Moral Competence in Robots?
Bertram F. MALLE
Brown University

Abstract. I start with the premise that any social robot must have moral competence. I offer a framework for what moral competence is and sketch the prospects for it to be developed in artificial agents. After considering three proposals for requirements of “moral agency” I propose instead to examine moral competence as a broader set of capacities. I posit that human moral competence consists of five components and that a social robot should ideally instantiate all of them: (1) A system of norms; (2) a moral vocabulary; (3) moral cognition and affect; (4) moral decision making and action; and (5) moral communication.

Keywords. Human-robot interaction, moral psychology, social robotics

Introduction
Any robot that collaborates with, looks after, or helps humans—in short, a social robot—must have moral competence. But is moral competence in robots possible?

Answering this question requires an integrated philosophical, psychological, computational, and engineering approach. For one, we must decide what would broadly count as moral competence; for another, we must gather psychological theories and evidence to ensure that a robot’s moral competence is suitable for living among humans; finally, we must determine whether it is feasible to implement moral competence in computational architectures and actual robotic machines.

Typically scholars treat moral competence as moral agency—the capacity to act according to what is right and wrong. Moral competence certainly includes this capacity; but it goes further. Here I offer a framework for what moral competence is and sketch the prospects for it to be developed in artificial agents. The framework is grounded in conceptual and scientific work on moral psychology and is inspired by an emerging collaborative project on “Moral Competence in Computational Architectures for Robots” [1], [2].

1. Competence and Moral Competence

It is perhaps a curious fact that the English word competence originates from the Latin word competentia, which in post-classical Latin meant meeting together, agreement, but it also stems from competere, which means to compete or rival (OED). We see here the fundamental tension of human social living: between cooperation and competition. Today, competence is an aptitude, a qualification, a dispositional capacity to deal adequately with certain tasks. The question is: what tasks?

Uncontroversially, moral competence must deal with the task of moral decision making and action. From Aristotle to Kant to Kohlberg, morality has been “doing the right thing.” But there is quite a bit more.

Acknowledgement. This project was supported by a grant from the Office of Naval Research, No. N00014-13-1-0269. The opinions expressed here are our own and do not necessarily reflect the views of ONR.
In psychology, moral cognition has been the focus of post-Kohlbergian literature [3], and cognition includes such phenomena as judgments of permissibility, wrongness, blame; and to different degrees (depending on the theory), affect or emotion.

Further, psychologists, linguists, and sociologists have studied moral communication, including such phenomena as negotiating blame through justification and excuses, apology, and forgiveness.

Finally, these three components require two things in the first place: 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 the ingredients of moral competence:

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

2. Moral Norms

Morality’s function is to regulate human social behavior in contexts in which biological desires no longer guarantee individual and collective well-being [4], [5]. Human communities perform this regulation by motivating and deterring certain behaviors through the imposition of norms. Being equipped with a norm system thus constitutes a first critical element in human moral competence [6].

Yet even though having a norm system is an essential characteristic of moral competence, little is known about norms in human psychology. How are norms acquired? How are they represented in the mind? What properties do they have that allow them to be so context-sensitive and mutually adjusting as we know them to be? I will briefly examine these questions.

Though evidence is limited on the development of norms, 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 [7]: “that’s not nice! That was naughty!”; “He did something wrong.”

At the same time, children are able to induce rather general rules from concrete instances, such as “bombs hurt people” [7, p. 77] or even abstract principles such as the act-omission distinction [8]. In addition, the set of capacities labeled social cognition or theory of mind adds a powerful supportive structure for the acquisition of moral norms and moral judgments. This capacity set includes the segmentation of intentional behaviors [9]; mastery over the concept of goal [10]; the distinction between beliefs and desires [11], between desires and intentions [12], and a variety of emotions [13], as well as the rich linguistic expressions of these distinctions [14].

Representation. Another largely ignored question is how norms are represented in the human mind. Are they networks of concepts? Goal concepts—which typically are explicitly represented in robotic architectures—seem close to norm concepts; but are they really the same? Perhaps the most unique category of norms are values, and there are indications that they cannot simply be reduced to goals [15]. Moreover, if Jon
Elster is correct in claiming that “social norms provide an important kind of motivation for action that is irreducible to rationality or indeed any other form of optimizing mechanism” [16, p. 15], then a simple goal-based action control system will not do for moral robots.

Contextual activation. Another important facet of norms is that they appear to be rapidly activated, because people detect norm violations within a few hundred milliseconds [17]. How is this possible? Perhaps physical or linguistic contexts activate subsets of action-specific norms (what one is or is not permitted to do in the particular context), and with a well-tuned event parsing mechanism [18] violations can then be quickly detected. The relation of these concrete norms to more abstract norms may be constructed offline, through conversation and conceptual reorganization [19].

