When Will People Regard Robots as Morally Competent Social Partners?*

Bertram F. Malle, Brown University Matthias Scheutz, Tufts University

Abstract— **We propose that moral competence consists of five distinct but related elements: (1) having a system of norms; (2) mastering a moral vocabulary; (3) exhibiting moral cognition and affect; (4) exhibiting moral decision making and action; and (5) engaging in moral communication. We identify some of the likely triggers that may convince people to ascribe each of these elements of moral competence to robots. We suggest that humans will treat robots as moral agents (who have some rights, obligations, and are targets of blame) if they perceive them to have at least element (1) and one or more of elements (2)-(5).**

I. Introduction

The question posed in the title of this paper may be interpreted either temporally or conditionally: When—in the timeline of rapidly advancing progress in robotics—will people regard social robots as morally competent social partners? And under what conditions will they do so? The first question may be speculative and best left to futurists. However, if we merely wait until future societies regard robots as morally competent without examining what it would *take* for robots to be competent in this way, we miss out on a critical opportunity to design robots that are psychologically safe and that could benefit humanity not only through their technical proficiency but also through their moral and social value.

The focus of this paper will be on the conditional meaning of the title question, which pursues just this opportunity: What would it *take* for robots to be morally competent? To answer this question we first need to establish what moral competence consists of. We introduce here a framework (first developed in [1], [2] and recently expanded [3]) that integrates extant literatures on moral psychology, moral philosophy, and social cognitive science. This framework does not determine what "true" moral competence is but tries to enumerate the capacities that ordinary people expect of one another in their social relationships. At least some of these capacities people will expect of social robots as well. We must therefore, in a clearly multi-disciplinary endeavor, analyze the psychological nature of these capacities in humans, develop ways to implement at least some of them in computational architectures and physical machines, and continuously examine whether robots with such emerging moral competence are in fact suitable for and accepted as social partners [4]. The moral standing and abilities of machines will therefore emerge from and be constrained by the relations that people are willing to form with them [5].

II. Elements of Moral Competence

A. The Framework in Overview

A competence is an aptitude, a qualification, a dispositional capacity to deal adequately with certain tasks. Moral competence must, uncontroversially, deal with the task of *moral decision making and action*. From Aristotle to Kant to Kohlberg, morality has been about "doing the right thing." And recent questions about moral properties of robots have centered on decisions about life and death [6], often in situations of moral dilemmas [7], which have prominence in psychology as well [8]–[10].

But there is quite a bit more to moral competence than just moral decision making. For one, *moral cognition* has been a primary focus of recent theoretical and experimental work in psychology, examining such phenomena as judgments of permissibility, wrongness, and blame [11]–[15]. The capacity of moral cognition is engaged when an agent witnesses or interacts with another agent that performs a morally relevant behavior—behaviors that are most frequently moral norm violations but can also be ones that meet or exceed moral norms. In addition, the role of *affect and emotion* in those judgments has been widely debated and investigated [13], [16]–[19].

Further, psychologists, sociologists, and philosophers have studied how moral cognition and affect leads to *moral communication,* including socially expressing moral criticism [20]–[22] and negotiating this criticism through justifications, excuses, and apologies [23]–[27].

The criteria by which moral decision making is evaluated and the standards against which morally relevant behavior is assessed are *moral norms*; so having and mastering a system of norms is a necessary requirement for moral competence. One might argue that decision making, judgment, and communication are all made *moral* by virtue of their reliance on and their intertwinement with moral norms.

Finally, one other foundational, perhaps obvious element of moral competence is having a *moral vocabulary* that

* This project was supported in part by a grant from the Office of Naval Research, No. N00014-14-l-0144. The opinions expressed here are our own and do not necessarily reflect the views of ONR.

Bertram F. Malle is with the Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, 190 Thayer St., Providence, RI 02912 USA (phone: 401-863-6820; fax: 401-863-2255; e-mail: bfmalle@ brown.edu).

