thus exhibit basic moral evaluations based on their perception of the interacting shapes (Hamlin, 2013).

This human propensity to project agency onto self-propelled objects (Premack, 1990), presumed to be an evolutionary adaptation that allowed humans to anticipate dangers from other agents, is particularly consequential for the development of robots. For robots are self-propelled objects that typically move about the environment in ways that suggest goal-driven behavior to human observers (Tremoulet and Feldman, 2000). There is evidence that even simple robots like the Roomba vacuum cleaner in the form of a disk with no life-like features (such as eyes or arms) can trigger the "human agency detector" (Scheutz, 2012). Hence, it stands to reason that humans not only project agency but may, under some circumstances, project *moral* characteristics onto such machines (Malle and Scheutz, 2016).

Researchers working in the field of human-robot interaction have investigated human reactions to robots violating norms of varying severity. (Strait et al., 2014), for example, investigated a violation of the social norm to "be polite." They examined whether people preferred robot tutors that were polite by giving hedged instructions as opposed to robots that used imperatives in their instructions. People had no such preference in their own interactions with robots, but when they observed other people interact with the robot they preferred the polite one.

(Short et al., 2010) examined a violation of a more serious norm—"tell the truth." A robot and human repeatedly played the game "rock-paper-scissors," and the robot had to announce the winner of each round. In one of the conditions, the robot announced that it won the

round even though it had lost. Humans found the "cheating" robot to be more engaging and made greater attributions of mental states to that robot in the conditions in which it cheats.

To the extent that people ascribe mental states to a robot they may also grant the robot certain rights and protect it from unfair treatment. The evidence shows that both children and adults do so. In (Kahn, Jr. et al., 2012), children interacted with a robot that was suddenly locked away in a closet by the experimenter because it "wasn't needed anymore." The robot protested, but to no avail. The researchers documented that children viewed the robot as having mental states and believed that the robot deserved to be treated fairly. (Briggs and Scheutz, 2014) investigated in a series of studies to what extent people themselves would be responsive to a robot's moral appeals—protesting an instruction the participants gave the robot but that it deemed unfair. People were significantly less likely to insist on the robot following that instruction when the robot protested, regardless of whether the protesting robot was a victim or witness of the unfairness, and independent of the robot's appearance. However, some features of the robot seem to matter, as people are more reluctant to physically hit robots that are more intelligent (Bartneck et al., 2007) and robots that are described in a personalized manner (Darling et al., 2015).

In addition to live human-robot interaction experiments, recent studies have started to look at situations that cannot be examined in a laboratory context—either because they go beyond what is ethically acceptable in a research study or because they depict robots that do not yet exist.

For example, (Scheutz and Arnold, 2016a) surveyed participants about their attitudes toward sex robots. They found a consistent gender difference in what people considered appropriate uses for sex robots, with women less inclined than men to consider them socially

useful. However, there were also convergences between men and women on what sex robots are like and how sex with them is to be classified.

(Malle et al., 2015) examined people's responses to a situation that cannot be studied in the lab and is also not yet part of our reality: a robot itself making a moral decision about life and death. Participants read a narrative describing an agent (human or robot) caught in a moral dilemma: either (a) to intervene in a dangerous situation and save four persons while sacrificing the life of one, or (b) to stand back and let four persons die. People considered it more permissible if the robot sacrificed one person for the good of many than if the human did it. Moreover, when people were confronted with the agent's actual choice, they blamed a human who intervened more than a human who stood back, but they blamed a robot that intervened no more or even less than a robot that stood back. Recently, researchers have begun to probe people's responses to self-driving cars that might, in the near future, face similar moral dilemmas. (Bonnefon et al., 2016) found a contrast between people's judgments of what would be the morally right action for a self-driving car (namely, to sacrifice one pedestrian to save many) and what kind of car people would buy or like to see around the neighborhood (namely, one that doesn't intervene).

