For *Lethal Autonomous Weapons: Re-Examining the Law & Ethics of Robotic Warfare*

May Machines Take Lives to Save Lives?
Human Perceptions of Autonomous Robots (with the Capacity to Kill)

Matthias Scheutz
Department of Computer Science
200 Boston Avenue
Tufts University
Medford, MA 02155, USA

Bertram F. Malle
Department of Cognitive, Linguistic, and Psychological Sciences
190 Thayer Streeet,
Brown University
Providence, RI 02912, USA

Keywords

moral psychology, moral dilemmas, human-robot interaction, blame, moral justifications

Abstract

In the future, artificial agents are likely to make life-and-death decisions about humans. Ordinary people are the likely arbiters of whether these decisions are morally acceptable. We summarize research on how ordinary people evaluate artificial (compared to human) agents that make life-and-death decisions. The results suggest that many people are inclined to morally evaluate artificial agents' decisions, and when asked how the artificial and human agents should decide, they impose the same norms on them. However, when confronted with how the agents did in fact decide, people judge the artificial agents' decisions differently from those of humans. This difference is best explained by the justifications people grant the human agents (imagining their experience of the decision situation) but not the artificial agent (whose experience they cannot imagine). If people fail to infer the decision processes and justifications of artificial agents, these agents will have to communicate such justifications to people, so they can understand and accept their decisions.

100 word bio-sketches

Matthias Scheutz is Professor of Computer and Cognitive Science in the Department of Computer Science at Tufts University and Senior Gordon Faculty Fellow in Tuft's School of Engineering. He earned a Ph.D. in Philosophy from the University of Vienna in 1995 and a Joint Ph.D. in Cognitive Science and Computer Science from Indiana University Bloomington in 1999. He has over 300 peer-reviewed publications in artificial intelligence, artificial life, agent-based computing, natural language processing, cognitive modeling, robotics, human-robot interaction, and foundations of cognitive science. His research interests include multi-scale agent-based models of social behavior and complex cognitive and affective autonomous robots with natural language and ethical reasoning capabilities for natural human-robot interaction. His lab page is https://hrilab.tufts.edu.

Bertram F. Malle is Professor of Cognitive, Linguistic, and Psychological Sciences and Co-Director of the Humanity-Centered Robotics Initiative at Brown University. Trained in psychology, philosophy, and linguistics at the University of Graz, Austria, he received his Ph.D. in psychology from Stanford University in 1995. He received the Society of Experimental Social Psychology Outstanding Dissertation award in 1995, a National Science Foundation (NSF) CAREER award in 1997, and he is past president of the Society of Philosophy and Psychology. Malle's research, focuses on social cognition, moral psychology, and human-robot interaction. He has distributed his work in 150 scientific publications and several books. His lab page is http://research.clps.brown.edu/SocCogSci.

Introduction

The prospect of developing and deploying autonomous “killer robots”—robots that use lethal force—has occupied news stories now for quite some time, and it is also increasingly being discussed in academic circles, by roboticists, philosophers, and lawyers alike. The arguments made in favor or against using lethal force on autonomous machines range from philosophical first principles (Sparrow, 2007, 2011), to legal considerations (Asaro, 2012; Pagallo, 2011), to practical effectiveness (Bringsjord, 2019) to concerns about computational and engineering feasibility (Arkin, 2009, 2015).

The purposeful application of lethal force, however, is not restricted to military contexts, but can equally arise in civilian settings. In a well-documented case, for example, police used a tele-operated robot to deliver and detonate a bomb to kill a man who had previously shot five police officers (Sidner and Simon, 2016). And while this particular robot was fully tele-operated, it is not unreasonable to imagine that an autonomous robot could be instructed using simple language commands to drive up to the perpetrator and set off the bomb there. The technology exists for all involved capabilities, from understanding the natural language instructions, to autonomously driving through parking lots, to performing specific actions in target locations.

Lethal force, however, does not entail the use of weapons. Rather, a robot can apply its sheer physical mass to inflict significant, perhaps lethal harm on humans, as can a self-driving car when it fails to avoid collisions with other cars or pedestrians. The context of autonomous driving has received particular attention recently, because life-and-death decisions will inevitably have to be made by autonomous cars, and it is highly unclear how they should be made. Much of the discussion here builds on the Trolley Dilemma (Foot, 1967; Thomson, 1976), which used to be restricted to human decision-makers but has been extended to autonomous cars: They, too, can face life-and-death decisions involving their passengers as well as pedestrians on the street, such as when avoiding a collision with four pedestrians is not possible without colliding with a single pedestrian or without endangering the car’s passenger (Awad et al., 2018; Bonnefon et al., 2016; Li et al., 2016; Wolkenstein, 2018; Young and Monroe, 2019).