Matthias Scheutz is with the Department of Computer Science, Tufts University, 161 College Avenue, Medford, MA 02155 USA (e-mail: matthias.scheutz@tufts.edu).

allows the agent to represent norms, use them in judgments and decisions, and communicate about them.

These five key elements of moral competence are depicted in Fig. 1, ordered tentatively from prerequisites (on the bottom) to those that build on the prerequisites (top).

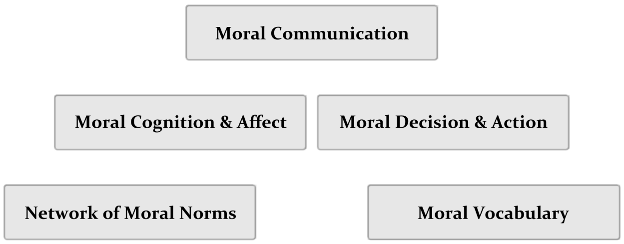

Figure 1. The five constitutive elements of moral competence, ordered from elements that are prerequisites (bottom) to elements that build on the latter (top).

B. Two Corollaries of the Framework

Two important claims follow from this framework of moral competence:

First, moral competence is not to be equated with “moral agency”—the topic most heavily discussed in the current social robotics literature [28]–[30]. Scholars have suggested a variety of criteria for being an *agent*, including embodiment, consciousness, soul, free will—criteria that raise more conceptual questions than they are intended to answer [31]. Similarly, asking what makes an agent *moral* leads to perhaps even more difficult problems. Often a moral agent is characterized as an entity that can act according to what is right and wrong [28], [29] or could be held responsible for its actions [30], which mixes several elements together: capacity for decision and action; mastery of norms; and properties that invite or justify others to morally judge the agent’s behavior. Keeping these elements separate offers a cleaner approach to the abstract notion of moral agency, and it enables us to accept that, in various applications of social robots (e.g., for health and social assistance or for safety and security), a moral agent in one application may look different from a moral agent in another application.

Second, an entity may be partially morally competent [32], [33] by having, say, a capacity for moral judgment but not moral decision making, or the capacity for moral judgment and decision making without the capacity for moral communication. In addition, by investigating in detail what makes up each element of moral competence and what convinces people to ascribe a given element to another agent (especially robots), we can also begin to explicate how at least some capacities can come in degrees, like they do in children and in patients with certain clinical deficits.

Together, these claims paint the picture of a dynamic, constantly updated endeavor of developing a “moral robot” [34], with no single target or criterion of success but with multiple possible functions and uses [35]. Contributions from theoretical, empirical, and computational scholars will have to be integrated to accommodate the goals of developing robot capacities that are indeed tailored to interactions with ordinary humans. For we need to know what capacities and demands humans themselves exhibit; we need to design the computational architectures and physical implementations of robot analogs of some of these capacities; and we need to assess whether these analogs are accepted and thus genuinely tailored to the needs of humans, especially the needs of vulnerable populations.

We now turn to the main goal of this contribution: identifying some of the likely features that may convince people to ascribe each of these elements of moral competence to future robots.

III. A NETWORK OF NORMS

Having a moral norm system is a prerequisite for other elements of moral competence and may be demonstrated most convincingly in the presence of one of those other elements—most obviously when a robot takes certain norms into account while making a moral decision or when it detects and evaluates a norm violation. From a design perspective, however, building a norm network must precede the full development of other capacities. Unfortunately, the empirical literature does not provide much guidance for how norm networks are represented in the human mind. What we know, from reflection and limited research [36]–[38], is that norms exemplify a unique set of properties: They are: (A) activated in highly context-specific ways; (B) consistently updated; and (C) organized in flexible hierarchies.

