Learning How to Behave

Moral Competence for Social Robots

Bertram F. Malle and Matthias Scheutz

Contents

Abstract

We describe a theoretical framework and recent research on one key aspect of robot ethics: the development and implementation of a robot's moral competence. As autonomous machines take on increasingly social roles in human communities, these machines need to have some level of moral competence to ensure safety, acceptance, and justified trust. We review the extensive and complex elements of human moral competence and ask how analogous competences could be implemented in a robot. We propose that moral competence consists of five elements, two constituents (moral norms and moral vocabulary) and three activities (moral judgment, moral action, and moral communication). A robot's computational representations of social and moral norms is a prerequisite for all three

B. F. Malle (✉)
Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, Providence, RI, USA
E-Mail: bfmalle@brown.edu

M. Scheutz
Department of Computer Science, Tufts University, Medford, MA, USA
E-Mail: matthias.scheutz@tufts.edu

O. Bendel (Hrsg.), *Handbuch Maschinenethik*, Springer Reference Geisteswissenschaften, https://doi.org/10.1007/978-3-658-17484-2_17-1

moral activities. However, merely programming in advance the vast network of human norms is impossible, so new computational learning algorithms are needed that allow robots to acquire and update the context-specific and graded norms relevant to their domain of deployment. Moral vocabulary is needed primarily for moral communication, which expresses moral judgments of others' violations and explains one's own moral violations – to justify them, apologize, or declare intentions to do better. Current robots have at best rudimentary moral competence, but with improved learning and reasoning they may begin to show the kinds of capacities that humans will expect of future social robots.

Keywords
Norms · Robot ethics · Machine morality · Moral action · Moral judgment · Machine learning · Explanations

1 Introduction

Robot ethics is concerned with two classes of questions: those that probe the ethical dimensions of humans designing, deploying, and treating robots, and those that probe what ethical and moral capacities a robot should have and how these capacities could be implemented. The first class of questions is concerned with ethical design in engineering (Flanagan et al. 2008; Wynsberghe 2013), values of implementation (Hofmann 2013), and considerations of robot rights (Gunkel 2014; Petersen 2007). The second set of questions, more often labeled "machine morality" (Sullins 2011) or "machine ethics" (Moor 2006) is concerned with criteria for moral agency (Floridi and Sanders 2004), justification for lethal military robots (Arkin 2009), and mathematical proofs for moral reasoning (Bringsjord et al. 2006). We consider these questions distinct but interacting (Malle 2016): ethical design of safe robots must include design *of* moral capacities in robots (Malle and Scheutz 2014), treatment of robots must take into account the robot's own social and moral capacities, and advancing a robot's moral capacities will make reference to a number of moral concepts and phenomena (e.g., norms, values, moral judgement; Anderson and Anderson 2011; Malle et al. 2017; Wallach and Allen 2008). Our focus here will be on the second question: what might constitute a robot's moral competence and how such competence could be implemented in computational architectures.

There are concrete concerns about society's readiness for the advent of new types of robots, especially increasingly autonomous learning machines that become members of human communities. A social robot is one that interacts, collaborates with, looks after, or helps humans. These responsibilities pose serious challenges to robot design and deployment, especially when the human in the interaction is vulnerable and trusts the robot. How should we develop robots that practice reading with a 3rd-grader, look after a sleeping infant, or make a medication adjustment for a patient in unbearable pain? When people take on these roles, they must have the social-cognitive and moral capacities that keep others safe, valued, and respected;

when robots take on these roles, they must too. Thus we ask: what sort of moral capacities are required of social robots?

We are not concerned here with the philosophical debates over whether a robot can be a "genuine" moral agent (Floridi and Sanders 2004), can be "truly" responsible for its actions (Matthias 2004; Sparrow 2007), or can make "real" ethical decisions (Miller et al. 2017; Purves et al. 2015). Our aim is more descriptive: what matters most for the project of designing safe and competent social robots is whether *people treat* those robots as targets of their moral judgments and decision, and whether people *expect* a robot to be moral. In spite of claims that robots cannot be blamed or punished (e.g., Funk et al. 2016; Levy 2015; Sparrow 2007), there is mounting evidence that people do treat robots as targets of moral judgments (Bartneck et al. 2007; Darling et al. 2015; Kahn et al. 2012; Malle et al. 2015). We have therefore defined a robot as moral "if it has one or more relevant competences that people consider important for living in a moral community" (Scheutz and Malle 2017, p. 365). Consequently we first need to sketch what moral competences humans have and whether similar ones could be implemented in robots.

One thing is clear: moral competence is not a single capacity. Many psychological phenomena have been studied that could be called "moral competence": decision making about moral dilemmas (Mikhail 2007; Greene et al. 2001); self-regulation of emotion and prosocial behavior (Eisenberg 2000); moral judgments and associated emotions (Haidt 2001; Alicke 2000); as well as responding to others' moral criticism by means of explanation, justification, or defense (Antaki 1994; Dersley and Wootton 2000; Semin and Manstead 1983). To integrate these and related elements we rely on a framework we have developed over the past years (Malle 2016; Malle and Scheutz 2014; Scheutz 2014; Scheutz and Malle 2014) that lays out the moral capacities that ordinary people exhibit and expect of one another in their social relationships. We assume that people are bound to expect some or all of these capacities in advanced social robots as well and that the moral abilities of machines will emerge from, and be in part constrained by, the relations that people are willing to form with them (Coeckelbergh 2010). An ethically responsible science of social robotics must, therefore, be knowledgeable about these human capacities, 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 (Fridin 2014).

2 A Framework of Moral Competence

We propose that moral competence consists of at least five elements that are closely connected in humans but may come apart in artificial agents (Fig. 1). Three of the elements are activities: moral action, moral judgment, and moral communication; two elements are core constituents of the three activities: moral norms and a moral vocabulary.

Fig. 1 Five elements of moral competence, which can be divided into constituents (lower two) and activities (upper three)

1. *A system of norms* encompasses a community's standards for behavior. They guide an agent's decisions to behave in certain ways (→moral action) and shape others' judgments of those behaviors (→moral judgment).
2. A *moral vocabulary* allows the agent to conceptually and linguistically represent both norms and morally significant behaviors and their appropriate judgments, as well as fuel communication in response to them (→moral communication).
3. *Moral action* is action in compliance with norms and thus is adapted to and coordinated with other community members who operate under the same norms.
4. *Moral judgment* is evaluation of behavior relative to norms and information processing that leads to the specific judgment (e.g., permissibility, wrongness, degrees of blame).
5. *Moral communication* expresses, often supported by affect and emotion, people's moral judgments and their attempts to identify, explain, or defend norm violations, as well as negotiate or repair social estrangement after a norm violation.

3 A System of Norms

Morality's function is to regulate individual behavior so that it complies with community interests. Some individual goals clash with community interests, and when there is no biological mechanism that inhibits pursuit of those goals, social-moral regulation must step in (Churchland 2012; Joyce 2006; Ullmann-Margalit 1977). Humans achieve this regulation by motivating and deterring certain behaviors through the imposition of norms and, if these norms are violated, by levying sanctions (Alexander 1987; Bicchieri 2006). Being equipped with a norm system thus constitutes a necessary element in human moral competence (Sripada and Stich 2006; Nichols and Mallon 2006).

We can conceptualize norms cognitively as instructions to act whereby the agent keenly takes into account that (1) a sufficient number of individuals in the community expect and demand of each other to follow the instruction, and (2) a sufficient

number of individuals in the community in fact follow the instruction (Malle et al. 2017; *cf.* Bicchieri 2006; Brennan et al. 2013). Norms are distinct from other guides of behavior, such as goals or habits. Goals might be pursued even if nobody demands the agent to do so, whereas norms are typically followed *even though* the individual has a goal to do otherwise (e.g., standing in line at the coffee shop even though one would rather put in one's order right away). Collective habits are behaviors that many people perform because they all want to, not because they demand of each other to do so (e.g., eating more food when other people are around; Wenk 2015).

Many fascinating topics arise with respect to norms:

- How do individuals represent norms? (are they concepts? action programs?)
- How are norms organized? (hierarchically? as spreading activation networks?)
- How are norms activated by specific contexts (and how do people identify the context they are in?)
- How do people acquire norms (by observation? instruction? reinforcement?)

From among these fascinating topics we select two that are critically important for designing robots with norm capacity: norm representation and norm acquisition.

3.1 Norm Representation

Currently there is little research available on how norms are represented in the human mind. What has been suggested, from reflection and limited research, is that norms are highly context-specific, activated very quickly, and likely to be organized in some forms of knowledge structures (Aarts and Dijksterhuis 2003; Bicchieri 2006; Harvey and Enzle 1981; Sripada and Stich 2006; Tomasello and Vaish 2013).