This research is in its infancy, and subtle variations may shift people expectations and preferences (Malle et al., 2016). Nonetheless, the results so far suggest two conclusions, one firm, the second one more tentative. First, people readily direct moral expectations and moral judgments to robots, at least robots of sufficient cognitive complexity and behavioral abilities. Second, people may be more accepting of robots than of humans to make conflictual decisions (e.g., endangering one individual while trying to save multiple individuals). The exact reasons for such a potential transfer or responsibility are currently unclear. One hypothesis the deserves

consideration is that making such decisions normally carries significant emotional costs (such as guilt and trauma) and social costs (affecting one's relationships with others), as is suspected, for example, in drone pilots (Chatterjee, 2015). Having robots make such decisions would reduce those human costs. However, there is currently a significant debate over using autonomous machines as lethal weapons (Arkin, 2009; Asaro, 2011; Sparrow, 2007), and reducing current human costs is only one of many factors to consider. Without staking a position in this debate, we would like to emphasize the importance of investigating ordinary people's psychological responses to near-future robots that might make morally significant decisions. Some of people's responses may be inevitable (given the psychological mechanisms humans are equipped with; (Malle and Scheutz, 2016); other responses may change with instruction and experience. Either way, designers, engineers, and policy makers need to take those responses under advisement to guide the robots' proper development, deployment, and possible legal regulations for their behavior.

Options For Developing Moral Or Ethical Robots

Positing now that people expect robots to have at least some moral competences, the key question becomes what it would take to actually endow robots with moral competence. We consider three main options, all with their own advantages and disadvantages.

1. Implement *ethical theories* as proposed by philosophers
2. Implement *legal principles* as proposed by legal scholars
3. Implement *human-like moral competence* as proposed by psychologists

Implementing Ethical Theories

(Gips, 1995), among others, suggested we could equip a robot with one of the three major philosophical ethical theories. The first main theory is *virtue ethics*, which posits that ethical

thought and action is guided by a person's character, constituted by "virtues" such as wisdom, courage, temperance, and justice. Moor specifically links implicit ethical agents to virtue ethics when he says that "implicit ethical agents have a kind of built-in virtue – not built-in by habit but by specific hardware or programming" (Moor, 2013). In some cases, virtues can be directly implemented in robot behavior. "Courage," for instance, might be realized by the robot's willingness to engage in a risky action (possibly endangering its own existence) when that action might avert harm to a human. For example, an autonomous vehicle might initiate an evasive maneuver that would prevent colliding with a pedestrian but risk crashing into parked cars, thus likely damaging the robotic car. The implementation of other virtues is less obvious. For example, it is unclear how "wisdom" could be realized in a robotic system over and above demands of rational behavior, such as when a game-playing computer always picks the best move from its perspective.

The second main ethical theory, *deontology*, posits that ethical action is not based on a virtuous character but on explicit rules, which can sensibly be applied to machines. (Gert, 2005) proposed that one could characterize ethical behavior in terms of a set of basic rules, each with the following structure: "everyone is always to obey the rule except when a fully informed rational person can publicly allow violating it" (p. 203). Such rules might include "don't kill," "don't cause pain," "don't deceive," "obey the law," etc., which apply quite generally. Anyone who violates such a rule "when no rational person can publicly allow such a violation may be punished" (p. 203).

Setting aside important questions about what rules to select for a robot and what it would mean to punish or hold a robot responsible for a rule violation, robots that abide by a given set of ethical rules arguably behave ethically. To implement such a system, robot designers could

employ "deontic logics," which have specifically been developed to allow for reasoning with the core concepts of *obligation*, *permission*, *prohibition*, and *option*, all of which can be defined in terms of permissions. That is, an action α is *obligatory* if not doing it is not permitted, α is *prohibited* if doing it is not permitted, and α is *optional* if doing it or not doing it is permitted. Basic axioms and rules of inference can then enable logical derivations in a given context to determine what the robot ought to do. This works well as long as there are no conflicting obligations, such as when the robot is obligated to do α, obligated to do β, but cannot physically (or practically) do both α and β together. Not only does the logical approach not give any advice on what to do in such cases, but standard deontic logics will, more generally, allow the robot to infer that every action is obligated (e.g., (Goble, 2005), which is clearly not intended. Hence, aside from other questions about computational feasibility and scalability, a challenge with the formal deontic approach is to curb the impact of deontic conflicts to not render all inferences useless.