But autonomous systems can end up making life-and death decisions even without the application of physical force, namely, by sheer omission in favor of an alternative action. A search-and-rescue robot, for example, may attempt to retrieve an immobile injured person from a burning building but in the end choose to leave the person behind and instead guide a group of mobile humans outside, who might otherwise die because the building is about to collapse. Or a robot nurse assistant may refuse to increase a patient’s morphine drip even though the patient is in agony, because the robot is following protocol of not changing pain medication without an attending physician’s direct orders.

In all these cases of an autonomous system making life-and-death decisions, the system’s moral competence will be tested—its capacity to recognize the context it is in, recall the applicable norms, and make decisions that are maximally in line with these norms (Malle and Scheutz, 2019). The ultimate arbiter of whether the system passes this test will be ordinary people. If future artificial agents are to exist in harmony with human communities, their moral competence must be a reflection of the community’s norms, broad human values, and the psychology of moral behavior and moral judgment; only then will people accept those agents as partners in their everyday lives (Malle and Scheutz, 2015; Scheutz and Malle, 2014). In this chapter, we will summarize our recent empirical work on ordinary people’s evaluations of a robot’s moral competence in life-and-death dilemmas of the kinds inspired by the Trolley Dilemma (Malle et al., 2015, 2016; Malle, Scheutz, et al., 2019; Malle, Thapa, et al., 2019).

Specifically, we will compare people's *normative expectations* for how an artificial agent should act in such a dilemma with their expectations for how a human should act in the identical dilemma. In addition, we will assess people's *moral judgments* of artificial (or human) agents after they decided to act one way or another. Critically, we will examine the role of justifications that people consider when evaluating the agents' decisions. We will show that even when norms are highly similar for artificial and human agents, these justifications often differ, and consequently the moral judgments the agents are assigned will differ as well. From these results, it will become clear that artificial agents must be able to justify their decisions when they act in surprising and potentially norm-violating ways (de Graaf and Malle, 2017). For without such justifications, artificial systems will not be understandable, acceptable, and trustworthy to humans (Wachter et al., 2017; Wang et al., 2016). This is another high bar for artificial systems to meet, because these justifications have to navigate a thorny territory of mental states that underlie decisions and of conflicting norms that must be resolved when a decision is made. At the end of this chapter we will briefly sketch what kinds of architectures and algorithms would be required to meet this high bar.

Artificial Moral Agents

Some machines no longer act like machines (e.g., in personnel, military, or search and rescue domains). They make decisions on the basis of beliefs, goals, and other mental states, and their actions have direct impact on social interactions and individual human costs and benefits. Because many of these decisions have moral implications (e.g., harm or benefits to some but not others), people are inclined to treat the machines as moral agents—agents who are expected to act in line with society's norms and, when they do not, are proper targets of blame.

Some scholars do not believe that robots can be blamed or held responsible (e.g., Funk, Irrgang, & Leuteritz, 2016; Sparrow, 2007); but ordinary people are inclined to blame robots (Kahn, Jr. et al., 2012; Malle et al., 2015, 2016; Monroe et al., 2014). Moreover, there is good reason to believe that robots will soon become more sophisticated decision makers, and people will increasingly expect moral decision making from them. Thus we need insights from empirical science to anticipate how people will respond to such agents and explore how these responses should inform agent design . We have conducted several lines of research that examined these responses, and we summarize here two, followed by brief reference to two more.

In all studies we framed the decision problem the agents faces as moral dilemmas—situations in which every available action violates at least one norm. Social robots will inevitably face moral dilemmas (Bonnefon et al., 2016; Lin, 2013; Millar, 2014; Scheutz and Malle, 2014), some involving life-and-death situations, some not. Moral dilemmas are informative because each horn of a dilemma can be considered a norm violation, and it is such violations that prompt perceptions of robot autonomy and moral agency (Briggs and Scheutz, 2017; Harbers et al., 2017; Podschwadek, 2017). But it is not just a matter of perception; artificial agents must weigh the possible violations and resolve the dilemmas in ways that are acceptable to people. However, we do not currently understand whether such resolutions must be identical to those given by humans and, if not, in what features they might differ.