There is evidence for context-specific activation: Aarts and Dijksterhuis (2003) showed that the mere sight of a library can trigger the "be quiet" norm, with cognitive as well as behavioral effects. More detailed aspects of the environment can activate norms as well, such as a litter-free courtyard triggering the "don't litter" norm (Cialdini et al. 1990). Conversely, a lot of garbage on the floor indicates that the community does not obey the norm, which reduces the perceived community demand for the individual to follow the norm. We recently developed experimental methods to extract community norms from ordinary people's responses to everyday scenes (Kenett et al. 2016). We presented people with pictures of numerous distinct contexts (e.g., jogging path, board room) and asked them to generate, as quickly as possible, actions that one is "supposed to do here" (to elicit prescription norms) or "forbidden to do here" (to elicit prohibition norms). Context-specificity was very high. Among the top-7 actions mentioned as prescribed in each of the 8 scenes (56 total), only three such prescriptions were mentioned in more than one scene, making 95% of the generated norms specific to a single context. Somewhat less but still impressive, context-specificity of *prohibitions* was 75%.

In a subsequent study we examined the speed of context-specific activation. We selected the top-7 action norms that the previous participants had generated for a

given scene and presented new participants with these seven actions along with a picture of the scene. They were asked to consider one action at a time and quickly press a 'Yes' key if an action was "supposed to" or "should" be performed in this context (for prescriptions) or "forbidden to" or "not allowed to" be performed (for prohibitions). We presented these seven target norms along with seven foils – action norms that had been generated among top-7 in *other* scenes. We showed that people strongly differentiated the norms that were specific to a given context from those that stemmed from another context (signal detection parameter $d' = 1.16$). Moreover, people were surprisingly fast in detecting the context-specific norms, averaging around 1200 ms, which is as fast as, if not faster than, judgments of whether a person's behavior is intentional or reveals the person's goal (Malle and Holbrook 2012).

The specific organizational structure of norms is currently unknown. Some authors suggested that norms are "knowledge structures" (Aarts and Dijksterhuis 2003; Harvey and Enzle 1981), but more needs to be said about what characterizes such structures. Flexible hierarchies would be necessary to handle the context-dependence of norm strength: A given norm may be more important than another norm in one context but the reverse ordering may be true in another context. Moreover, some norms may benefit from temporal organization (e.g., the restaurant wait staff must first guide customers to their table, then bring menus, then ask about food selections – a reverse order would be a notable norm violation).

Aside from these initial ideas, no detailed cognitive model is currently available for the process of norm activation or for the underlying norm representations that would facilitate such (apparently fast) activation. Context specificity, in particular, is a vexing computational problem (Ford and Hayes 1991). Humans seem to recognize contexts by being sensitive to a bundle of diagnostic indicators, among them physical spaces (e.g., office vs. bathroom), temporal markers (morning vs. evening), roles (boss vs. employee), relationships (stranger vs. friend), and goals (e.g., discussion vs. vote tallying in a business meeting). The indicators are likely to covary, so that recognizing certain objects surrounding people allows one to predict the relationships among those people (Wang et al. 2018).

Context also appears to determine at what level of abstraction norms are activated. Suppose a commercial airline pilot decides to no longer fly because of a recently diagnosed heart condition. What norm was activated? The specific norm that "pilots ought not to fly when they know they have a heart condition"? This is likely to be too specific, unless the pilot handbook specifies "heart condition" as one specific requirement for handing in one's resignation. Or that "people ought to protect human life"? This is likely to be too general. Perhaps the most likely scenario is that, upon learning about the heart condition, the pilot draws an inference that a heart condition may pose a safety risk, which activates the norm that "as a pilot one ought not to impose safety risks on one's passengers." Though speculative, the example illustrates that a complete model of norms (and even just a specific network of norms for a pilot) will be hierarchically organized and enormously complex. Nonetheless, somehow people comply with complex norms most of the time, so a computational

system, too, should in principle be able to represent and comply with such a system of norms.

3.2 Norm Acquisition

Children and adults learn norms in a variety of ways. First, most obviously, norms can be established or taught through direct expression (Edwards 1987), be it in verbal utterances or other symbols (e.g., signs). Surprisingly, children seem to be exposed to relatively few such explicit norm expressions (Wright and Bartsch 2008, pp. 74–77), so other paths of acquisition are paramount. Among them, second, is the inference of norms from moral evaluations of specific behaviors, be it from a frown or lashing or a verbal comment (e.g., "this is a mean thing to say"; "that's terrible!"). Third, children and adults infer norms from other people's behavior by imitating others' actions, particularly for novel objects (Casler et al. 2009), when the action is presented as familiar (Schmidt et al. 2011), or when a sufficient number of people perform the same behavior in the same context (Herrmann et al. 2013; Milgram et al. 1969). In addition, if the behavior is performed with high similarity from person to person (McNeill 1995) – perhaps even synchronously, as in a ritual – a norm is very likely to be present (Rossano 2012). Fourth, people take into account the costs and consequences of potentially norm-guided behavior. Behaviors reflect a norm when the agent accepted a cost in performing the behavior (Henrich 2009). Conversely, rare behaviors that have negative consequences for others are suggestive evidence for norms of prohibition (*cf.* Cialdini et al. 1990).

Little of this behavioral work has been translated into cognitive models of human norm learning. By contrast, some initial computational work has tackled the representation, structure, and learning of norms in artificial agents, to which we turn next.

3.3 Norms in Robots

Previous efforts to integrate norms into artificial agents took two main routes: the study of multi-agent systems and the design of formal reasoning systems. In the multi-agent literature, researchers have proposed that purely rational autonomous agents cannot form well-functioning societies, and as models of actual social communities, multi-agent system simulations must take seriously the critical role of norms (Andrighetto et al. 2013; Conte et al. 2013). In this literature, norms are typically treated as social mechanisms, which need to be specified, monitored, and enforced. Some authors have elaborated the internal states of modeled autonomous agents, adding norms as constituents of a mental architecture (e.g., Broersen et al. 2001; Governatori and Rotolo 2007). But even when norms are connected with beliefs, goals, and expectations, the form and properties of mental representation that norms require have remained unclear. Such cognitive details may not be needed for modeling multi-agent *systems* (societies), but they will be needed if we want to build

individual autonomous agents such as robots that interact with people in the real world and learn their norms.

In a second line of work on norms in artificial agents, scholars have offered logical frameworks for rule-based moral reasoning (Ågotnes and Wooldridge 2010; Arkin and Ulam 2009; Iba and Langley 2011), including processes to resolve conflicts (e.g., Pereira and Saptawijaya 2007) and analogical inference to apply previously learned moral judgments to novel scenarios (Blass and Forbus 2015). One prominent proposal for a cognitive architecture of ethical planning in robots combines multiple functions (e.g., "*ethical governor*," "*responsibility advisor*") to detect potential norm violations during action planning (Arkin and Balch 1997; Arkin and Ulam 2009). This system can handle only very specific, hard-coded moral decisions and lacks tools for novel ethical inferences, for reasoning through normative conflicts, or for the acquisition of new norms. Another robotic architecture (Briggs and Scheutz 2015, 2013) detects potential norm violations before action and can offer justifications for why it refuses to follow an unethical user command. Though the system includes general inference algorithms that work with explicit representations of normative principles, it cannot yet acquire new norms or principles from interactions and observations.

Because work on artificial agents has so far overlooked the important process of learning and updating norms we recently set out to develop computational algorithms that can learn the kinds of context-specific norms that humans bring to every social situation (Malle et al. 2017; Sarathy et al. 2017a, b). Building on our theoretical and computational proposals for norm representations and norm inference, we introduced novel learning algorithms that allow agents to learn explicit formal norm representations from observation (e.g., from data on how many people endorse a norm in a particular context). Specifically, we proposed a norm representation scheme that introduces a novel deontic modal operator, imbued with context-specificity and uncertainty, within the Dempster-Shafer theory of evidence (Shafer 1976). The learning algorithm works by integrating previous evidence for a particular norm with new incoming evidence (e.g., the previous hypothesized norm that one has to be "silent" in a library with the new incoming evidence of a person speaking in the library). The Dempster-Shafer (DS) framework is able to systematically incorporate evidence (possibly contradictory) from different information sources (e.g., observations, direct instruction) and update the confidence the learner has in a particular norm. Simulation studies with the DS-based norm representations demonstrate that when an artificial agent observes consistent norm-conforming behavior, the agent will be able to learn those prohibition or obligation norms with high levels of confidence. Conversely, if norm adherence is low, or if the observed behavior is not due to a norm but rather a habit, the learner's confidence in norm representations will not converge over time and thus indicate that the observed behavior may not be due to an underlying norm.