The third ethical theory, *consequentialism*, is historically the newest and also the one that meshes best with computational mechanisms already implemented in robotic control systems: expected utility theory. The basic idea is to always choose an action that *maximizes the good for everybody involved*. Formally, this means that the robot would consider all available actions α together with their probability of success $p(\alpha)$ and their associated utilities $u(\alpha, i)$ for all agents i and then compute the best action:

$$argmax_{\alpha} \sum_{\alpha,i} p_i(\alpha) \cdot u(\alpha, j)$$

This way of determining the action that maximizes overall utility (the "overall good") is closely related to policy-based decision algorithms based on *Partially Observable Markov Decision*

Processes (POMDPs), which select the best action given the available knowledge the robot has. The main difference between consequentialism and such algorithms is that the consequentialist robot would have to compute not only its own discounted utilities but also those of the relevant in-group (a necessary restriction on the notion of "all" agents' utility). However, at least two significant challenges arise. First, a well-known problem for consequentialist models independent of robotic implementations is how to handle the knowledge limitations any agent has (i.e., knowing *how* good an action will be for others, how many others to take into considerations, etc.). Second, there are open questions about how, in the robot's representational system, moral values should be figured into utilities and traded off with the costs of all possible actions (Scheutz, 2014).

Overall, the main problem associated with implementing philosophical ethical theories is that there is still no consensus among philosophers about which approach is the normatively correct one. And since the different theories sometimes make different recommendations for how to act in certain situations, one would have to take a philosophical moral stance to decide which recommendation to follow. Moreover, whether a robot adopting the chosen system would be acceptable to community members is entirely unclear, as none of the ethical theories claim to be, or have shown to be, correct descriptions of human moral psychology and thus of human expectations of robot moral psychology (Powers, 2013).

Implementing Legal Theories

Another option of equipping a robot with ethical behavior is to implement the most systematic agreed-on moral principles in a society: the laws defined by the legal system. For social robots interacting with humans one could, for example, focus on the four bedrock norms specified in the US tort law, the "intentional torts against the person":

1. *false imprisonment* (impeding a person's free physical movement);
2. *battery* (harmful or offensive bodily contact);
3. *assault* (putting someone in a position where they perceive harmful or offensive contact to be imminent, even if no battery occurs); and
4. *intentional infliction of emotional distress* (extreme and outrageous conduct that causes severe distress).

One could then carefully examine the legal definitions of these torts and distill the ingredients needed for a robot to determine when, say, harmful contact and thus battery might occur (Mikhail, 2014). Such an approach would require the definition of possible circumstances and behaviors that would trigger such legal principles, which would then have to be implemented in the robotic system. Part of the effort would be to make legal terms such as "intent" or "imminent" or "distress" computational—that is, provide algorithms that detect intent, perceptions of imminence, or distressed emotional states. Moreover, the legal concept of a "rational person" would have to be formalized to be able to use it in cases where the law specifically refers to the decisions and actions performed by a rational person. It is currently unclear how this could be done without requiring robot designers to solve the "AI problem"—without having to replicate human-like understanding and reasoning capabilities.

Implementing Human-Like Moral Competence

The third approach does not implement an ethical theory or a set of legal principles in a robot. Instead, it analyzes the various capacities that make up human moral competence and attempts to replicate at least some of these capacities in machines, without necessarily replicating all of them (or replicating all of human cognition). On this approach one might investigate, for example, how humans learn, represent, and reason about *moral norms*, and once a sufficient empirical

understanding of the hypothesized norm capacity is available, one could develop computational models of learning, representing, and reasoning about norms that could be integrated into robotic architectures (Malle et al., 2017). Such models would allow robots not only to behave in human-like ways (with respect to the particular capacity) but also to make reasonable predictions about human behavior that is guided by this capacity—such as when and how humans acquire new norms or under what circumstances they might break norms. Such modeling can significantly improve human-robot interactions because the robot can better adapt to the interaction and the human would feel better understood. These benefits are difficult to obtain with the other two approaches.

Another advantage is that the kinds of moral competences under consideration go far beyond a list of principles, mechanisms, or laws. Obviously, moral robots need to have a sophisticated norm system (Malle et al., 2017), but they may also need to make moral judgments of behavior relative to those norms and engage in moral communication—from explaining one's actions to expressing moral criticism to accepting an apology. Human moral competence is a cognitive as well as a social phenomenon.