A Robot in a Life-Saving Mining Dilemma

In the first line of work (Malle et al., 2015; Malle, Scheutz, et al., 2019) we examined a variant of the classic trolley dilemma. In our case, a runaway train with four mining workers on board is about to crash into a wall, which would kill all four unless the protagonist (a repairman or repair robot) performs an action that saves the four miners: redirecting the train onto a side track. As a (known but unintended) result of this action, however, a single person working on this side track would die (he cannot be warned). The protagonist must make a decision to either (a) take an action that saves four people but causes a single person to die ("Action") or (ii) take no action and allow the four to die ("Inaction"). In all studies, the experimental conditions of Agent (human or robot) and Decision (action or inaction) were manipulated between subjects. We assessed several kinds of judgments, which fall into two main classes. The first class assesses the *norms* people impose on the agent: "What should the [agent] do?" "Is it permissible for the [agent] to redirect the train?"; the second assesses *evaluations* of the agent's actual decision: "Was it morally wrong that the [agent] decided to [not] direct the train onto the side track?"; "How much blame does the person deserve for [not] redirecting the train onto the side track?" *Norms* were assessed in three/four studies, *decision evaluations* in all studies. In addition, we asked participants to explain why they made the particular moral judgments (e.g., "Why does it seem to you that the [agent] deserves this amount of blame?"). All studies had a 2 (Agent: human repairmen or robot) × 2 (Decision: Action or Inaction) between-subjects design, and we summarize here the results of six studies from around 3000 online participants.

Before we analyzed people's moral responses to robots, we examined whether they treated robots as moral agents in the first place. We systematically classified people's explanations of their moral judgments and identified responses that either expressly denied the robot's moral capacity (e.g., "doesn't have a moral compass," "it's not a person," "it's a machine," "merely programmed,") or mentioned the programmer or designer as the fully or partially responsible agent. Automated text analysis followed by human inspection showed that about one third of U.S. participants denied the robot moral agency, leaving two thirds who accepted the robot as a proper target of blame. Though all results still hold in the entire sample, it made little sense to include data from individuals who explicitly rejected the premise of the study—to evaluate an artificial agent's moral decision. Thus, we further analyzed the data of only those participants who accepted the premise.

First, when probing participants' normative expectations, we found virtually no human-robot differences. Generally, people were equally inclined to find the Action permissible for the human (61%) and the robot (64%), and when asked to choose, they recommended that each agent should take the *Action*, both the human (79%) and the robot (83%).

Second, however, when we analyzed moral judgments we identified a robust human-robot asymmetry across studies (we focus here on blame judgments, but very similar results hold for wrongness judgments). Whereas robots and human agents were blamed equally after deciding to "act" (i.e., sacrifice one person for the good of four)—44.3 and 42.1, respectively, on a 0-100 scale—humans were blamed less ($M = 23.7$) than robots ($M = 40.2$) after deciding to *not* act. Five of the six studies found this pattern to be statistically significant. The average effect size of the relevant interaction term was $d = 0.25$, and the effect size of the human-robot difference in the Inaction condition was $d = 0.50$.

What might explain this asymmetry? It cannot be a preference for a robot to make the "utilitarian" choice and the human to make the deontological choice. Aside from the difficulty of neatly assigning each choice option to these traditions of philosophical ethics, it is actually not the case that people expected the robot to act any differently from humans, as we saw from the highly comparable norm expectation data (permissibility and should). Furthermore, if robots were preferred to be utilitarians, then a robot's *Action* decision would be welcomed and should receive less blame—but in fact, blame for human and robot agents was consistently similar in this condition.

A better explanation for the pattern of less blame for human than robot in the case of Inaction might be that people's *justifications* for the two agents differed. Justifications are the agent's reasons for deciding to act, and those reasons represent the major determinant of blame when causality and intentionality are held constant (Malle et al., 2014), which we can assume is true for the experimental narratives. What considerations might justify the lower blame for the human agent in the Inaction case? We explored people's verbal explanations following their moral judgments and found a pattern of responses that provided a candidate justification: the impossibly difficult decision situation made it understandable and thus somewhat acceptable for the human to decide *not* to act. Indeed, across all studies, people's spontaneous characterizations of the dilemma as "difficult," "impossible," and the like, were more frequent for the Inaction condition (12.1%) than the Action condition (5.8%), and more frequent for the human protagonist (11.2%) than the robot protagonist (6.6%). Thus, it appears that participants notice, or even vicariously feel, this "impossible situation" primarily when the human repairman decides not to act, and that is why the blame levels are lower.

A further test of this interpretation was supportive: When considering those among the 3000 participants who mentioned the decision difficulty, their blame levels were almost 14 points lower, and among this group there was no longer a human-robot asymmetry for the Inaction decision. The candidate explanation for this asymmetry in the whole sample is then that participants more readily consider the decision difficulty for the human agent, especially in the Inaction condition, and when they do, blame levels decrease. Fewer participants consider the decision difficulty for the robot agent, and as a result, little net blame mitigation occurs.

We then turned to an experimental test of this hypothesis. We attempted to highlight the decision difficulty even for robot agents by describing the robot as "struggling with the difficult decision." This manipulation weakened the human-robot asymmetry for Inactions from 0.25 (average of all other studies) to $d = .09$ and to nonsignificance.