Sarathy et al. (2018) also demonstrated that an artificial agent can learn explicit norm representations for observable properties from explicit natural language instructions. For example, the social norm of safely handing over a knife to a human requires the handler to grasp it by the blade so that the recipient can receive it by the

handle. A robot could thus be taught the norm with the utterance "To hand over a knife, grasp it by the blade." From this utterance, the robot can infer the norm context ("handing over an object") as well as the appropriate action ("grasp knife by the blade" as opposed to some other region on the object, like the handle). This type of explicit instruction is particularly useful for cases where an agent needs to quickly learn the appropriate behavior and where experimenting with alternatives might not be safe. Critically, in our approach, the norms acquired through observation and the norms acquired through instruction share the same representational format. As a result, the artificial agent can systematically reason with and talk about its acquired norms in a unified manner.

4 Moral Vocabulary

Some rudimentary moral capacities may operate without language, such as the recognition of prototypically prosocial and antisocial actions (Hamlin 2013) or foundations for moral action in empathy and reciprocity (Flack and de Waal 2000). But a morally competent human needs a vocabulary to represent a myriad of social and moral norms, to express moral judgments, and to instantiate moral practices such as blaming, justifying, and excusing. A moral vocabulary supporting these functions is not merely a list of words but presupposes an ontology, hierarchical categories within that ontology, and agents' representations of themselves and their role in varying relationships (Parthemore and Whitby 2013). The study of moral language can help reveal this rich structure. Much like conceptual distinctions of time (Casasanto and Boroditsky 2008), space (Gentner et al. 2013), theory of mind (de Villiers 2007), and personality (Saucier and Goldberg 1996), core moral concepts and distinctions are likely to be carved into natural languages and can be revealed in systematic linguistic patterns.

In our initial research we mined a variety of public texts, extant scholarly work, and lay informant reports and found that moral vocabulary falls into three broad domains, each with at least two distinct subcategories (see Fig. 2):

1. A **normative frame** includes language referring to conceptual variants of norms (e.g., "obligation," "value," "principle"), certain near-universal contents of norms (e.g., "harm" "reciprocity"), and properties of norms (e.g., "prohibited," "recommended"), as well as agent qualities that are normatively supported, both as categories (e.g., "virtue, character) and concrete attributes (e.g., "fair," "honest").
2. A language of **norm violation** characterizes attributes of violations (e.g., "wrong," "break") and attributes of violators (e.g., "culpable," "thief").
3. A language of **responses to violations** includes cognitive and behavioral responses from witnesses and victims (e.g., "blame," "forgive") as well as from the norm violator (e.g., "excuse," "shame").

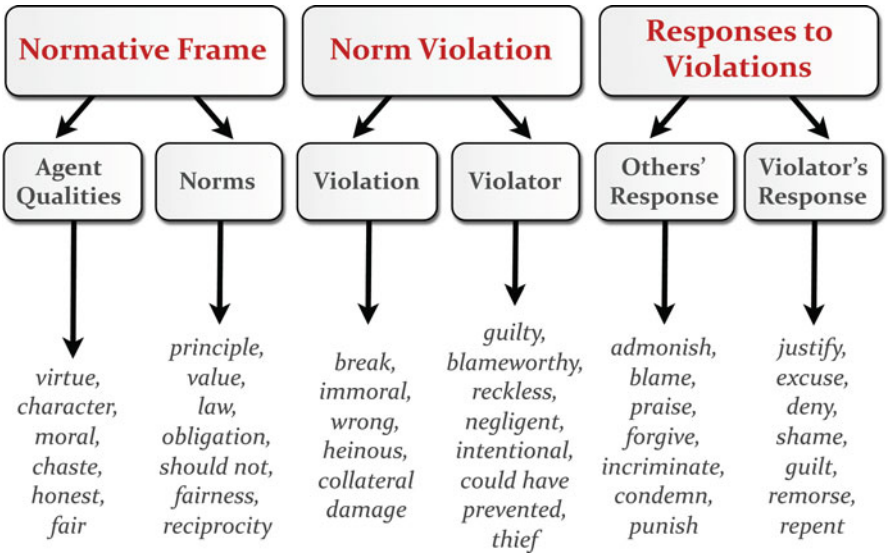

Fig. 2 Three major domains of moral vocabulary, with important subcategories and samples of word instances underneath

We validated this category system by populating it with two pre-existing concise English lexicons and showing that the words can be reliably classified in the system by both expert coders and community members. We then formed a core set of 352 lexemes, selected for being frequently chosen by community members as good representatives of their respective category or subcategory. Using this lexicon we (Voiklis et al. 2016) mined several text corpora to test whether the first- and second-level categories can make meaningful distinctions among certain types of texts, both to each other and relative to a baseline of word frequency from the Corpus of Contemporary American English (COCA, Davies 2010). For example, we analyzed sermons considered "liberal" (Unitarian-Universalist) and ones considered "conservative" (Southern Baptist) and found that both sets of texts referred far more than baseline to virtues and values (category 1) as well as to expressions of moral admonishment, criticism, and praise (category 3). In addition, the Baptist sermons were even more extreme in their emphasis on virtues, admonishment, and criticism, whereas the unitarian sermons tended to emphasize values and principles, permissions, and reconciliatory communication. Overall, a linear discriminant function correctly classified any randomly drawn sermon with an accuracy rate of 82%.

Documenting the discriminative power of this moral vocabulary is only a first step. Equipped with such vocabulary, an artificial intelligence could mine large swaths of verbal data, focus on morally dense passages therein, and learn frequent syntactic constructions, collocations, and substitutability relations. The key challenge is semantic interpretation of this vocabulary, an aim of AI that famously has been met with skepticism (Searle 1980). But language meaning is closely connected with language use (Clark 1985), and if social robots indeed engage in social

interactions, acquire and practice language in discourse contexts (Baldwin and Tomasello 1998), then meaning may no longer be out of reach (Stahl 2004).

5 Moral Judgment

Moral judgment is the set of psychological processes that evaluate behaviors relative to social and moral norms. Among these processes we need to distinguish between at least two kinds (Malle et al. 2014; Monin et al. 2007). First, people evaluate *events* (e.g., a dead boy on the street, a thrown punch) as bad, good, wrong, or (im) permissible. Second, they judge the *agent* who committed the violation as morally responsible, deserving blame or praise. The two kinds of processes differ not just in the object of judgment (event vs. agent) but in the amount of information processing that normally underlies each judgment. Whereas event judgments merely register that a norm has been violated (or met or exceeded), agent judgments such as blame (or praise) take into account the agent's causal contributions and mental states (Malle et al. 2014).

For a robot to register that an event violated a norm is not a trivial endeavor. At a minimum, it must be able to segment the stream of events (e.g., behaviors), identify the context in which the events occur, know which norms apply to this context, and identify those events that violate one or more of the applicable norms. Beyond detecting norm violations, the process of making agent-directed moral judgments such as blame requires much more: assessing an agent's causal contributions; determining whether the agent acted intentionally; assuming she acted intentionally, what reasons she had; assuming the event was not intentional, whether the agent could and should have prevented it (Gilbert et al. 2015; Malle et al. 2014). Because of this complex array of processes underlying moral judgment, a number of researchers have suggested that moral cognition is no unique "module" or "engine" but derives from ordinary cognition (Cushman and Young 2011; Guglielmo et al. 2009). What makes moral judgment unique is not so much a particular process but the fact that events are analyzed and evaluated with respect to *norms*.

Where in all this is affect? The specific roles of affective phenomena in moral judgment are still debated. There is little doubt that the detection of a norm violation often leads to a negative affective response – an evaluation that *something is bad*, perhaps accompanied by physiological arousal and facial expressions. But exactly what this affective response sets in motion is unclear: A strengthened motivation to find the cause of the bad event (Knobe and Fraser 2008)? A biased search for evidence that allows the perceiver to blame somebody (Alicke 2000)? Or merely a marker that something important occurred (Damasio 1994) that now requires further information processing (Guglielmo and Malle 2017)? Nobody would deny that affective phenomena often accompany moral judgments and that they probably facilitate learning moral norms; but there is little direct evidence for the claim that they are *necessary* or *constitutive* of those judgments (May 2018; Pizarro et al. 2011). People can make moral judgments without much affect at all (Harenski et al. 2010; Cima et al. 2010), and moral emotions such as anger or resentment require

specific cognitive processes (Hutcherson and Gross 2011). Even the familiar assumption that emotions or a desire to blame routinely bias moral judgments have recently come into question (Cusimano et al. 2017; Gawronski et al. 2018; Horne and Powell 2016; Monroe and Malle 2018).