However, attempting to implement human-like moral competence is challenging, for it is not yet clear exactly what perceptual, cognitive, affective, communicative and behavioral components underwrite human moral competence (Cushman et al., 2010; Guglielmo, 2015; Malle and Scheutz, 2014). For example, is it necessary to be able to simulate another person's decision-making in order to judge whether that person behaved morally? Are affective responses essential ingredients of moral judgment and decision making? And is the highly context-specific human norm system logically inconsistent, which might make it computationally intractable? Moreover, there are important ethical questions as to whether we should attempt to replicate

human morality in a machine. Human moral behavior can be suboptimal at times, and one might expect robots to be morally superior, i.e., show *supererogatory* performance (Scheutz and Arnold, 2016b). However, we need to differentiate replicating moral *competence* from replicating moral *performance.* Known sources of human performance decrements — such as strong intense affect, personal stakes, and group identity — can be explicitly omitted in designing moral robots. Few people would consider a robot less genuinely moral if it didn't get angry, selfish, or prejudiced. In fact, humans might look to such robots as reminders or models of norms and behaviors they would under normal circumstances fully endorse. In addition, replicating moral competence is a functional notion, leaving ample room for distinct implementations of the competence depending on the specific properties of the organism or platform.

Regardless of which approach for realizing ethical behavior will be taken, it is critical to ensure that the robots' moral decisions are understandable to people, especially if those decisions do not perfectly match people's own expectations or preferences. Without such understanding people would not trust robots and would be unwilling to collaborate with them.

Approaches towards Developing Moral Artificial Agents

Much of the discussion on what it takes for robots to count as moral has occurred outside the fields of robotics (Bringsjord and Taylor, 2012; Kahn, Jr. et al., 2006; Sullins, 2006; Wallach and Allen, 2008). Additionally, some scholars within the cognitive systems community have set out to build cognitive architectures in which to model human moral decision-making (Blass and Forbus, 2015; Dehghani et al., 2008). For example, Blass and Forbus (2015) showed how analogical reasoning can be used to apply previously learned moral judgments to novel scenarios. In addition, some in the logic-based community have started to investigate normative

reasoning in single agent and multi-agent systems (e.g., (Ågotnes et al., 2007; Andrighetto et al., 2010; Pereira and Saptawijaya, 2009).

One of the most prominent proposals for developing architectures explicitly incorporating mechanisms for ethical behavior in robots is an extended version of the *Autonomous Robot Architecture* AuRA (Arkin and Balch, 1997). Augmented by an *ethical governor*, a *responsibility advisor*, and an *ethical adaptor*, the system allows for modifications of the robot's behavioral repertoire in case unethical behaviors are observed. Specifically, the ethical adaptor uses a scalar "guilt" value that monotonically increases over time as unanticipated ethical violations are detected by the system (Arkin and Ulam, 2009); as a result, actions with harmful potential are subsequently disallowed. The current system can handle only very specific, hard-coded moral decisions, but it can also advise human operators ahead of a mission about possible ethical conflicts in a limited way (Arkin et al., 2009). It does, however, lack the formal representations of norms, principles, values, etc. to allow it to perform general ethical inferences and reason through normative conflicts.

Similarly, the mechanisms proposed by (Briggs and Scheutz, 2013, 2015) for the robotic *Distributed Integrated Affect Reflection and Cognition* (DIARC) architecture (Scheutz et al., 2006) can detect potential norm violations that would result from carrying out human instructions that are in conflict with given normative principles. In that case, the robot can engage the human operator in a brief dialogue about why it is not permitted to carry out the instruction and offer a justification for its refusal. Different from the ethical extensions to the AuRA architecture, the DIARC extension is based on general inference algorithms that work with *explicit* representations of normative principles. However, the current system can handle only simple potential, but no actual norm conflicts (i.e., conflicts that could arise if it were to

follow a particular command and execute an action that would be in conflict with its existing principles). Moreover, it cannot yet acquire new norms or principles from interactions and observations.

Research and development of mechanisms for ensuring normative behavior in autonomous robots has just begun, but it is poised to expand, judging from the increasing number of workshops and special sessions devoted to robot ethics and related topics (Malle, 2015). The prospects of autonomous weapon systems have fueled discussion and spurred the development of systems capable of making ethically licensed decisions, but other morally charged applications (e.g., robots for eldercare or robots for sex) have come into focus and are likely to contribute to a broadening of the discussion and the efforts to design robots with moral capacities.