In sum, we learned two related lessons from these studies. First, people can have highly similar normative expectations regarding the (prospectively) "right thing to do" for both humans and robots in life and death scenarios, but people's (retrospective) moral judgments of actually made decisions may still differ for human and robot agents. That is because, second, people's justifications of human decisions and robot decisions can differ. In the reported studies, the difference stemmed from the ease of imagining the dilemma's difficulty for the human protagonist, which seemed to somewhat justify the decision to *not* act and lower its associated blame. This kind of imagined difficulty and resulting justification was rarer in the case of a robot protagonist. Observers of these response patterns from ordinary people may be worried about the willingness to decrease blame judgments when one better "understands" a decision (or the difficulty surrounding a decision). But that is not far from the reasonable person standard in contemporary law (e.g., Baron, 2011). The law, too, reduces punishment when the defendant's decision or action is understandable and reasonable. When "anybody" would find it difficult to sacrifice one person for the good of many (even if it were the right thing to do), then nobody should be strongly blamed for refraining from that action. Such a reasonable agent standard is not available for robots, and people's moral judgments reflect this inability to understand, and consider reasonable, a robot's action. This situation can be expected for the foreseeable future, until reasonable robot standards are established or people better understand how the minds of robots work, struggling or not.

AI and Drones in a Military Strike Dilemma

In the second line of work (Malle, Thapa, & Scheutz, 2019), we presented participants with a moral dilemma scenario in a military context inspired by the film *Eye in the Sky* (Hood, 2016).[1] The dilemma is between either (i) launching a missile strike on a terrorist compound but risking the life of a child, or (ii) canceling the strike to protect the child but risking a likely terrorist attack. Participants considered one of three decision makers: an artificial intelligence (AI) agent, an autonomous drone, or a human drone pilot. We embedded the decision maker within a command structure, involving military and legal commanders who provided guidance on the decision.

We asked online participants (a) what the decision maker should do (norm assessment), (b) whether the decision was morally wrong and how much blame the person deserves, and (c) why participants assigned the particular amount of blame. As above, the answers to the third question were content analyzed to identify participants who did not consider the artificial agents proper targets of blame. Across three studies, seventy-two percent of respondents were comfortable making moral judgments about the AI in this scenario and fifty-one percent were comfortable making moral judgments about the autonomous drone. We analyzed the data of the remaining participants for norm and blame responses.

In the first of three studies, we examined whether any asymmetry exists between a human and artificial moral decision maker in the above military dilemma. The study had a 3 × 2 between-subjects design that crossed a three-level *Agent* factor (human pilot vs. drone vs. AI) with a two-level *Decision* factor (launch the strike vs. cancel the strike). Online participants considered the missile strike dilemma and made two moral judgments: whether the agent's decision was morally wrong (Yes vs. No) and how much blame the agent deserved for the decision (on a 0-100 scale). After the latter judgment, participants explained their judgments ("Why does it seem to you that the [agent] deserves this amount of blame?"). After removing participants who expressed serious doubts about the AI's or drone's eligibility for moral evaluation, 501 participants remained for analysis.

When asked about *moral wrongness,* more people regarded the human pilot's decision to *cancel* as wrong (25.8%) than the decision to *launch* (14.8%), whereas the reverse was true for the two artificial agents: more people considered the drone's or AI's decision to *launch* as wrong (27.0%) than the decision to cancel (19.4%), interaction term $p = 0.014$, to $d = .18$. The rates for drone and AI did not differ.

When asked to provide *blame judgments*, people blamed the human pilot far more for *cancelling* ($M = 54.2$) than for launching ($M = 35.3$), whereas they blamed the artificial agents roughly the same amount for cancelling ($M = 43.4$) as for launching ($M = 41.5$); interaction $p = .007$, $d = 0.25$. Blame for the two artificial agents did not differ.

Thus, people appear to grant the human pilot some mitigation when launching the strike (recommended, though not commanded by superiors) but exacerbate blame when he cancels the strike. For artificial agents this differentiation of blame did not appear.

In a next study we tested one hypothesis that could explain this asymmetric blame—namely, that people simply impose different norms on human and artificial agents: In the given scenario, the human *should launch*, and when he cancels, he is blamed for his norm violation; an

[1] This scenario was, and the details of narratives, questions, and results for all studies can be found at http://research.clps.brown.edu/SocCogSci/AISkyMaterial.pdf.

AI or a drone my not have be imposed the same preferential norm about what they should do, which explains the roughly equal blame they receive for the two decisions.