If emotions are not necessary or constitutive of moral judgments, then robots can produce moral judgments even if they do not have emotions. As long as artificial agents can approximate human judgments' sensitivity to critical information (i.e., severity of norm violation, causality, intentionality, etc.), the absence of affective responses will be of little significance. In fact, this absence may be welcome because it averts the potential distorting impact that emotions sometimes have on moral judgments. A problem may arise, however, when affect is entirely absent in the *communication* of those moral judgments, and we return to this problem in Sect. 7.

6 Moral Action

Moral action, we suggested earlier, is action in compliance with moral norms. Such action will typically be intentional and grounded in planning and decision making processes that are not themselves moral (Cushman and Young 2011; May 2018). Thus, we focus here on the additional elements that need to be in place for ordinary planning and action systems to execute genuinely moral behaviors. We have already sketched the importance and complexity of norm systems in human and artificial moral agents; we now address specific challenges that actions in compliance with norms face.

Human moral decision making has received a fair amount of attention in the research literature, with two dominant foci: determinants of prosocial or anti-social behavior (Bandura 1999; FeldmanHall et al. 2015; Rand and Nowak 2013) and processes involved in solving moral dilemmas (Kohlberg 1984; Paxton et al. 2012; Mikhail 2007). Much of what the latter studies try to clarify 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 (Greene et al. 2004). But evidence against this override view is increasing (Koop 2013; Royzman et al. 2011; Moretto et al. 2010; Davis et al. 2009). People's responses rather seem to involve a package of affective and cognitive processes that simultaneously deal with *decision conflicts*. Such decision conflicts will arise in many everyday situations when every possible action violates some norm and the decision maker must trade off the unavoidable norm violations against each other – for example, by minimizing an overall violation cost function (Kasenberg and Scheutz 2018) or by searching for an ordering of norms in which averting to violate a more important norm justifies violation of a less important norm (Malle 2018). If an artificial agent tracks these tradeoffs it would be able to justify its conflict resolution to a human observer (Scheutz et al. 2015). Such machines may meet the hope for logical consistency and transparent verifiability (Bringsjord 2015; Dennis et al. 2016) and, as some have argued, might even render robots superior to humans in some domains (Arkin 2009).

Stepping back from the special case of moral dilemmas, logical consistency is not the only requirement for a morally competent community member. In fact, the list of psychological processes underpinning human (im)moral action is long, certainly including deliberate, reasoned action but also personality dispositions, momentary affective states such as empathy or greed, susceptibility to social pressure, and imitation of others' behavior. The latter two factors are often painted as woes of human decision making, as in the cases of obedience and lynch mobs. But without social conformity, human communities could not survive. Consider the modern situation of waiting to board an airplane and then residing for many hours in a tiny space without privacy; it seems remarkable how well behaved these groups of hundreds typically are. Individual deviations are dealt with swiftly by polite moral criticism ("Sir, this line is for first-class passengers only; you need to wait your turn") or removal (even of famous individuals when they, say, pee in a bottle; Todd 2011), and the vast majority follows a host of norms without protest. Even when airlines disrupt the mutual contract of norm following and leave passengers stranded with delayed and cancelled flights, people – though frustrated and angry – still stand in line for hours to get rebooked. Such social-moral action requires imitation, social pressure, and of course fear of sanctions.

Mechanisms such as imitation and conformity have evolved in humans partly to counteract the individual human agents' selfish interests (e.g., acquiring the last stand-by seat, being first on the plane). By contrast, just as robots may not need a host of emotions, they also don't need to be equipped with a host of selfish goals. Robots can be designed to follow social and moral norms that serve the community without having to handle conflicts with goals that serve only themselves. That superintelligent agents "discover" self-serving and self-preserving goals is often assumed in science fiction and speculations about the future (e.g., Bostrom 2014), but it is an assumption that we borrow from our experience with living beings; and it is an assumption that can be eliminated by hard limitations on how we design artificial agents of the future.

Human empathy and care, too, have evolved to counteract selfishness. But a robot's lack of selfishness does not necessitate its lack of empathy or caring. Empathy, at least the human kind, is parochial and strongly reduced by temporal and spatial distance (Bloom 2016), and we should perhaps be reluctant to design robots with this form of empathy. But sensitivity to another's needs and recognition of the costs of human suffering would contribute to a robot being a trustworthy social partner. Thus, such a social robot may have to demonstrate to human observers that it *values* things (Scheutz 2012), that it *cares* about certain outcomes (Wynsberghe 2013). Whether caring entails affect or emotion is currently unclear, but important ingredients of a robot's caring will include willingness to prioritize, attend to, exert effort to help, and so on. There is a potential tension, however, between a robot that has no selfish goals and a robot that cares about things. Part of what it means for a human to care is that one puts one's own needs aside in favor of the other's needs; if the robot does not have any needs of its own, how can its actions be interpreted as demonstrating that it cares? Perhaps it is the result that counts: if the robot succeeds

in rescuing, reviving, or just cheering up a person, it will *feel* as if the robot cares. Is that enough?

It is clear, that we want social robots to do the right thing (whether they can do it for the right reasons or not (Purves et al. 2015). If a robot can make morally preferred decisions, it will still make mistakes, violate norms, and perhaps make the wrong trade-offs in the eyes of some of its human partners. In such cases, most people seem ready to assign blame to a robot – in imagined scenarios (Malle et al. 2015; Monroe et al. 2014) and actual interactions (Kahn et al. 2012). In such cases, people don't rely on a theory of moral agency; rather, blame comes naturally as an act of social regulation that provides the norm violator with an opportunity to do better next time, to not violate the norm again (Cushman 2013; Malle et al. 2014). Thus, human blame could be used to regulate robot behavior if robots were able to take the received blame into account, update their norm system, and make better choices next time. Such feedback is not without risks, because not every moral teacher can be trusted, as we know from cases such as Chappie 2015, or *Tay* (West 2016). Moreover, even learning from established traces of human culture, such as novels or fairy tales (Riedl and Harrison 2016), does not always teach the right moral lessons (Mullan 2017). A robot's learning of a community's norm system must be an iterative process, relying on initial constraints that limit learning of unacceptable lessons and relying on multiple checks and balances (such as the robot consulting a panel of trusted community members when updating its norm system).

7 Moral Communication

Even if a robot is equipped with the cognitive tools that enable moral judgment and moral action, it will still fall short of a critical function of morality: to regulate people's behavior before or after they violate norms and to negotiate the social impact of norm violations. For that, moral communication is needed. Human community members often express their moral judgments to the suspected offender or to other community members (Dersley and Wootton 2000; Feinberg et al. 2012; Weingart et al. 2014); moral decision makers have to explain their actions to others (Antaki 1994); and social estrangement may need to be repaired through conversation or compensation (Walker 2006; McKenna 2012). Social robots need to have rudimentary communicative skills of this sort, to at least express the detection of a person's norm violation and explain their own norm violations to others.

Expressing moral judgments will not be insurmountable for robots that have moral cognition capacity and 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; Voiklis et al. 2014). On the positive side, the anger and outrage that sometimes accompanies human expressions of moral criticism can be avoided. This may be particularly important when the robot is a collaborative partner with a human, such as with a police officer on patrol or a teacher in the classroom. Here the robot may point out its partner's looming violation but preferably remain inaudible to others and take a calm, nonjudgmental tone.

Without the kind of affect that often makes humans defensive when being targets of criticism, the criticism may be more effective.

A problem could arise, however, when the robot coldly utters moral assessments such as, "He deserves a significant amount of blame for beating the prisoner." People normally expect community members not only to notice and point out norm violations but to do so with appropriate displays of concern or outrage (Drew 1998; Fehr and Fischbacher 2004). When a person fails to convey moral criticism with appropriate affective intensity, the audience will be confused, suspicious, perhaps consider such absent expression a norm violation itself. Whether humans have equal expectations for robots to affectively express and communicate moral judgments is currently unknown, so empirical research is needed to provide insight into this question.

Another challenge to a robot communicating its moral judgments is that, in some communities, a robot that reports observed violations might violate trust or loyalty norms. For example, a serious challenge in the military is that soldiers are reluctant to report unit members' unethical behavior, including human rights violations (MHAT-IV 2006); the same pattern is well known as the "Blue Code of Silence" in police units (Westmarland 2005). A robot may not be susceptible to such pressures of loyalty, but if it routinely reports violations it may not find its way into the tight social community of soldiers or police officers, being rejected as a snitch. Robots may have to first earn a level of trust that authorizes them to monitor and enforce norms. Then they would need to explicitly communicate their obligation to report norm violations, using this communication as a reminder of the applicable norms and an admonishment to obey them.