A. Context Specificity

Context specificity is a vexing computational problem [39], but humans can recognize contexts by being sensitive to a bundle of context-defining stimuli, among them physical spaces (e.g., office vs. bathroom), temporal markers (morning vs. evening), roles (boss vs. employee), relationships (stranger vs. friend), and goal projects (e.g., discussion vs. vote tallying in a business meeting). Each of these stimuli serve as cues for a bundled set of norms and norm-conforming habits.

B. Continuous Updating

Though humans are often described as cognitively conservative (holding to beliefs in the face of counterevidence [40], [41]), they seem to be finely attuned to changes in norms across time and contexts, following observations of other people’s expectations, their sanctioning behavior, and the reliability of norm conformity. Anybody can experience this partial updating of one’s norm system by entering a new culture, which involves high demands on processing exactly these data: the locals’ (often implicit) expectations, anticipated or experienced sanctions, and the prevalence and reliability of norm-conforming behavior.

C. (Flexible) Hierarchical Organization

The norm system contains concrete behavioral rules (e.g., don’t eat with your fingers), mid-level principles (e.g., be polite to the elderly), and abstract values (e.g., respect, fairness). These levels are connected vertically such that mid-level principles implement values, and behavioral rules implement mid-level principles (and values). In addition, at each level, some norms are more important than others.

Despite this double hierarchy, we can be certain that norms do not occupy fixed positions in this hierarchy. Each context activates subsets of norms and slightly rearranges the hierarchy for this subset—adjusting which lower-level norms instantiate which higher-level ones and which norms, at a given level, are more important in this context than others.

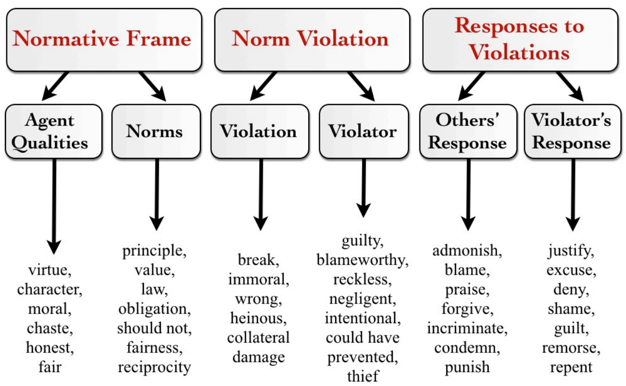

Figure 1. A sketch of moral vocabulary, displaying three major ontological categories, major subcategories, and a small sample of word instances under each.

From these characteristics, we can formulate a few features that may invite humans to grant an artificial agent some mastery of norms: (A) when the agent knows which norms to apply in what context; (B) when the agent can learn new rules or adjustments to familiar rules in new contexts; and (C) when the agent computes vertical norm instantiations and horizontal norm orderings with some (context-dependent) flexibility. None of these capacities is out of reach for current AI, especially if restricted to a manageable set of contexts and norms, such as elderly home care.

IV. Moral Vocabulary

Morally competent adults need a vocabulary to represent (conceptually and linguistically) the norms of their community and to teach, learn, and reason about these norms. They also need this vocabulary to express and instantiate various moral practices—including blaming, justifying, excusing, acquitting, or forgiving. Once more, the higher-level capacities of moral decision making, judgment, and communication would be compelling indicators of an AI's mastery of a moral vocabulary. But there should be easier, earlier guide posts en route to such high-level mastery. An initial design step would be to build a structured set of keywords as an "ontology" that is able to categorize and interconnect large numbers of words in text mining. According to our initial research, such keywords fall into three ontological categories, with at least one level of subcategories (see Fig. 2): The normative frame (with vocabulary for agent qualities as well as for norms); the norm violation (with vocabulary for the violation as well as for the violator); and responses to violations (with vocabulary for others' responses as well as for the violator's response).

V. 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. At the basic level, people's well-practiced norm network allows them to quickly detect violations (e.g., a dead body on the street), leading to a judgment of badness or wrongness. It takes more complex information processing, however, to form a judgment about the agent who committed the violation. Most prominently, judgments of blame take into account the agent's causal contributions, intentionality and mental states, and counterfactuals about what the agent should have and could have done differently [14].