The study had a 3 (Agent: human pilot, AI, drone) × 2 (Decision: launch vs. cancel) between-subjects design, with two dependent variables: *should* and *blame*. Online participants read the same narratives as in in the first study except that the word "autonomous" was removed from all but the first sentence of both the AI and the drone narrative, because the word may have increased the machines' seeming independence from the command structure. Participants first answered the question "*What should the [agent] do?*" (98% of participants provided a response easily verbally classifiable as *launch* or *cancel*). Then people provided blame judgments on a 0-100 scale and offered explanations of their blame judgments. After removing participants who expressed doubts about the artificial agents' moral eligibility, 541 participants remained for analysis.

When asked about what the agent *should* do, people did not impose different norms onto the three agents. Launching the strike was equally obligatory for the human ($M = 83.0\%$), the AI ($M = 83.0\%$), and the drone ($M = 80\%$). Neither human and artificial agents ($p = .45$) nor AI and drone ($p = .77$) differed from one another.

When asked to provide *blame judgments,* people again blamed the human pilot more for cancelling ($M = 52.4$) than for launching ($M = 31.9$), whereas the artificial agents together received more similar levels of blame for cancelling ($M = 44.6$) as for launching ($M = 36.5$), interaction $p = .046$, $d = 0.19$. However, while the *cancel–launch* blame difference for the human pilot was strong, $d = 0.58$, that for the drone was still $d = 0.36$, above the AI's ($d = 0.04$), though not significantly so, $p = .13$.

We then considered a second explanation for the human-machine asymmetry—that people apply different moral justifications for the human's and the artificial agents' decisions. Structurally, this explanation is similar to the case of the mining dilemma, but the specific justifications differ. Specifically, the human pilot may have received less blame for launching than canceling the strike because launching was more strongly justified by the commanders' approval of this decision. Being part of the military command structure, the human pilot thus has justifications available that modulate blame as a function of the pilot's decision. These justifications may be cognitively less available to respondents when they consider the decisions of artificial agents, in part because it is difficult to mentally simulate what duty to one's superior, disobedience, ensuing reprimands, and so forth might look like for an artificial agent and its commanders.

People's verbal explanations following their blame judgments in Studies 1 and 2 provided support for this hypothesis. Across the two studies, participants who evaluated the human pilot offered more than twice as many remarks referring to the command structure (26.7%) as did those who evaluated artificial agents (11%), $p = .001$, $d = .20$. More striking, the *cancel–launch* asymmetry for the human pilot was amplified among those 94 participants who referred to the command structure ($M_{diff} = 36.9$, $d = 1.27$), compared to those 258 who did not ($M_{diff} = 13.3$, $d = 0.36$), interaction $p = .004$. And a *cancel–launch* asymmetry appeared even for the artificial agents (averaging AI and drone) among those 76 participants who referenced the command structure ($M_{diff} = 36.7$, $d = 1.16$), not at all among those 614 who did not make any such reference ($M_{diff} = 1.3$, $d = 0.01$), interaction $p < .001$.

A final study tested the hypothesis more directly that justifications explain the human-machine asymmetry. We increased the human pilot's justification to cancel the strike by including in the narrative the military lawyers' and commanders' affirmation that either decision is supportable, thus explicitly authorizing the pilot to make his own decision (labeled the "decision freedom" manipulation). As a result, the human pilot is now equally justified to cancel or launch the strike, and no relatively greater blame for canceling than launching should emerge.

Two samples combined to make up 522 participants. In the first sample, the decision freedom manipulation reduced the previous cancel-launch difference of 20 points ($d = 0.58$, $p < .001$ in Study 2) to 9 points ($d = 0.23$, $p = .12$). In the second sample, we replicated the 21-point *cancel-launch* difference in the standard condition ($d = 0.69$, $p < .001$) and reduced it to a 7-point difference ($d = 0.21$, $p = .14$) in the decision freedom condition.

In sum, we were able to answer three questions. First, do people find it appropriate to treat artificial agents as targets of moral judgment? Indeed, a majority of people do. Compared to 60-70% of respondents who felt comfortable blaming a robot in our mining dilemmas, 72% across the three missile strike dilemma studies felt comfortable blaming an AI and 51% blamed the autonomous drone. Perhaps the label "drone" is less apt to invoke the image of an actual agent with choice capacity that does good and bad things and deserves praise or blame. In other research we have found that autonomous vehicles, too, may be unlikely to be seen as moral agents (Li et al., 2016). Thus, in empirical studies on artificial agents, we cannot simply assume that people will treat machines as moral decision making agents; it depends on the kind of machine, and we need to actually measure these assumptions.

Second, what norms do people impose on human and artificial agents in a life-and-death dilemma situation? In the present scenarios (as in the mining dilemma), we found no general differences in what actions are normatively expected of human and artificial agents. However, other domains and other robot roles may show differentiation of applicable norms, such as education, medical care, and other areas in which personal relations play a central role.