Explaining one's own norm-violating behaviors is a second important moral communication capacity, directly derived from the capacity to explain behaviors in general, which is relatively well understood in psychology (Hilton 2007). A robot that explains its behavior must be intelligible to its human partners, so whatever its process architecture, it must formulate explanations within the conceptual framework of lay behavior explanations (de Graaf and Malle 2017). In particular, people's explanations of intentional behaviors are conceptually distinct from those of unintentional behaviors. People 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, which are seen as generating effects without the involvement of reasoned beliefs, desires and intentions (Malle 1999, 2011). Correspondingly, explaining one's intentional moral violations amounts to offering reasons that justify the violating action, whereas explaining unintentional moral violations amounts to offering causes that excuse one's involvement in the violation (Malle et al. 2014). Unique to the moral domain, such unintentional violations are evaluated by counterfactuals: Blame for unintentional negative events increases if the person should have and could have acted differently to prevent the event. As a result, moral criticism involves simulation of the past (what alternative paths of prevention the agent may have had available) and simulation of the future (how one is expected to act differently to prevent repeated offenses), both feasible computations (Bello 2012).

Explanations of one's own intentional actions of course 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 (Nisbett and Wilson 1977), but these doubts weaken in the case of reasons for intentional action (Malle 2011). A robot, in any case, should have solid access to its own reasoning. Once it retrieves the trace of its reasoning, it must translate this reasoning into humanly comprehensible ways (e.g., reason explanations) (Cox 2011; de Graaf and Malle 2017). This amounts to one last form of simulation: modeling what relevant community members would need to know so as to understand, and deem justified, the robot's decision in question. A robot with this capacity would have a community-validated moral decision criterion at hand: it would simulate in advance possible human challenges to its planned actions and determine whether a community-accepted explanation is available. If not, the action is unacceptable; if so, then the action has passed a social criterion for moral behavior.

8 Conclusion

In light of the extensive and complex elements of human moral competence, designing robots with such competence is an awe-inspiring challenge. The key steps will be to build computational representations of norm systems and incorporate moral concepts and vocabulary into the robotic architecture. Once norms and concept representations are available in the architecture, the next step is to develop algorithms that can computationally capture moral cognition and moral action. The development of these processes will take time, as will the development of a robot's moral communication skills. New computational learning algorithms will need to be developed to help robots acquire the vast network of human norms and the requisite language that enables moral discourse. To this end, we will need to move from robots that are merely programmed to robots that interact with human communities over time, so they can update their programs and adapt to contexts and demands that designers could not anticipate. However, devising learning algorithms that acquire norms only from trusted or representative community members, not from ones that try to direct the system toward their individual goals, is a difficult problem that will require comprehensive theoretical and creative engineering work.

Acknowledgments This project was supported by a grant from the Office of Naval Research (ONR), No. N00014-13-1-0269, and from the Defense Advanced Research Projects Agency (DARPA), SIMPLEX 14-46-FP-097. The opinions expressed here are our own and do not necessarily reflect the views of ONR or DARPA.

References

Aarts, Henk, and Ap Dijksterhuis. 2003. The silence of the library: Environment, situational norm, and social behavior. *Journal of Personality and Social Psychology* 84(1): 18–28. https://doi.org/10.1037/0022-3514.84.1.18.

Ågotnes, Thomas, and Michael Wooldridge. 2010. Optimal social laws. In *Proceedings of the 9th international conference on autonomous agents and multiagent systems (AAMAS 2010)*, ed. van der Hoek, Kaminka, Lesprance, Luck, and Sen, 667–674.

Alexander, Richard D. 1987. *The biology of moral systems*. Hawthorne: Aldine de Gruyter.

Alicke, Mark D. 2000. Culpable control and the psychology of blame. *Psychological Bulletin* 126 (4): 556–574. https://doi.org/10.1037//0033-2909.126.4.556.

Anderson, Michael, and Susan Leigh Anderson, eds. 2011. *Machine ethics*. New York: Cambridge University Press.

Andrighetto, Giulia, Guido Governatori, Pablo Noriega, and Leendert W. N van der Torre, eds. 2013. *Normative multi-agent systems*, Dagstuhl Follow-Ups 4. Saarbrücken/Wadern: Dagstuhl Publishing. http://nbn-resolving.de/urn:nbn:de:0030-drops-39972

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

Arkin, Ronald C. 2009. *Governing lethal behavior in autonomous robots*. Boca Raton: CRC Press.

Arkin, Ronald C., and T. Balch. 1997. AuRA: Principles and practice in review. *Journal of Experimental and Theoretical Artificial Intelligence* 9(2): 175–189.

Arkin, Ronald C., and P. Ulam. 2009. An Ethical adaptor: Behavioral modification derived from moral emotions. In *Computational intelligence in robotics and automation (CIRA), 2009 IEEE international symposium on*, 381–387. Piscataway: IEEE Press.

Baldwin, Dare A., and Michael Tomasello. 1998. Word learning: A window on early pragmatic understanding. In *The proceedings of the twenty-ninth annual child language research forum*, ed. Eve V. Clark, 3–23. Chicago: Center for the Study of Language and Information.

Bandura, Albert. 1999. Moral disengagement in the perpetration of inhumanities. *Personality and Social Psychology Review* 3(3): 193–209. https://doi.org/10.1207/s15327957pspr0303_3.

Bartneck, Christoph, Marcel Verbunt, Omar Mubin, and Abdullah Al Mahmud. 2007. To kill a mockingbird robot. In *Proceedings of the ACM/IEEE international conference on human-robot interaction*, 81–87. New York: ACM Press. https://doi.org/10.1145/1228716.1228728.

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

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

Blass, Joseph A., and Kenneth D. Forbus. 2015. Moral decision-making by analogy: Generalizations versus exemplars. In *Twenty-Ninth AAAI conference on artificial intelligence*, 501–507. AAAI. http://www.aaai.org/ocs/index.php/AAAI/AAAI15/paper/view/9765.

Bloom, Paul. 2016. *Against empathy: The case for rational compassion*. New York: Ecco.

Bostrom, Nick. 2014. *Superintelligence: Paths, dangers, strategies*. Oxford: Oxford University Press.

Brennan, Geoffrey, Lina Eriksson, Robert E. Goodin, and Nicholas Southwood. 2013. *Explaining norms*. New York: Oxford University Press.

Briggs, Gordon, and Matthias Scheutz. 2013. A hybrid architectural approach to understanding and appropriately generating indirect speech acts. In *Proceedings of twenty-seventh AAAI conference on artificial intelligence*, 1213–1219.

Briggs, Gordon, and Matthias Scheutz. 2015. "'Sorry, I can't do that:' Developing mechanisms to appropriately reject directives in human-robot interactions." In *Proceedings of the 2015 AAAI fall symposium on AI and HRI*, 32–36. Palo Alto, CA: AAAI Press.

Bringsjord, Selmer. 2015. A vindication of program verification. *History and Philosophy of Logic* 36(3): 262–277. https://doi.org/10.1080/01445340.2015.1065461.

Bringsjord, Selmer, Konstantine Arkoudas, and Paul Bello. 2006. Toward a general logicist methodology for engineering ethically correct robots. *Intelligent Systems, IEEE* 21(4): 38–44.

Broersen, Jan, Mehdi Dastani, Joris Hulstijn, Zisheng Huang, and Leendert van der Torre. 2001. The BOID architecture: Conflicts between beliefs, obligations, intentions and desires. In *Proceedings of the fifth international conference on autonomous agents*, AGENTS '01, 9–16. New York: ACM. https://doi.org/10.1145/375735.375766.

Casasanto, Daniel, and Lera Boroditsky. 2008. Time in the mind: Using space to think about time. *Cognition* 106(2): 579–593. https://doi.org/10.1016/j.cognition.2007.03.004.

Casler, Krista, Treysi Terziyan, and Kimberly Greene. 2009. Toddlers view artifact function normatively. *Cognitive Development* 24(3): 240–247. https://doi.org/10.1016/j.cogdev.2009.03.005.

Chappie. 2015. Motion picture. Sony Pictures Home Entertainment.

Churchland, Patricia S. 2012. *Braintrust: What neuroscience tells us about morality*. Princeton: Princeton University Press.

Cialdini, Robert B., Raymond R. Reno, and Carl A. Kallgren. 1990. A focus theory of normative conduct: Recycling the concept of norms to reduce littering in public places. *Journal of Personality and Social Psychology* 58(6): 1015–1026. https://doi.org/10.1037/0022-3514.58.6.1015.

Cima, Maaike, Franca Tonnaer, and Marc D. Hauser. 2010. Psychopaths know right from wrong but don't care. *Social Cognitive and Affective Neuroscience* 5(1): 59–67. https://doi.org/10.1093/scan/nsp051.