What would it take for a robot to credibly engage in moral cognition of morally significant events? Assuming the robot is equipped with a norm network and moral vocabulary as discussed earlier, it needs to be able to segment event streams and identify those events (behaviors and states) that violate one or more of the relevant norms. At a minimum, this identification process would have to succeed for verbal event streams by classifying sentences as norm violating or not. The more sophisticated the machine's norm system, the more refined the classifications, including fine differentiations into cases where the same behavior is acceptable in one context but unacceptable in another. The recognition of context and recruitment of context-relevant norms is the biggest challenge here, but adapting a parser to search for physical, temporal, role, etc. information and activate norm bundles relevant to those situations is not an insurmountable problem. Consider three norm violations:

(1) Sarah faked an injury after an automobile accident.

The phrase "after an automobile accident" should be a strong trigger for a norm bundle, though it is unclear whether "after" means right after the accident or days and weeks later. The behavior of "faked an injury" co-occurring with "auto accident" may be sufficient to trigger insurance contexts and their associated prohibitions against faking injury.

(2) Paul got fired and, in response, entered the personnel manager's office and shot her.

Two of the three component events ("got fired" and "entered the personnel manager's office") are not norm violations, but the third clearly is. Verbs such as "shot" will have high priors for severe norm violations no matter the context, and the AI may search for mitigating information in such a case. In fact, the preceding two component events would have to be considered as potential justifications, but neither should be acceptable (whereas, for example, "the personnel manager held two hostages in her office" might).

Beyond sentences, segmentation of visual events would of course be more impressive. Identifying the relevant event within such rich stimulus arrays is likely to be more difficult, but culling context cues may be easier, especially if scenes are restricted to environments in which the robot is actually going to interact with humans (e.g., in a hospital room, an apartment, an office building).

If people observed a robot do reasonably well in this classification task, they may have some trust in it as a security monitor, crime detector, and similar moral cognition roles. Trusting it as an autonomous decision maker will take more (see below).

An agent that detects norm violations does not necessarily have the capacity to make sophisticated moral judgments. To convincingly demonstrate agent-directed judgments such as blame, additional information detection and integration would be expected: taking into account causal contributions, intentionality and mental states, and counterfactual reasoning. At the verbal level, a simplified approach might succeed that codes sentence components for these factors. Consider the following example:

(3) Sharon took a t-shirt out of the store without paying.

(3.1) She was homeless and needed a shirt.

(3.2) She forgot to take it off before leaving the fitting room.

About 75 percent of people judge the conjunctive behavior in (3), taking a t-shirt and not paying, as intentional. The remaining uncertainty is resolved in favor of intentionality in (3.1) by the phrase "needed a shirt," which reveals itself as the reason for intentionally stealing it. In (3.2), however, lack of intentionality is favored by the phrase "forgot to take it off," whose major verb implies that taking the t-shirt out of the store without paying was unintentional. A credible AI analysis would have to recognize the reasonably high intentionality prior in (3) but then resolve the uncertainty differently when faced with either (3.1) or (3.2). Moreover, the norm violation of "stealing" with "t-shirt" as the object would have to be assigned a higher disapproval value than "unintentional taking away the t-shirt."

And where is affect? The literature on human psychology is rather unclear over the exact role that affect and specific emotions play in moral judgment. 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: Some say it simply marks that something important occurred [42]; other suggest that it motivates the perceiver to find the cause of the bad event [43]; yet others warn that such affect biases the search for evidence that specifically enables the perceiver to blame somebody [11]. However, people can make moral judgments without much affect at all [44], and currently no compelling evidence supports the claim that affective phenomena are necessary or constitutive of those judgments [18], [45].