Third, how do people morally evaluate a human or artificial agent's decision in such a dilemma? We focused on judgments of blame, which are the most sophisticated moral judgments and take into account all available information (about the norm violation, causality, intentionality, and the agent's reasons for acting; Malle et al., 2014; Monroe & Malle, 2017). Our results show that people's blame judgments differ between human and artificial agents, and these differences appear to arise from different moral justifications that people have available for, or grant to, artificial agents. People mitigated their blame for the human pilot when the pilot launched the missile strike, because he was going along with the superiors' recommendation and therefore had justification to launch the strike; by contrast, people exacerbated blame when the pilot cancelled the strike, because he was going against the superiors' recommendations. Blame judgments differed less to not at all for artificial agents, and our hypothesis is that most people did not grant the agents justifications that referred back to the command structure they were part of. In fact, it is likely that many people simply did not think of the artificial agents as embedded in social-institutional structures and, as a result, they explained and justified those agents' actions, not in terms of the roles they occupied, but in terms of the inherent qualities of the decision.

Discussion

Overall, our empirical results suggest that many (though not all) human observers will form moral judgments about artificial systems that make decisions in life-and-death situations. People tend to apply very similar norms to human and artificial agents about how the agents *should* decide, but when they judge the moral quality of the agents' actual decision, their judgments tend to differ; and that is likely because these moral judgments are critically dependent on the kinds of justifications people grant the agents. People seem to imagine the psychological and social situation that a human agent is in and can therefore detect, and perhaps vicariously experience, the decision conflict the agent endures and the social pressures or social support the agent receives. This process can invoke justifications for the human's decision and thus lead to blame mitigation (though sometimes to blame exacerbation). In the case of artificial agents, by contrast, people have difficulty imagining the agent's decision process or "experience," and justification or blame mitigation will be rare. As a result, artificial and human agents' decisions may be judged differently even if the *ex-ante* norms are the same.

If people fail to infer the decision processes and justifications of artificial agents, these agents will have to generate justifications for their decisions and actions, especially when the latter are unintuitive or violate norms. While it is an open question what kinds of justifications will be acceptable to humans, it is clear that these justifications need to make explicit recourse to normative principles that humans uphold. That is because justifications often clarify why one action, violating a less serious norm, was preferable over the alternative, which would have violated a more serious norm. This requirement for justifications, in turn, places a significant constraint on the design of architectures for autonomous agents: Any approach to agent decision making that only *implicitly* encodes decisions or action choices will come up short on the justification requirement, because it cannot link choices to principles. This shortcoming applies to agents governed by Reinforcement Learning algorithms (Abel et al., 2016) and even sophisticated Cooperative Inverse Reinforcement Learning approaches (Hadfield-Menell et al., 2016), because the agents learn how to act from observed behaviors without ever learning the reasons for any of the behaviors.

It follows that artificial agents must know at least some of the normative principles that guide human decisions in order to be able to generate justifications that are acceptable to humans. Perhaps agents could rely on such principles in generating justifications even when the behavior in reality was not the result of decisions involving those principles. Such an approach may succeed for cases in which the agent's behavior aligns with human expectations (because, after all, the system did the right thing), but it is likely to fail when no obvious alignment can be established (precisely because the agent did not follow any of the principles for making its decisions; see also (Kasenberg et al., 2018). But this approach is at best posthoc-rationalization and, if discovered, is likely to be considered deceptive, jeopardizing human trust in the decision system. In our view, a better approach would be for artificial agents to ground their decisions in human normative principles in the first place; then generating justifications amounts to pointing to the obeyed principles, and when a norm conflict occurs, the justification presents that the chosen option obeyed the more important principles. (Kasenberg and Scheutz, 2018) have started to develop an ethical planning and reasoning framework with explicit norm representations that can handle ethical decision making even in cases of norm conflicts. Within this framework, dedicated algorithms will allow for justification dialogues in which the artificial agent can be asked, in natural language, to justify its actions, and it does with recourse to normative principles in factual and counterfactual situations (Kasenberg et al., 2019).

Conclusion

Human communities work best when members know the shared norms, largely comply with them, and are able to justify a decision to violate one norm in service of a more important one. As artificial agents become part of human communities, we should make similar demands on them. Artificial agents embedded in human communities will not be subject to exactly the same norms as humans are, but they will have to be aware of the norms that apply to them and comply with the norms to the extent possible. However, moral judgments are based not only on an action's norm compliance but also on the reasons for the action. If people find a machine's reasons opaque, the machines must make themselves transparent, which includes justifying their actions by reference to applicable norms. If machines enter society that make life-and-death decisions or at least assume socially influential roles, they will have to demonstrate their ability to act in norm-compliant ways, express their knowledge of applicable norms before they act, and offer appropriate justifications, especially in response to criticism, after they acted. It is up to us how to design artificial agents, and endowing them with this form of moral or at least norm competence will be a safeguard for human societies, ensuring that artificial agents will be able to improve the human condition.