Conclusion

Moral robots are necessary to ensure effective and safe human-robot interactions. Given that a massive deployment of social robots in human societies is already predictable, we need to start developing algorithms and mechanisms for such robots to meet human expectations of moral competence and behave in ethical ways. We must determine what capacities are needed for the wide variety of tasks that social robots are expected to take on, and we must implement such capacities within one of the three paradigms discussed above—the philosophical, the legal, or the psychological. Each of the paradigms has strengths and weaknesses, and perhaps some symbiotic combination can be found in the near future. We are at the beginning of a long path towards developing machines that have moral capacities. Yet, it is clear that we have to take this route if we want to ensure that robot technology will serve humanity and not emerge as one of its primary threats.

References

Ågotnes, T., Hoek, W.V.D., Rodriguez-Aguilar, J.A., Sierra, C. and Wooldridge, M. (2007), "On the Logic of Normative Systems", *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI '07)*, pp. 1181–1186.

Andrighetto, G., Villatoro, D. and Conte, R. (2010), "Norm Internalization in Artificial Societies", *AI Communications*, Vol. 23 No. 4, pp. 325–339.

Arkin, R.C. (2009), *Governing Lethal Behavior in Autonomous Robots*, CRC Press, Boca Raton, FL.

Arkin, R.C. and Balch, T. (1997), "AuRA: Principles and Practice in Review", *Journal of Experimental and Theoretical Artificial Intelligence*, Vol. 9 No. 2, pp. 175–189.

Arkin, R.C. and Ulam, P. (2009), "An Ethical Adaptor: Behavioral Modification Derived from Moral Emotions", *Computational Intelligence in Robotics and Automation (CIRA), 2009 IEEE International Symposium on*, IEEE, pp. 381–387.

Arkin, R.C., Wagner, A.R. and Duncan, B. (2009), "Responsibility and lethality for unmanned systems: Ethical pre-mission responsibility advisement", *Proceedings of the 2009 IEEE Workshop on Roboethics*, Georgia Institute of Technology.

Asaro, P.M. (2011), "Remote-Control Crimes", *Robotics & Automation Magazine, IEEE*, Vol. 18 No. 1, pp. 68–71.

Asaro, P.M. (2012), "A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics", in Lin, P., Abney, K. and Bekey, G. (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, pp. 169–186.

Asimov, I. (1942), "Runaround", *Astounding Science Fiction*.

Bartneck, C., Verbunt, M., Mubin, O. and Al Mahmud, A. (2007), "To Kill a Mockingbird Robot", *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, presented at the HRI '07, ACM Press, New York, NY, pp. 81–87.

Blass, J.A. and Forbus, K.D. (2015), "Moral Decision-Making by Analogy: Generalizations versus Exemplars", *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pp. 501–507.

Bonnefon, J.-F., Shariff, A. and Rahwan, I. (2016), "The Social Dilemma of Autonomous Vehicles", *Science*, Vol. 352 No. 6293, pp. 1573–1576.

Briggs, G. and Scheutz, M. (2013), "A Hybrid Architectural Approach to Understanding and Appropriately Generating Indirect Speech Acts", *Proceedings of Twenty-Seventh AAAI Conference on Artificial Intelligence*, pp. 1213–1219.

Briggs, G. and Scheutz, M. (2014), "How Robots Can Affect Human Behavior: Investigating the Effects of Robotic Displays of Protest and Distress", *International Journal of Social Robotics*, Vol. 6 No. 2, pp. 1–13.

Briggs, G. and Scheutz, M. (2015), "'Sorry, I Can't Do That:' Developing Mechanisms to Appropriately Reject Directives in Human-Robot Interactions", *Proceedings of the 2015 AAAI Fall Symposium on AI and HRI*.

Bringsjord, S. and Taylor, J. (2012), "The Divine-Command Approach to Robot Ethics", in Lin, P., Bekey, G. and Abney, K. (Eds.), *Anthology on Robo-Ethics*, MIT Press.

Chatterjee, P. (2015), "Is Drone Warfare Fraying at the Edges?", *Www.tomdispatch.com*, 8 March, available at: http://www.tomdispatch.com/post/175964/tomgram%3A_pratap_chatterjee,_is_drone_warfare_fraying_at_the_edges/ (accessed 24 July 2016).

Coeckelbergh, M. (2010), "Robot Rights? Towards a Social-Relational Justification of Moral Consideration", *Ethics and Information Technology*, Vol. 12 No. 3, pp. 209–221.

Cushman, F., Young, L. and Greene, J.D. (2010), "Multi-System Moral Psychology", *The Moral Psychology Handbook*, Oxford University Press, Oxford, UK.