So would robots need to show any affect as part of their moral judgments? If artificial agents can approximate human judgments in their sensitivity to critical information (i.e., severity of norm violation, causality, intentionality, etc.), their absence of affective responses will be of little relevance. A problem may arise, however, in the *communication* of those moral judgments. A coldly stated assessment, "He deserves a significant amount of blame for hitting the child in the face," could upset a human partner. That is because people expect that community members not only adhere to shared norms but also censor those who violate those norms, and do so with appropriate displays of concern or outrage [21], [46]. Failure to express moral criticism with appropriate intensity may be itself a norm violation and can be interpreted as a defect in the nonexpressive moral judge. Some scaled intensity of expressed moral judgment may thus be necessary for robots to be acceptable social partners.

VI. Moral Decision and Action

Having a norm network and using it to detect potential norm violations does not ipso facto allow an agent to integrate norms into complex action planning. Young children are able to detect a number of norm violations (in part because of their uncanny ability for statistical learning [47], but they often have a difficult time integrating norms into their own action planning. Individuals on the autism spectrum, too, are able to detect violations of norms [48], [49], but their everyday behavior often breaks through norms and conventions.

Competent moral action requires decision making, not just imitating norm-conforming behavior. Robots that meet expectations for moral decision making must exhibit what human partners conceptualize as *reasoned choice*. Some skeptics of the possibility of moral competence in robots assume that reasoned choice presupposes some kind of nondeterministic "free will" [50], [51]. 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 [52], [53]. If a robot acquires and uses knowledge about the world to guide its actions in line with its goals, it effectively displays this capacity for choice and intentional action [34]. When such a robot commits a norm violation, people readily assign blame to it—in simulated scenarios [54], [55] and actual interaction [56]. 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 for being granted the ability to make competent moral decisions; metaphysical free will is not.

Clear signs of reasoned choice capacity include representing the problem at hand (e.g., which goals are given, which action options are available), searching for relevant information (e.g., about means to achieve those actions and their consequences), integrating the information through appropriate weighting of values and probabilities, and settling on a course of action—firm, though revisable in light of new information. In effect, moral decision making is no different from other decision making, except that action selection is guided and constrained by a system of moral norms.

LaChat [57] argued that a robot with moral decision making capacity must have "empathy", the ability to feel the pain of others. But the significance of others' pain does not lie merely in feeling it. When somebody feels the pain of others and nonetheless acts immorally, we surely would think something went wrong. The hidden assumption here is that ordinary humans, *because* they feel the pain of others, make moral decisions that minimize others' pain. So if an agent makes moral decisions that minimize others' pain through some other means—for example, by careful consideration of states of the world and emotional states of individuals—few people would insist that this agent doesn't act morally. Once more, to be trusted and accepted, a robot may have to communicate to human observers (verbally or through physical expression) that it *values* things [58], [59], that it *cares* about certain outcomes [60]. To this end we need not invent a unique computational structure (a "caring state") but rather analyze carefully what kinds of behaviors people treat as diagnostic of caring: willingness to prioritize, attend to, invest in, help, and so on. If the robot's caring is clearly revealed in its actions, those actions will count more than the specific internal structure that brings them about.

VII. MORAL COMMUNICATION

As important as the cognitive tools are that enable moral cognition and moral decision making, they are not sufficient to achieve the socially most important function of morality: to engage in regulating each other's behavior. For that, moral communication is needed. It should be feasible for a robot to express its own moral judgments and decisions to others, provided it has well-developed natural language skills. However, social obstacles may stand in the way of robots being welcomed as moral communicators. For one thing, people will regard robots as low in status, and those low in status are not always free to voice their moral criticism. So robots will need to be aware of the norms of blaming [14] and sometimes hold back social acts of moral criticism unless, for example, safety concerns override those norms. Similarly, robots embedded in teams of search and rescue, military, or police may have to earn a level of trust that licenses them to monitor and enforce norms, but if they do, they could strengthen the ethical standards of those teams [6].