Clark, Herbert H. 1985. Language use and language users. In *Handbook of social psychology*, ed. Gardner Lindzey and Eliot Aronson, 179–231. New York: Random House.

Coeckelbergh, Mark. 2010. Robot rights? Towards a social-relational justification of moral consideration. *Ethics and Information Technology* 12(3): 209–221. https://doi.org/10.1007/s10676-010-9235-5.

Conte, Rosaria, Giulia Andrighetto, and Marco Campenni. 2013. *Minding norms: Mechanisms and dynamics of social order in agent societies*. New York: Oxford University Press.

Cox, Michael T. 2011. Metareasoning, monitoring, and self-explanation. In *Metareasoning*, ed. Michael T. Cox and Anita Raja, 131–149. Cambridge, MA: The MIT Press. http://mitpress.universitypressscholarship.com/view/10.7551/mitpress/9780262014809.001.0001/upso-9780262014809-chapter-9

Cushman, Fiery. 2013. The functional design of punishment and the psychology of learning. In *Psychological and environmental foundations of cooperation*, Signaling, commitment and emotion, ed. Richard Joyce, Kim Sterelny, B. Calcott, and B. Fraser, vol. 2. Cambridge, MA: MIT Press.

Cushman, Fiery, and Liane Young. 2011. Patterns of moral judgment derive from nonmoral psychological representations. *Cognitive Science* 35(6): 1052–1075. https://doi.org/10.1111/j.1551-6709.2010.01167.x.

Cusimano, Corey, Stuti Thapa Magar, and Bertram F. Malle. 2017. Judgment before emotion: People access moral evaluations faster than affective states. In *Proceedings of the 39th annual conference of the cognitive science society*, ed. G. Gunzelmann, A. Howes, T. Tenbrink, and E. J. Davelaar, 1848–1853. Austin: Cognitive Science Society.

Damasio, Antonio R. 1994. *Descartes' error: Emotion, reason, and the human brain*. New York: Putnam.

Darling, Kate, Palash Nandy, and Cynthia Breazeal. 2015. Empathic concern and the effect of stories in human-robot interaction. In *Proceedings of the 24th IEEE international symposium on robot and human interactive communication (RO-MAN)*, 770–775. Kobe: IEEE. https://doi.org/10.1109/ROMAN.2015.7333675.

Davies, Mark. 2010. The corpus of contemporary American english as the first reliable monitor corpus of english. *Literary and Linguistic Computing* 25(4): 447–464. https://doi.org/10.1093/llc/fqq018.

Davis, Tyler, Bradley C. Love, and W. Todd Maddox. 2009. Anticipatory emotions in decision tasks: Covert markers of value or attentional processes? *Cognition* 112(1): 195–200. https://doi.org/10.1016/j.cognition.2009.04.002.

Dennis, Louise, Michael Fisher, Marija Slavkovik, and Matt Webster. 2016. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems* 77(March): 1–14. https://doi.org/10.1016/j.robot.2015.11.012.

Dersley, Ian, and Anthony Wootton. 2000. Complaint sequences within antagonistic argument. *Research on Language and Social Interaction* 33(4): 375–406. https://doi.org/10.1207/S15327973RLSI3304_02.

Drew, Paul. 1998. Complaints about transgressions and misconduct. *Research on Language & Social Interaction* 31(3/4): 295–325.

Edwards, Carolyn Pope. 1987. Culture and the construction of moral values: A comparative ethnography of moral encounters in two cultural settings. In *The emergence of morality in young children*, ed. Jerome Kagan and Sharon Lamb, 123–151. Chicago: University of Chicago Press.

Eisenberg, Nancy. 2000. Emotion, regulation, and moral development. *Annual Review of Psychology* 51: 665–697.

Fehr, Ernst, and Urs Fischbacher. 2004. Third-Party punishment and social norms. *Evolution and Human Behavior* 25(2): 63–87. https://doi.org/10.1016/S1090-5138(04)00005-4.

Feinberg, Matthew, Joey T. Cheng, and Robb Willer. 2012. Gossip as an effective and low-cost form of punishment. *Behavioral and Brain Sciences* 35(01): 25–25. https://doi.org/10.1017/S0140525X11001233.

FeldmanHall, Oriel, Tim Dalgleish, Davy Evans, and Dean Mobbs. 2015. Empathic concern drives costly altruism. *NeuroImage* 105(January): 347–356. https://doi.org/10.1016/j.neuroimage.2014.10.043.

Flack, J. C., and Frans B. M. de Waal. 2000. Any animal whatever'. Darwinian building blocks of morality in monkeys and apes. *Journal of Consciousness Studies* 7(1–2): 1–29.

Flanagan, Mary, Daniel C. Howe, and Helen Nissenbaum. 2008. Embodying values in technology: Theory and practice. In *Information technology and moral philosophy*, Cambridge studies in philosophy and public policy, ed. Jeroen van den Hoven and John Weckert, 322–353. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511498725.017.

Floridi, Luciano, and J. W. Sanders. 2004. On the morality of artificial agents. *Minds and Machines* 14(3): 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.

Ford, Kenneth M., and Patrick J. Hayes. 1991. *Reasoning agents in a dynamic world: The frame problem*. Greenwich: JAI Press.

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

Funk, Michael, Bernhard Irrgang, and Silvio Leuteritz. 2016. Enhanced information warfare and three moral claims of combat drone responsibility. In *Drones and responsibility: Legal, philosophical and socio-technical perspectives on remotely controlled weapons*, ed. Ezio Di Nucci and Filippo Santoni de Sio, 182–196. London: Routledge.

Gawronski, Bertram, Paul Conway, Joel Armstrong, Rebecca Friesdorf, and Mandy Hütter. 2018. Effects of incidental emotions on moral dilemma judgments: An analysis using the CNI model. *Emotion* (February). https://doi.org/10.1037/emo0000399.

Gentner, Dedre, Asli Ozyürek, Ozge Gürcanli, and Susan Goldin-Meadow. 2013. Spatial language facilitates spatial cognition: Evidence from children who lack language input. *Cognition* 127(3): 318–330. https://doi.org/10.1016/j.cognition.2013.01.003.

Gilbert, Elizabeth A., Elizabeth R. Tenney, Christopher R. Holland, and Barbara A. Spellman. 2015. Counterfactuals, control, and causation: Why knowledgeable people get blamed more. *Personality and Social Psychology Bulletin*, March, 0146167215572137. https://doi.org/10.1177/0146167215572137.

Governatori, Guido, and Antonino Rotolo. 2007. BIO logical agents: Norms, beliefs, intentions in defeasible logic. In *Normative Multi-Agent Systems*, ed. Guido Boella, Leon van der Torre, and Harko Verhagen. Dagstuhl seminar proceedings. Dagstuhl, Germany: Internationales Begegnungs- und Forschungszentrum für Informatik (IBFI), Schloss Dagstuhl, Germany. http://drops.dagstuhl.de/opus/volltexte/2007/912.

de Graaf, Maartje, and Bertram F. Malle. 2017. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall symposium series technical reports*, FS-17-01, 19–26. Palo Alto: AAAI Press.

Greene, Joshua D., R. B. Sommerville, Leigh E. Nystrom, John M. Darley, and Jonathan D. Cohen. 2001. An FMRI investigation of emotional engagement in moral judgment. *Science* 293(5537): 2105–2108. https://doi.org/10.1126/science.1062872.

Greene, Joshua D., Leigh E. Nystrom, Andrew D. Engell, John M. Darley, and Jonathan D. Cohen. 2004. The neural bases of cognitive conflict and control in moral judgment. *Neuron* 44(2): 389–400. https://doi.org/10.1016/j.neuron.2004.09.027.

Guglielmo, Steve, and Bertram F. Malle. 2017. Information-acquisition processes in moral judgments of blame. *Personality and Social Psychology Bulletin* 43(7): 957–971. https://doi.org/10.1177/0146167217702375.

Guglielmo, Steve, Andrew E. Monroe, and Bertram F. Malle. 2009. At the heart of morality lies folk psychology. *Inquiry: An Interdisciplinary Journal of Philosophy* 52(5): 449–466. https://doi.org/10.1080/00201740903302600.

Gunkel, David J. 2014. A vindication of the rights of machines. *Philosophy & Technology* 27(1): 113–132. https://doi.org/10.1007/s13347-013-0121-z.

Haidt, Jonathan. 2001. The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review* 108(4): 814–834. https://doi.org/10.1037/0033-295X.108.4.814.

Hamlin, J. Kiley. 2013. Moral judgment and action in preverbal infants and toddlers: Evidence for an innate moral core. *Current Directions in Psychological Science* 22(3): 186–193. https://doi.org/10.1177/0963721412470687.