Darling, K., Nandy, P. and Breazeal, C. (2015), "Empathic Concern and the Effect of Stories in Human-Robot Interaction", *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, IEEE, pp. 770–775.

Dehghani, M., Tomai, E., Iliev, R. and Klenk, M. (2008), "MoralDM: A Computational Modal of Moral Decision-Making", *Proceedings of the 30th Annual Conference of the Cognitive Science Society (CogSci)*, Washington, D.C.

Gert, B. (2005), *Morality: Its Nature and Justification*, Revised., Oxford University Press, New York, NY.

Gips, J. (1995), "Toward the Ethical Robot", in Ford, K.M., Glymour, C. and Hayes, P.J. (Eds.), *Android Epistemology*, MIT Press, Cambridge, MA, USA, pp. 243–252.

Goble, L. (2005), "A logic for deontic dilemmas", *Journal of Applied Logic*, Vol. 3 No. 3–4, pp. 461–483.

Guglielmo, S. (2015), "Moral judgment as information processing: An integrative review", *Frontiers in Psychology*, Vol. 6, available at: http://doi.org/10.3389/fpsyg.2015.01637.

Gunkel, D.J. (2014), "A Vindication of the Rights of Machines", *Philosophy & Technology*, Vol. 27 No. 1, pp. 113–132.

Gunkel, D.J., Bryson, J.J. and Torrance, S. (Eds.). (2012), *The Machine Question: AI, Ethics and Moral Responsibility*, The Society for the Study of Artificial Intelligence and Simulation of Behaviour.

Hamlin, J.K. (2013), "Moral Judgment and Action in Preverbal Infants and Toddlers: Evidence for an Innate Moral Core", *Current Directions in Psychological Science*, Vol. 22 No. 3, pp. 186–193.

Heider, F. and Simmel, M. (1944), "An Experimental Study of Apparent Behavior", *The American Journal of Psychology*, Vol. 57 No. 2, pp. 243–259.

Kahn, Jr., P.H., Ishiguro, H., Friedman, B. and Kanda, T. (2006), "What Is a Human? Toward Psychological Benchmarks in the Field of Human-Robot Interaction", *The 15th IEEE International Symposium on Robot and Human Interactive Communication, 2006. ROMAN 2006*, presented at the 15th IEEE International Symposium on Robot and Human Interactive Communication, 2006. ROMAN 2006, pp. 364–371.

Kahn, Jr., P.H., Kanda, T., Ishiguro, H., Freier, N.G., Severson, R.L., Gill, B.T., Ruckert, J.H., et al. (2012), "'Robovie, You'll Have to Go into the Closet Now': Children's Social and Moral Relationships with a Humanoid Robot", *Developmental Psychology*, Vol. 48 No. 2, pp. 303–314.

Malle, B.F. (2015), "Integrating Robot Ethics and Machine Morality: The Study and Design of Moral Competence in Robots", *Ethics and Information Technology*, available at:http://doi.org/10.1007/s10676-015-9367-8.

Malle, B.F. and Scheutz, M. (2014), "Moral Competence in Social Robots", *Proceedings of IEEE International Symposium on Ethics in Engineering, Science, and Technology, Ethics'2014*, IEEE, Chicago, IL, pp. 30–35.

Malle, B.F. and Scheutz, M. (2016), "Inevitable Psychological Mechanisms Triggered by Robot Appearance: Morality Included?", *2016 AAAI Spring Symposium Series Technical Reports SS-16-03*, AAAI Press, Palo Alto, CA, pp. 144–146.

Malle, B.F., Scheutz, M., Arnold, T., Voiklis, J. and 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.*, ACM, New York, NY, pp. 117–124.

Malle, B.F., Scheutz, M. and Austerweil, J.L. (2017), "Networks of Social and Moral Norms in Human and Robot Agents", in Ferreira, M.I.A., Sequeira, J.S., Tokhi, M.O., Kadar, E. and Virk, G.S. (Eds.), *A World with Robots*, Springer, Berlin/Heidelberg, Germany.

Malle, B.F., Scheutz, M., Forlizzi, J. and Voiklis, J. (2016), "Which Robot Am I Thinking About? The Impact of Action and Appearance on People's Evaluations of a Moral Robot", *Proceedings of the Eleventh Annual Meeting of the IEEE Conference on Human-Robot Interaction, HRI'16*, IEEE Press, Piscataway, NJ, pp. 125–132.