Besides expressing moral judgments, morally competent robots would also explain their own behaviors to others. Following the importance of intentionality discussed earlier, the robot would have to distinguish between its own intentional actions (which it executed the way it planned them) and its unintentional, accidental behaviors (which occurred as deviations from its planning process), most prominently collateral damage. People expect very different kinds of explanations for intentional and unintentional behaviors [61], and robots would have to mirror these differences in order to be understood and accepted. That is, explaining intentional moral violations would require offering reasons that justify the violating action, whereas explaining unintentional moral violations would require offering causes that excuse one's involvement in the violation [14]. 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. Simulating past and future possible worlds may be computationally feasible [62], but running such a system will be extremely challenging unless the causal and normative domains of consideration are constrained in some way and their distributions learned through repeated exposure.

VIII. CODA

In a compelling study of a reasonably autonomous robot interacting with toddlers in a child care setting, children initially responded very differently to the robot than to other children but treated the robot as a peer after a few months [63]. Even if emerging and future robots are no better than "reasonably" autonomous, people's repeated interactions with them will improve the robots' capacities and bring into relief under what conditions people will treat the robot as morally component. People's expectation that robots *should* be morally competent will come into relief even earlier. We hope that robot designers will anticipate this expectation and recognize moral competence as one of the necessary attributes of future social robots.

REFERENCES

[1] B. F. Malle and M. Scheutz, "Moral competence in social robots," in *IEEE International Symposium on Ethics in Engineering, Science, and Technology*, Chicago, IL, 2014, pp. 30–35.

[2] M. Scheutz and B. F. Malle, "'Think and do the right thing': A plea for morally competent autonomous robots.," in *IEEE International Symposium on Ethics in Engineering, Science, and Technology, June*, Chicago, IL, 2014, pp. 36–39.

[3] B. F. Malle, "Moral competence in robots?," in *Sociable robots and the future of social relations: Proceedings of Robo-Philosophy 2014*, J. Seibt, R. Hakli, and M. Nørskov, Eds. Amsterdam, Netherlands: IOS Press, 2014, pp. 189–198.

[4] M. Fridin, "Kindergarten social assistive robot: First meeting and ethical issues," *Computers in Human Behavior*, vol. 30, pp. 262–272, Jan. 2014.

[5] M. Coeckelbergh, "Robot rights? Towards a social-relational justification of moral consideration," *Ethics Inf Technol*, vol. 12, no. 3, pp. 209–221, Sep. 2010.

[6] R. C. Arkin, "The case for ethical autonomy in unmanned systems," *Journal of Military Ethics*, vol. 9, no. 4, pp. 332–341, 2010.

[7] P. Lin, "The ethics of autonomous cars," *The Atlantic*, 08-Oct-2013. [Online]. Available: http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/. [Accessed: 30-Sep-2014].

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

[9] F. Cushman and L. Young, "The psychology of dilemmas and the philosophy of morality," *Ethical Theory and Moral Practice*, vol. 12, no. 1, pp. 9–24, 2009.

[10] J. F. Christensen and A. Gomila, "Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review," *Neuroscience & Biobehavioral Reviews*, vol. 36, no. 4, pp. 1249–1264, Apr. 2012.

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

[12] F. Cushman, "Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment," *Cognition*, vol. 108, no. 2, pp. 353–380, Aug. 2008.

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

[14] B. F. Malle, S. Guglielmo, and A. E. Monroe, "A theory of blame," *Psychological Inquiry*, vol. 25, no. 2, pp. 147–186, 2014.

[15] J. M. Mikhail, *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. New York, NY: Cambridge University Press, 2011.

[16] M. D. Alicke, "Blaming badly," *Journal of Cognition and Culture*, vol. 8, pp. 179–186, Apr. 2008.

[17] B. Monin, D. A. Pizarro, and J. S. Beer, "Reason and emotion in moral judgment: Different prototypes lead to different theories," in *Do emotions help or hurt decision making? A hedgefoxian perspective*, K. D. Vohs, R. F. Baumeister, and G. Loewenstein, Eds. New York, NY: Russell Sage Foundation, 2007, pp. 219–244.