Acknowledgments

This project was supported by a grant from the Office of Naval Research (ONR), No. N00014-16-1-2278. The opinions expressed here are our own and do not necessarily reflect the views of ONR

References

Abel D, MacGlashan J and Littman ML (2016) Reinforcement learning as a framework for ethical decision making. In: *AAAI Workshop: AI, Ethics, and Society, volume WS-16-02 of 13th AAAI Workshops*, 2016. AAAI Press.

Arkin R (2015) The case for banning killer robots: Counterpoint. *Communications of the ACM* 58(12): 46–47. DOI: 10.1145/2835965.

Arkin RC (2009) Governing Lethal Behavior in Autonomous Robots. Boca Raton, FL: CRC Press.

Asaro PM (2012) A body to kick, but still no soul to damn: Legal perspectives on robotics. In: Lin P, Abney K, and Bekey G (eds) *Robot Ethics: The Ethical and Social Implications of Robotics*. MIT Press, pp. 169–186. Available at: http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=6733967 (accessed 5 October 2015).

Awad E, Dsouza S, Kim R, et al. (2018) The moral machine experiment. *Nature* 563(7729): 59–64. DOI: 10.1038/s41586-018-0637-6.

Baron M (2011) The standard of the reasonable person in the criminal law. In: R.A. Duff, Lindsay Farmer, S.E. Marshall, et al. (eds) *The Structures of the Criminal Law*. Oxford, UK: Oxford University Press, pp. 11–35.

Bonnefon J-F, Shariff A and Rahwan I (2016) The social dilemma of autonomous vehicles. *Science* 352(6293): 1573–1576. DOI: 10.1126/science.aaf2654.

Briggs G and Scheutz M (2017) The case for robot disobedience. *Scientific American* 316(1): 44–47. DOI: 10.1038/scientificamerican0117-44.

Bringsjord S (2019) Commentary: Use AI to stop carnage. Available at: https://www.timesunion.com/opinion/article/Commentary-Use-AI-to-stop-carnage-14338001.php (accessed 8 October 2019).

de Graaf M and Malle BF (2017) How people explain action (and autonomous intelligent systems should too). In: *2017 AAAI Fall Symposium Series Technical Reports*. FS-17-01. Palo Alto, CA: AAAI Press, pp. 19–26.

Foot P (1967) The problem of abortion and the doctrine of double effect. *Oxford Review* 5: 5–15.

Funk M, Irrgang B and Leuteritz S (2016) Enhanced information warfare and three moral claims of combat drone responsibility. In: Nucci ED and Sio FS de (eds) *Drones and Responsibility: Legal, Philosophical and Socio-Technical Perspectives on Remotely Controlled Weapons*. London, UK: Routledge, pp. 182–196.

Hadfield-Menell D, Russell SJ, Abbeel P, et al. (2016) Cooperative inverse reinforcement learning. In: Lee DD, Sugiyama M, Luxburg UV, et al. (eds) *Advances in Neural Information Processing Systems 29*. Curran Associates, Inc., pp. 3909–3917. Available at: http://papers.nips.cc/paper/6420-cooperative-inverse-reinforcement-learning.pdf.

Harbers M, Peeters MMM and Neerincx MA (2017) Perceived autonomy of robots: Effects of appearance and context. In: *A World with Robots*. Intelligent Systems, Control and Automation: Science and Engineering. Springer, Cham, pp. 19–33. DOI: 10.1007/978-3-319-46667-5_2.

Hood G (2016) *Eye in the sky*. Bleecker Street Media, New York, NY. Available at: http://www.imdb.com/title/tt2057392/ (accessed 30 June 2017).

Kahn, Jr. PH, Kanda T, Ishiguro H, et al. (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, New York, NY, 2012, pp. 33–40. ACM. DOI: 10.1145/2157689.2157696.

Kasenberg D and Scheutz M (2018) Norm conflict resolution in stochastic domains. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Kasenberg D, Arnold T and Scheutz M (2018) Norms, rewards, and the intentional stance: comparing machine learning approaches to ethical training. *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES'18*. AIES '18: 184–190. DOI: 10.1145/3278721.3278774.

Kasenberg D, Roque A, Thielstrom R, et al. (2019) Generating justifications for norm-related agent decisions. In: *12th International Conference on Natural Language Generation (INLG), Tokyo, Japan*, October 2019.