Harenski, Carla L., Keith A. Harenski, Matthew S. Shane, and Kent A. Kiehl. 2010. Aberrant neural processing of moral violations in criminal psychopaths. *Journal of Abnormal Psychology* 119(4): 863–874.

Harvey, Michael D., and Michael E. Enzle. 1981. A cognitive model of social norms for understanding the transgression–helping effect. *Journal of Personality and Social Psychology* 41(5): 866–875. https://doi.org/10.1037/0022-3514.41.5.866.

Henrich, Joseph. 2009. The evolution of costly displays, cooperation and religion: Credibility enhancing displays and their implications for cultural evolution. *Evolution and Human Behavior* 30(4): 244–260. https://doi.org/10.1016/j.evolhumbehav.2009.03.005.

Herrmann, Patricia A., Cristine H. Legare, Paul L. Harris, and Harvey Whitehouse. 2013. Stick to the script: The effect of witnessing multiple actors on children's imitation. *Cognition* 129(3): 536–543. https://doi.org/10.1016/j.cognition.2013.08.010.

Hilton, Denis J. 2007. Causal explanation: From social perception to knowledge-based causal attribution. In *Social psychology: Handbook of basic principles*, 2nd ed., ed. Arie W. Kruglanski and E. Tory Higgins, 232–253. New York: Guilford Press.

Hofmann, Bjørn. 2013. Ethical challenges with welfare technology: A review of the literature. *Science and Engineering Ethics* 19(2): 389–406.

Horne, Zachary, and Derek Powell. 2016. How large is the role of emotion in judgments of moral dilemmas? *PLOS ONE* 11(7): e0154780. https://doi.org/10.1371/journal.pone.0154780.

Hutcherson, Cendri A., and James J. Gross. 2011. The moral emotions: A social–functionalist account of anger, disgust, and contempt. *Journal of Personality and Social Psychology* 100(4): 719–737. https://doi.org/10.1037/a0022408.

Iba, W. F., and P. Langley. 2011. Exploring moral reasoning in a cognitive architecture. In *Proceedings of the 33rd annual meeting of the cognitive science society.* Austin, TX: Cognitive Science Society

Joyce, Richard. 2006. *The evolution of morality*. Cambridge, MA: MIT Press.

Kahn, Jr., Peter H., Takayuki Kanda, Hiroshi Ishiguro, Brian T. Gill, Jolina H. Ruckert, Solace Shen, Heather E. Gary, Aimee L. Reichert, Nathan G. Freier, and Rachel L. Severson. 2012. 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*, 33–40. New York: ACM. https://doi.org/10.1145/2157689.2157696.

Kasenberg, Daniel, and Matthias Scheutz. 2018. Norm conflict resolution in stochastic domains. In *Proceedings of the thirty-second AAAI conference on artificial intelligence*, 32–36. Palo Alto, CA: AAAI Press.

Kenett, Yoed N., M. M. Allaham, Joseph L. Austerweil, and Bertram F. Malle. 2016. The norm fluency task: Unveiling the properties of norm representation. In *Poster presented at the 57th annual meeting of the psychonomic society, Boston, MA, November 2016*. Boston.

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

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

Koop, Gregory J. 2013. An assessment of the temporal dynamics of moral decisions. *Judgment and Decision making* 8(5): 527–539.

Levy, David. 2015. When robots do wrong. In *Cognitive robotics*, ed. Hooman Samani, 3–22. Boca Raton: CRC Press. https://doi.org/10.1201/b19171-4.

Malle, Bertram F. 1999. How people explain behavior: A new theoretical framework. *Personality and Social Psychology Review* 3(1): 23–48. https://doi.org/10.1207/s15327957pspr0301_2.

Malle, Bertram F. 2011. Time to give up the dogmas of attribution: A new theory of behavior explanation. In *Advances of experimental social psychology*, ed. Mark P. Zanna and James M. Olson, vol. 44, 297–352. San Diego: Academic Press.

Malle, Bertram F. 2016. Integrating robot ethics and machine morality: The study and design of moral competence in robots. *Ethics and Information Technology* 18(4): 243–256. https://doi.org/10.1007/s10676-015-9367-8.

Malle, Bertram F. 2018. From binary deontics to deontic continua: The nature of human (and Robot) norm systems. *Paper presented at the third international robo-philosophy conference, University of Vienna, Austria*.

Malle, Bertram F., and Jess Holbrook. 2012. Is there a hierarchy of social inferences? The likelihood and speed of inferring intentionality, mind, and personality. *Journal of Personality and Social Psychology* 102(4): 661–684. https://doi.org/10.1037/a0026790.

Malle, Bertram F., and Matthias Scheutz. 2014. Moral competence in social robots. In *Proceedings of IEEE international symposium on ethics in engineering, science, and technology, Ethics'2014*, 30–35. Chicago: IEEE.

Malle, Bertram F., Steve Guglielmo, and Andrew E. Monroe. 2014. A theory of blame. *Psychological Inquiry* 25(2): 147–186. https://doi.org/10.1080/1047840X.2014.877340.

Malle, Bertram F., Matthias Scheutz, Thomas Arnold, John Voiklis, and Corey Cusimano. 2015. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on Human-Robot Interaction, HRI'15*, 117–124. New York: ACM.

Malle, Bertram F., Matthias Scheutz, and Joseph L. Austerweil. 2017. Networks of social and moral norms in human and robot agents. In *A world with robots: International Conference on Robot Ethics: ICRE 2015*, ed. Maria Isabel Aldinhas Ferreira, Joao Silva Sequeira, Mohammad Osman Tokhi, Endre E. Kadar, and Gurvinder Singh Virk, 3–17. Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-46667-5_1.

Matthias, Andreas. 2004. The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology* 6(3): 175–183. https://doi.org/10.1007/s10676-004-3422-1.

May, Joshua. 2018. Précis of regard for reason in the moral mind. *Behavioral and Brain Sciences*.

McKenna, Michael. 2012. Directed blame and conversation. In *Blame: Its nature and norms*, ed. D. Justin Coates and Neal A. Tognazzini, 119–140. New York: Oxford University Press.

McNeill, William Hardy. 1995. *Keeping together in time: Dance and drill in human history*. Cambridge, MA: Harvard University Press http://hdl.handle.net/2027/heb.04002.

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

Mikhail, John. 2007. Universal moral grammar: Theory, evidence and the future. *Trends in Cognitive Sciences* 11(4): 143–152. https://doi.org/10.1016/j.tics.2006.12.007.

Milgram, Stanley, Leonard Bickman, and Lawrence Berkowitz. 1969. Note on the drawing power of crowds of different size. *Journal of Personality and Social Psychology* 13(2): 79–82. https://doi.org/10.1037/h0028070.

Miller, Keith W., Marty J. Wolf, and Frances Grodzinsky. 2017. This 'Ethical Trap' is for roboticists, not robots: On the issue of artificial agent ethical decision-making. *Science and Engineering Ethics* 23(2): 389–401. https://doi.org/10.1007/s11948-016-9785-y.

Monin, Benoît, David A. Pizarro, and Jennifer S. Beer. 2007. Deciding versus reacting: Conceptions of moral judgment and the reason-affect debate. *Review of General Psychology* 11(2): 99–111. https://doi.org/10.1037/1089-2680.11.2.99.

Monroe, Andrew E., and Bertram F. Malle. 2018. People systematically update moral judgments of blame. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspa0000137.

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

Moor, James H. 2006. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems* 21(4): 18–21. https://doi.org/10.1109/MIS.2006.80.

Moretto, Giovanna, Elisabetta Làdavas, Flavia Mattioli, and Giuseppe di Pellegrino. 2010. A psychophysiological investigation of moral judgment after ventromedial prefrontal damage. *Journal of Cognitive Neuroscience* 22(8): 1888–1899. https://doi.org/10.1162/jocn.2009.21367.

Mullan, John. 2017. We need robots to have morals. Could shakespeare and austen help? *The Guardian*, July 24, 2017, sec. Opinion. https://www.theguardian.com/commentisfree/2017/jul/24/robots-ethics-shakespeare-austen-literature-classics

Nichols, Shaun, and Ron Mallon. 2006. Moral dilemmas and moral rules. *Cognition* 100(3): 530–542. https://doi.org/10.1016/j.cognition.2005.07.005.

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

Parthemore, Joel, and Blay Whitby. 2013. What makes any agent a moral agent? Reflections on machine consciousness and moral agency. *International Journal of Machine Consciousness* 4: 105–129.

Paxton, Joseph M., Leo Ungar, and Joshua D. Greene. 2012. Reflection and reasoning in moral judgment. *Cognitive Science* 36(1): 163–177. https://doi.org/10.1111/j.1551-6709.2011.01210.x.