[18] B. Huebner, S. Dwyer, and M. Hauser, "The role of emotion in moral psychology," *Trends in Cognitive Sciences*, vol. 13, no. 1, pp. 1–6, Jan. 2009.

[19] J. D. Greene, “Emotion and morality: A tasting menu,” *Emotion Review*, vol. 3, no. 3, pp. 227–229, Jul. 2011.

[20] M. McKenna, “Directed blame and conversation,” in *Blame: Its nature and norms*, D. J. Coates and N. A. Tognazzini, Eds. New York, NY: Oxford University Press, 2012, pp. 119–140.

[21] P. Drew, “Complaints about transgressions and misconduct,” *Research on Language & Social Interaction*, vol. 31, no. 3/4, pp. 295–325, Jul. 1998.

[22] C. Bennett, “The expressive function of blame,” in *Blame: Its nature and norms*, D. J. Coates and N. A. Tognazzini, Eds. New York, NY: Oxford University Press, 2012, pp. 66–83.

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

[24] M. E. McCullough, R. Kurzban, and B. A. Tabak, “Putting revenge and forgiveness in an evolutionary context,” *Behavioral and Brain Sciences*, vol. 36, no. 01, pp. 41–58, 2013.

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

[26] J. T. Tedeschi and M. Reiss, “Verbal strategies as impression management.,” in *The psychology of ordinary social behaviour*, C. Antaki, Ed. London: Academic Press, 1981, pp. 271–309.

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

[28] A. M. DeBaets, “Can a robot pursue the good? Exploring artificial moral agency.,” *Journal of Evolution & Technology*, vol. 24, no. 3, pp. 76–86, Sep. 2014.

[29] L. Floridi and J. W. Sanders, “On the morality of artificial agents,” *Minds and Machines*, vol. 14, no. 3, pp. 349–379, Aug. 2004.

[30] J. Parthemore and B. Whitby, “What makes any agent a moral agent? Reflections on machine consciousness and moral agency,” *International Journal of Machine Consciousness*, vol. 4, pp. 105–129, 2013.

[31] D. J. Gunkel, “A vindication of the rights of machines,” *Philos. Technol.*, vol. 27, no. 1, pp. 113–132, Mar. 2014.

[32] C. Allen, “The future of moral machines,” *The New York Times: Opinionator*, 25-Dec-2011. [Online]. Available: http://opinionator.blogs.nytimes.com/2011/12/25/the-future-of-moral-machines/. [Accessed: 29-Dec-2014].

[33] J. H. Moor, “The nature, importance, and difficulty of machine ethics,” *IEEE Intelligent Systems*, vol. 21, no. 4, pp. 18–21, Jul. 2006.

[34] T. M. Powers, “Incremental machine ethics,” *Robotics & Automation Magazine, IEEE*, vol. 18, no. 1, pp. 51–58, Mar. 2011.

[35] P. M. Asaro, “What should we want from a robot ethic?,” *International Review of Information Ethics*, vol. 6, no. 12, pp. 9–16, 2006.

[36] M. Tomasello and A. Vaish, “Origins of human cooperation and morality,” *Annual Review of Psychology*, vol. 64, no. 1, pp. 231–255, 2013.

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

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

[39] K. M. Ford and P. J. Hayes, *Reasoning agents in a dynamic world: The frame problem*. Greenwich, CT: JAI Press, 1991.

[40] A. G. Greenwald, “The totalitarian ego: Fabrication and revision of personal history,” *American Psychologist*, vol. 35, pp. 603–618, 1980.

[41] S. T. Fiske and S. E. Taylor, *Social cognition*, 2nd ed. New York, NY: McGraw-Hill, 1991.

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

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

[44] C. L. Harenski, K. A. Harenski, M. S. Shane, and K. A. Kiehl, “Aberrant neural processing of moral violations in criminal psychopaths,” *Journal of Abnormal Psychology*, vol. 119, no. 4, pp. 863–874, Nov. 2010.