Mikhail, J. (2014), "Any Animal Whatever? Harmful Battery and Its Elements as Building Blocks of Moral Cognition", *Ethics*, Vol. 124 No. 4, pp. 750–786.

Monroe, A.E., Dillon, K.D. and 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*, Vol. 27, pp. 100–108.

Moor, J.H. (2006), "The Nature, Importance, and Difficulty of Machine Ethics", *IEEE Intelligent Systems*, Vol. 21 No. 4, pp. 18–21.

Moor, J.H. (2013), "Four Kinds of Ethical Robots", *Philosophy Now*, No. 72.

Murphy, R. and Woods, D.D. (2009), "Beyond Asimov: The Three Laws of Responsible Robotics", *IEEE Intelligent Systems*, Vol. 24 No. 4, pp. 14–20.

Pagallo, U. (2011), "Robots of Just War: A Legal Perspective", *Philosophy & Technology*, Vol. 24 No. 3, pp. 307–323.

Pereira, L.M. and Saptawijaya, A. (2009), "Modelling Morality with Prospective Logic", *International Journal of Reasoning-Based Intelligent Systems*, Vol. 1 No. 3/4, pp. 209–221.

Powers, T.M. (2013), "Machines and Moral Reasoning", *Philosophy Now*, No. 72, available at: https://philosophynow.org/issues/72/Machines_and_Moral_Reasoning (accessed 26 July 2016).

Premack, D. (1990), "The Infant's Theory of Self-Propelled Objects.", *Cognition*, Vol. 36 No. 1, pp. 1–16.

Scheutz, M. (2012), "The Inherent Dangers of Unidirectional Emotional Bonds Between Humans and Social Robots", in Lin, P., Bekey, G. and Abney, K. (Eds.), *Anthology on Robo-Ethics*, MIT Press, Cambridge, MA, pp. 205–221.

Scheutz, M. (2014), "The Need for Moral Competency in Autonomous Agent Architectures", in Müller, V.C. (Ed.), *Fundamental Issues of Artificial Intelligence*, Springer, Berlin, pp. 515–525.

Scheutz, M. and Arnold, T. (2016a), "Are We Ready for Sex Robots?", *Proceedings of the 11th ACM/IEEE International Conference on Human-Robot Interaction*.

Scheutz, M. and Arnold, T. (2016b), "Feats Without Heroes: Norms, Means, and Ideal Robotic Action", *Frontiers in Robotics and AI*, Vol. 3, available at:http://doi.org/10.3389/frobt.2016.00032.

Scheutz, M. and Malle, B.F. (2014), "'think and Do the Right Thing': A Plea for Morally Competent Autonomous Robots.", *Proceedings of the IEEE International Symposium on Ethics in Engineering, Science, and Technology, Ethics'2014*, Curran Associates/IEEE Computer Society, Red Hook, NY, pp. 36–39.

Scheutz, M., Schermerhorn, P., Kramer, J. and Anderson, D. (2006), "First Steps Toward Natural Human-Like Hri", *Autonomous Robots*, Vol. 22 No. 4, pp. 411–423.

Short, E., Hart, J., Vu, M. and Scassellati, B. (2010), "No Fair!! An Interaction with a Cheating Robot", *Proceedings of the 5th ACM/IEEE International Conference on Human-Robot Interaction*, Osaka, Japan.

Sparrow, R. (2007), "Killer Robots", *Journal of Applied Philosophy*, Vol. 24 No. 1, pp. 62–77.

Strait, M., Canning, C. and Scheutz, M. (2014), "Let me tell you! Investigating the Effects of Robot Communication Strategies in Advice-Giving Situations based on Robot Appearance, Interaction Modality, and Distance", *Proceedings of 9th ACM/IEEE International Conference on Human-Robot Interaction*, pp. 479–486.

Sullins, J. (2006), "When Is a Robot a Moral Agent?", *International Review of Information Ethics*, Vol. 6 No. 12, pp. 23–30.

Tremoulet, P.D. and Feldman, J. (2000), "Perception of Animacy from the Motion of a Single Object", *Perception*, Vol. 29 No. 8, pp. 943–951.

Wallach, W. and Allen, C. (2008), *Moral Machines: Teaching Robots Right from Wrong*, Oxford University Press, New York, NY.