Pereira, Luís Moniz, and Ari Saptawijaya. 2007. Modelling morality with prospective logic. In *Progress in artificial intelligence*, Lecture Notes in Computer Science, ed. José Neves, Manuel Filipe Santos, and José Manuel Machado, 99–111. Springer Berlin/Heidelberg. http://link.springer.com/chapter/10.1007/978-3-540-77002-2_9

Petersen, Stephen. 2007. The ethics of robot servitude. *Journal of Experimental & Theoretical Artificial Intelligence* 19(1): 43–54. https://doi.org/10.1080/09528130601116139.

Pizarro, David A., Yoel Inbar, and Chelsea Helion. 2011. On disgust and moral judgment. *Emotion Review* 3(3): 267–268. https://doi.org/10.1177/1754073911402394.

Purves, Duncan, Ryan Jenkins, and Bradley J. Strawser. 2015. Autonomous machines, moral judgment, and acting for the right reasons. *Ethical Theory and Moral Practice* 18(4): 851–872. https://doi.org/10.1007/s10677-015-9563-y.

Rand, David G., and Martin A. Nowak. 2013. Human cooperation. *Trends in Cognitive Sciences* 17(8): 413–425. https://doi.org/10.1016/j.tics.2013.06.003.

Riedl, Mark O., and Brent Harrison. 2016. Using stories to teach human values to artificial agents. In *AI, ethics, and society, papers from the 2016 AAAI Workshop, Phoenix, Arizona, USA, February 13, 2016*. http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12624

Rossano, Matt J. 2012. The essential role of ritual in the transmission and reinforcement of social norms. *Psychological Bulletin* 138(3): 529–549. https://doi.org/10.1037/a0027038.

Royzman, Edward B., Geoffrey P. Goodwin, and Robert F. Leeman. 2011. When sentimental rules collide: 'Norms with feelings' in the dilemmatic context. *Cognition* 121(1): 101–114. https://doi.org/10.1016/j.cognition.2011.06.006.

Sarathy, Vasanth, Matthias Scheutz, Yoed N. Kenett, M. M. Allaham, Joseph L. Austerweil, and Bertram F. Malle. 2017a. Mental representations and computational modeling of context-specific human norm systems. In *Proceedings of the 39th annual meeting of the Cognitive Science Society, London*, 1035–1040. Austin: Cognitive Science Society.

Sarathy, Vasanth, Matthias Scheutz, and Bertram F. Malle. 2017b. Learning behavioral norms in uncertain and changing contexts. In *Proceedings of the 2017 8th IEEE international conference on cognitive infocommunications (CogInfoCom)*, 301–306. Pascataway, NJ: IEEE Press.

Sarathy, Vasanth, Bradley Oosterveld, Evan Krause, and Matthias Scheutz. 2018. Learning cognitive affordances from natural language instructions. *Advances in Cognitive Systems* 6: 1–20.

Saucier, Gerard, and Lewis R. Goldberg. 1996. The language of personality: Lexical perspectives on the five-factor model. In *The five-factor model of personality: Theoretical perspectives*, ed. Jerry S. Wiggins, 21–50. New York: Guilford Press.

Scheutz, Matthias. 2012. The affect dilemma for artificial agents: Should we develop affective artificial agents? *IEEE Transactions on Affective Computing* 3(4): 424–433.

Scheutz, Matthias. 2014. The need for moral competency in autonomous agent architectures In *Fundamental issues of artificial intelligence*, ed. Vincent C. Müller, 515–525 Berlin: Springer.

Scheutz, Matthias, and Bertram F. Malle. 2014. 'Think and do the right thing': A plea for morally competent autonomous robots. In *Proceedings of the IEEE international symposium on ethics in engineering, science, and technology, Ethics'2014*, 36–39. Red Hook: Curran Associates/IEEE Computer Society.

Scheutz, Matthias, and Bertram F. Malle. 2017. Moral Robots. In *The routledge handbook of neuroethics*, ed. L. Syd M. Johnson and Karen Rommelfanger, 363–377. New York: Routledge.

Scheutz, Matthias, Bertram F. Malle, and Gordon Briggs. 2015. Towards morally sensitive action selection for autonomous social robots In *Proceedings of the 24th IEEE international symposium on robot and human interactive communication (RO-MAN),* 492–497. Kobe: IEEE Press.

Schmidt, Marco F. H., Hannes Rakoczy, and Michael Tomasello. 2011. Young children attribute normativity to novel actions without pedagogy or normative language. *Developmental Science* 14(3): 530–539.

Searle, John R. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences* 3(3): 417–457.

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

Shafer, Glenn. 1976. *A mathematical theory of evidence*. Princeton: Princeton University Press.

Sparrow, Robert. 2007. Killer robots. *Journal of Applied Philosophy* 24(1): 62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x.

Sripada, Chandra Sekhar, and Stephen Stich. 2006. A framework for the psychology of norms. In *The innate mind (Vol. 2: Culture and cognition)*, ed. Peter Carruthers, Stephen Laurence, and Stephen Stich, 280–301. New York: Oxford University Press.

Stahl, Bernd Carsten. 2004. Information, ethics, and computers: The problem of autonomous moral agents. *Minds and Machines* 14(1): 67–83. https://doi.org/10.1023/B:MIND.0000005136.61217.93.

Sullins, John P. 2011. Introduction: Open questions in roboethics. *Philosophy & Technology* 24(3): 233. https://doi.org/10.1007/s13347-011-0043-6.

Todd, Ben. 2011. 'Non, Non! you can't wee wee here, monsieur': Gerard depardieu gets caught short in a plane. DailyMail.Com. August 18, 2011. http://www.dailymail.co.uk/tvshowbiz/article-2026990/Gerard-Depardieu-urinates-plane-hes-refused-permission-to-toilet.html.

Tomasello, Michael, and Amrisha Vaish. 2013. Origins of human cooperation and morality. *Annual Review of Psychology* 64(1): 231–255. https://doi.org/10.1146/annurev-psych-113011-143812.

Ullmann-Margalit, Edna. 1977. *The emergence of norms*, Clarendon Library of Logic and Philosophy. Oxford: Clarendon Press.

de Villiers, Jill. 2007. The interface of language and theory of mind. *Lingua. International Review of General Linguistics. Revue Internationale De Linguistique Generale* 117(11): 1858–1878. https://doi.org/10.1016/j.lingua.2006.11.006.

Voiklis, John, Corey Cusimano, and Bertram F. Malle. 2014. A social-conceptual map of moral criticism. In *Proceedings of the 36th annual conference of the cognitive science society*, ed. Paul Bello, M. Guarini, M. McShane, and Brian Scassellati, 1700–1705. Austin: Cognitive Science Society.

Voiklis, John, Corey Cusimano, and Bertram F. Malle. 2016. *Using moral communication to reveal moral cognition*. Paper presented at the international conference on thinking, providence, RI.

Walker, Margaret Urban. 2006. *Moral repair: Reconstructing moral relations after wrongdoing*. New York: Cambridge University Press.

Wallach, Wendell, and Colin Allen. 2008. *Moral machines: Teaching robots right from wrong*. New York: Oxford University Press.

Wang, Zhouxia, Tianshui Chen, Jimmy Ren, Weihao Yu, Hui Cheng, and Liang Lin. 2018. Deep reasoning with knowledge graph for social relationship understanding. *ArXiv:1807.00504 [Cs]*, July. http://arxiv.org/abs/1807.00504.

Weingart, Laurie R., Kristin J. Behfar, Corinne Bendersky, Gergana Todorova, and Karen A. Jehn. 2014. The directness and oppositional intensity of conflict expression. *Academy of Management Review* 40(2): 235–262. https://doi.org/10.5465/amr.2013.0124.

Wenk, Gary Lee. 2015. *Your brain on food: How chemicals control your thoughts and feelings*. New York: Oxford University Press.

West, John. 2016. Microsoft's disastrous tay experiment shows the hidden dangers of AI. *Quartz*, April 2, 2016. https://qz.com/653084/microsofts-disastrous-tay-experiment-shows-the-hidden-dangers-of-ai/.

Westmarland, Louise. 2005. Police ethics and integrity: Breaking the blue code of silence. *Policing & Society* 15(2): 145–165.

Wright, Jennifer Cole, and Karen Bartsch. 2008. Portraits of early moral sensibility in two children's everyday conversations. *Merrill-Palmer Quarterly* 54(1): 56–85. https://doi.org/10.2307/23096079.

van Wynsberghe, Aimee. 2013. Designing robots for care: Care centered value-sensitive design. *Science and Engineering Ethics* 19(2): 407–433. https://doi.org/10.1007/s11948-011-9343-6.