For designing a morally competent robot, all these features of norms present serious challenges. But if norms are kinds of representations, 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.

3. Moral Vocabulary

Some rudimentary moral capacities may operate without language, such as the recognition of prototypically prosocial and antisocial behaviors [20] or foundations for moral action in empathy and reciprocity [21]. But language is paramount for human morality. A norm system is conceptually and linguistically demanding, requiring language for learning it, using it, and negotiating it. Further, a morally competent human will need a vocabulary to express moral concepts and instantiate moral practices—to blame or forgive others’ transgressions, to justify and excuse one’s behavior, to contest and negotiate the importance of one norm over another.

A moral vocabulary 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,” but also “intentional,” “knowingly,” etc.);

3. **Vocabulary of responses to violations**
   (“blame,” “reprimand,” “excuse,” “forgiveness,” etc.).

Within each domain, there are numerous distinctions, and some have surprisingly subtle differentiations. In the second domain, for example, my colleagues and I recently uncovered a two-dimensional organization of 28 verbs of moral criticism [22] that suggests people systematically differentiate among verbs to capture criticism of different intensity in either public or private settings (see Figure 1).
4. Moral Cognition and Affect

I have proposed that two of the major domains of moral vocabulary are (a) norm violations and (b) responses to norm violations. What psychological processes are involved in detecting and responding to such violations? These processes are usually treated under the label of moral judgment, but we need to distinguish between at least two kinds of moral judgment [23]. First, people evaluate events (outcomes, behaviors) as bad, good, wrong, or (im)permissible; second, they evaluate agents as morally responsible, deserving blame or praise. The key difference between the two types is the amount of information processing that normally underlies each judgment. Whereas event judgments merely register that a norm has been violated, agent judgments such as blame take into account the agent’s specific causal involvement, intentionality, and mental states.

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

To further form agent judgments, people search for causes of the detected norm-violating event; if the causes involve an agent, they wonder whether the agent acted intentionally; if she acted intentionally, they wonder what reasons she had; and if the event was not intentional, they wonder whether the agent could and should have prevented it [24]. The core elements here are causal and counterfactual reasoning and social cognition, and that is why a number of researchers suggest that moral cognition is no unique “module” or “engine” but derives largely from ordinary cognition [25], [26], but cognition within the context of norms.

Where in all this is affect? The specific roles of affective phenomena in moral judgment are still under debate. 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 exactly what this affective response sets in motion is unclear: Marking that something important occurred [27]? Strengthening motivation to find the cause of the bad event [28]? Biasing the search for evidence that allows the perceiver to blame somebody [29]? And what do we make of the fact that people can make moral judgments without much affect at all [30] or that moral emotions such as anger or resentment require specific cognitive processes [31]? 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. And if affect is not necessary or constitutive of moral judgments, then even an affectless robot can very much be moral.

5. Moral Decision Making and Action

Human moral decision making has received a fair amount of attention in the research literature, with a focus on how humans handle moral dilemmas [32]–[34]. Much of what these studies reveal is how people resolve difficult conflicts within their norm system (e.g., saving multiple lives by sacrificing one life). A popular theoretical view of such situations is that initial affective responses can be overridden by deliberation [35]. But evidence against this override view is increasing [36]–[38]. People’s judgments seem to involve a package of affective and cognitive processes that all deal simultaneously with the conflict set up by the experimenter. Further, how judgments about carefully constructed extreme dilemmas translate into everyday moral decisions is not entirely clear.

The list of psychological factors that influence any socially significant action is long, certainly including momentary affective states to personality dispositions, automatic imitation to group pressure, and heuristics to reasoned action [39], [40]. This overdetermination is no different from nonmoral actions [41]. What makes certain actions moral is the involvement of socially shared norms, not just individual goals. In humans, there is frequent tension between these social norms and the agent’s own goals, and it is this tension that brings into play two additional psychological factors that guide moral action: empathy and self-regulation [42], [43], both of which are designed to favor communal values over selfish interest.
In designing a robot capable of moral decisions and actions, the tension between self-interest and community benefits can probably be avoided from the start. If so, then empathy and self-regulation may be dispensable, because the robot does not have “temptation” to be selfish and ignore other’s needs. However, humans are highly sensitive to other people’s displays of empathy, and a robot that appears to coldly assess moral situations may not be trusted. A robot may be able to build trust, however, when it models how the human perceives a situation and when it communicates (in natural language) its understanding of the human’s perception—that is, the robot would display perspective taking rather than emotional empathy. Of course, these communications and attempts at perspective taking must not be merely 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 show that it values things.

Thus, we are back to the challenge of building a norm system in the robot, including values that make the machine care about certain outcomes and that guide the robot’s decision making and action, especially in the social world. The key concepts of caring or valuing will have to be spelled out in detail—what it means in computational terms for a machine to care about something, and whether this is different from “having a goal” (even a socially important goal).