Li J, Zhao X, Cho M-J, et al. (2016) From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. *Society of Automotive Engineers (SAE) Technical Paper 2016-01-0164*. DOI: 10.4271/2016-01-0164.

Lin P (2013) The ethics of autonomous cars. Available at: http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/ (accessed 29 September 2014).

Malle BF and Scheutz M (2015) When will people regard robots as morally competent social partners? In: *Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, pp. 486–491. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=7333667 (accessed 19 July 2016).

Malle BF and Scheutz M (2019) Learning how to behave: Moral competence for social robots. In: Bendel O (ed.) *Handbuch Maschinenethik [Handbook of Machine Ethics]*. Springer Reference Geisteswissenschaften. Wiesbaden, Germany: Springer. DOI: 10.1007/978-3-658-17484-2_17-1.

Malle BF, Guglielmo S and Monroe AE (2014) A theory of blame. *Psychological Inquiry* 25(2): 147–186. DOI: 10.1080/1047840X.2014.877340.

Malle BF, Scheutz M, Arnold T, et al. (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*. New York, NY: ACM, pp. 117–124.

Malle BF, Scheutz M, Forlizzi J, et al. (2016) Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. In: *Proceedings of the Eleventh Annual Meeting of the IEEE Conference on Human-Robot Interaction, HRI'16*. Piscataway, NJ: IEEE Press, pp. 125–132.

Malle BF, Thapa S and Scheutz M (2019) AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In: Aldinhas Ferreira MI, Silva Sequeira J, Singh Virk G, et al. (eds) *Robotics and Well-Being*. Intelligent Systems, Control and Automation: Science and Engineering. Cham: Springer International Publishing, pp. 111–133. DOI: 10.1007/978-3-030-12524-0_11.

Malle BF, Scheutz M, Komatsu T, et al. (2019) Moral evaluations of moral robots. Unpublished Manuscript, Brown University.

Millar J (2014) An ethical dilemma: When robot cars must kill, who should pick the victim? | Robohub. Available at: http://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim/ (accessed 27 September 2014).

Monroe AE and Malle BF (2017) Two paths to blame: Intentionality directs moral information processing along two distinct tracks. *Journal of Experimental Psychology: General* 146(1): 123–133. DOI: 10.1037/xge0000234.

Monroe AE, Dillon KD and Malle BF (2014) Bringing free will down to earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and Cognition* 27: 100–108. DOI: 10.1016/j.concog.2014.04.011.

Pagallo U (2011) Robots of just war: A legal perspective. *Philosophy & Technology* 24(3): 307–323. DOI: 10.1007/s13347-011-0024-9.

Podschwadek F (2017) Do androids dream of normative endorsement? On the fallibility of artificial moral agents. *Artificial Intelligence and Law* 25(3): 325–339. DOI: 10.1007/s10506-017-9209-6.

Scheutz M and Malle BF (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*. Red Hook, NY: Curran Associates/IEEE Computer Society, pp. 36–39.

Sidner S and Simon AF (2016) How robot, explosives took out Dallas sniper. Available at: https://www.cnn.com/2016/07/12/us/dallas-police-robot-c4-explosives/index.html (accessed 8 October 2019).

Sparrow R (2007) Killer robots. *Journal of Applied Philosophy* 24(1): 62–77. DOI: 10.1111/j.1468-5930.2007.00346.x.

Sparrow R (2011) Robotic weapons and the future of war. In: Wolfendale J and Tripodi P (eds) *New Wars and New Soldiers: Military Ethics in the Contemporary World*. Burlington, VA: Ashgate, pp. 117–133.

Thomson JJ (1976) Killing, letting die, and the trolley problem. *The Monist*: 204–217.

Wachter S, Mittelstadt B and Floridi L (2017) Transparent, explainable, and accountable AI for robotics. *Science Robotics* 2(6): eaan6080. DOI: 10.1126/scirobotics.aan6080.

Wang N, Pynadath DV and Hill SG (2016) Trust calibration within a human-robot team: Comparing automatically generated explanations. In: *The Eleventh ACM/IEEE International Conference on Human Robot Interaction, HRI '16*. Piscataway, NJ: IEEE Press, pp. 109–116. Available at: http://dl.acm.org/citation.cfm?id=2906831.2906852 (accessed 27 July 2017).

Wolkenstein A (2018) What has the trolley dilemma ever done for us (and what will it do in the future)? On some recent debates about the ethics of self-driving cars. *Ethics and Information Technology* 20(3): 163–173. DOI: 10.1007/s10676-018-9456-6.

Young AD and Monroe AE (2019) Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology* 85: 103870. DOI: 10.1016/j.jesp.2019.103870.