[45] Y. R. Avramova and Y. Inbar, “Emotion and moral judgment,” *Wiley Interdisciplinary Reviews: Cognitive Science*, vol. 4, no. 2, pp. 169–178, Mar. 2013.

[46] E. Fehr and U. Fischbacher, “Third-party punishment and social norms,” *Evolution and Human Behavior*, vol. 25, no. 2, pp. 63–87, Mar. 2004.

[47] N. Z. Kirkham, J. A. Slemmer, and S. P. Johnson, “Visual statistical learning in infancy: Evidence for a domain general learning mechanism,” *Cognition*, vol. 83, no. 2, pp. B35–B42, Mar. 2002.

[48] T. Zalla and M. Leboyer, “Judgment of intentionality and moral evaluation in individuals with high functioning autism,” *Rev.Phil.Psych.*, vol. 2, no. 4, pp. 681–698, Dec. 2011.

[49] T. Zalla, L. Barlassina, M. Buon, and M. Leboyer, “Moral judgment in adults with autism spectrum disorders,” *Cognition*, vol. 121, no. 1, pp. 115–126, Oct. 2011.

[50] S. Bringsjord, “But perhaps robots are essentially non-persons.,” *Erwägen Wissen Ethik*, vol. 20, no. 2, pp. 193–195, Apr. 2009.

[51] A. M. Johnson and S. Axinn, “The morality of autonomous robots,” *Journal of Military Ethics*, vol. 12, no. 2, pp. 129–141, 2013.

[52] A. E. Monroe and B. F. Malle, “From uncaused will to conscious choice: The need to study, not speculate about people's folk concept of free will,” *Review of Philosophy and Psychology*, vol. 1, no. 2, pp. 211–224, 2010.

[53] A. E. Monroe and B. F. Malle, “Free will without metaphysics,” in *Surrounding free will*, A. R. Mele, Ed. New York, NY: Oxford University Press, 2014.

[54] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano, “Sacrifice one for the good of many? People apply different moral norms to human and robot agents,” presented at the 2015 ACM/IEEE International Conference on Human-robot Interaction (HRI '15), 2015.

[55] A. E. Monroe, K. D. Dillon, and B. F. Malle, “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, Jul. 2014.

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

[57] M. R. La Chat, “Moral stages in the evolution of the artificial superego: A cost-benefits trajectory,” in *Cognitive, emotive and ethical aspects of decision making in humans and in artificial intelligence*, vol. 2, I. Smit, W. Wallach, and G. E. Lasker, Eds. Windsor, ON, Canada: International Institute for Advanced Studies in Systems Research and Cybernetics, 2003, pp. 18–24.

[58] M. L. Littman, “Value-function reinforcement learning in Markov games,” *Cognitive Systems Research*, vol. 2, no. 1, pp. 55–66, Apr. 2001.

[59] M. Scheutz, “The affect dilemma for artificial agents: should we develop affective artificial agents?,” *IEEE Transactions on Affective Computing*, vol. 3, no. 4, pp. 424–433, 2012.

[60] A. van Wynsberghe, “Designing robots for care: Care centered value-sensitive design,” *Sci Eng Ethics*, vol. 19, no. 2, pp. 407–433, Jun. 2013.

[61] B. F. Malle, “How people explain behavior: A new theoretical framework,” *Personality and Social Psychology Review*, vol. 3, no. 1, pp. 23–48, Feb. 1999.

[62] P. Bello, “Cognitive foundations for a computational theory of mindreading,” *Advances in Cognitive Systems*, vol. 1, pp. 59–72, 2012.

[63] F. Tanaka, A. Cicourel, and J. R. Movellan, “Socialization between toddlers and robots at an early childhood education center,” *PNAS*, vol. 104, no. 46, pp. 17954–17958, Nov. 2007.