Finally, according to numerous authors, “intuitive” processes play an important role in human moral decision making [19], [46]. Do robots need them? To the extent that these processes resolve a capacity limitation in humans, and to the extent that robots do not have this limitation, robots would not need moral intuitions or heuristics in their computational architecture. However, will humans be suspicious of robots that do not “intuit” right and wrong but reason logically over it? Once more, the robot’s ability to explain its reasoning in natural language may be a key to human trust, and this issue will have to be a topic of extensive empirical question.

### 6. Moral Communication

The suite of cognitive tools that enable moral judgment and decision making are insufficient to achieve the socially most important function of morality: to regulate other people’s behavior. For that, moral communication is needed. Moral perceivers often express their moral judgments to the alleged offender or to another community member [47], [48]; they sometimes have to provide evidence for their judgment [23]; the alleged offender may contest the charges or explain the action in question [49]; and social estrangement may need to be repaired through conversation or compensation [50], [51].

First a clarification: I am assuming that sufficiently capable robots are likely targets of moral blame—and evidence backs this assumption [52], [53]. People blame an agent to the extent that the agent has the capacity for choice and intentional action—whether the agent is a human or an advanced robot [53]. The capacity to choose also makes robots targets of blame for unintentional violations (e.g., due to negligence or mistakes) [54], because blame for unintentional violations accrues when the agent has the perceived intentional capacity to correct mistakes and prevent negative outcomes in the future. Because blame as a social act informs, corrects, and provides an opportunity to learn [23], [55], robots that can change and learn will be appropriate targets of blame.
Being an appropriate target of moral judgment also licenses the robot to express moral judgments. This should not be especially difficult for a robot whose moral cognition capacity is well developed and who has basic natural language skills. The subtle varieties of delivering moral criticism, however, may be difficult to master (e.g., the difference between scolding, chiding, or denouncing [22]. On the positive side, the anger and outrage that accompanies many human expressions of moral criticism will be absent. This may be particularly important when the robot is partnered with a human—such as with a police officer on patrol or with a teacher in a classroom—and points out (inaudibly to others) a looming violation. Without the kind of affect that would normally make the human partner defensive, the moral criticism may be more effective. In some communities, however, a robot that detects and, presumably, remembers and reports violations to others would itself violate norms of trust and 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 [56]. A robot would not be susceptible to such pressures of loyalty.

Besides expressing moral judgments, moral competence also requires explaining immoral behaviors (typically one’s own, but sometimes others’). This capacity is directly derived from explaining behaviors in general, which is relatively well understood in psychology [57], [58] but scarcely studied in robotics [59]. 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 [60]. 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 [23]. 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 do not question the deterministic order of the universe but 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 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 [61].

Explanations of one’s own intentional actions require more than causal analysis and simulation; they require access to one’s own reasoning en route to action. Some have famously doubted this capacity in humans [62], but these doubts do not apply in the case of reasons for action [63]. A robot, in any case, should have perfect access to its own reasoning via a system of meta-reasoning [64]. Even so, once the robot accesses the trace of its reasoning, it must articulate this reasoning in humanly comprehensible ways (as beliefs and desires), regardless of the formalism in which it performs the reasoning [59]. This amounts to one last form of simulation: modeling what a human would want to know so as to understand (and accept) the robot’s decision in question. In fact, if the robot can simulate in advance a possible human challenge to its planned action and has available an acceptable explanation, then the action has passed a social criterion for moral behavior.
7. Coda: Who is Responsible?

A question many people ask when thinking about robot ethics is this: “When the machine makes mistakes, who is responsible?” Obviously, as long as the “machine” doesn’t cross a boundary of autonomous decision making, we have just another case of user or product liability (not that those are trivial, but at least they are no philosophical puzzles). As long as robots are owned by someone, given tasks by someone, that person is responsible. But let’s go sci-fi. Suppose some robot is raised by a family, acquires the moral competence roughly outlined here, then goes out into the world, tries to find a job, meets new people. Then what?

We need to remember that “holding people responsible” is a tool of social regulation. People blame other humans because they assume or hope that they can change their behavior (there are other motives for blaming, too). Most nonpsychopathic humans are sensitive to criticism and to the threat of social rejection [65], [66] and are therefore inclined to change their future behavior when morally criticized. Blaming robots would make sense only if such a process of behavior change is built into their architecture. They don’t have to fear criticism, they don’t have to be hurt by blame; they just have to be responsive to reasons that a moral critic provides. If they are not responsive they have lost status as social beings and, just like nonresponsive humans are excluded from their community, robots, too, can be excluded. Rather than building fear into robots to motivate them into becoming moral beings, it may be better to equip robots with a desire to be the best they can be: to be the most morally competent being. Humans might learn a thing or two from that.

References


