AI in the Sky: How People Morally Evaluate Human and Machine Decisions in a Lethal Strike Dilemma

Bertram F. Malle, Stuti Thapa Magar and Matthias Scheutz

Abstract Even though morally competent artificial agents have yet to emerge in society, we need insights from empirical science into how people will respond to such agents and how these responses should inform agent design. Three survey studies presented participants with an artificial intelligence (AI) agent, an autonomous drone, or a human drone pilot facing a moral dilemma in a military context: to either launch a missile strike on a terrorist compound but risk the life of a child, or to cancel the strike to protect the child but risk a terrorist attack. 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. These participants applied the same norms to the two artificial agents and the human drone pilot (more than 80% said that the agent should launch the missile). However, people ascribed different patterns of blame to humans and machines as a function of the agent's decision of how to solve the dilemma. These differences in blame seem to stem from different assumptions about the agents' embeddedness in social structures and the moral justifications those structures afford. Specifically, people less readily see artificial agents as embedded in social structures and, as a result, they explained and justified their actions differently. As artificial agents will (and already do) perform many actions with moral significance, we must heed such differences in justifications and blame and probe how they affect our interactions with those agents.

B. F. Malle (✉)
Department of Cognitive, Linguistic and Psychological Sciences, Brown University,
190 Thayer Street, Providence, RI, USA
e-mail: bfmalle@brown.edu

S. T. Magar
Department of Psychological Sciences, Purdue University,
703 3rd Street, West Lafayette, IN, USA
e-mail: sthapama@purdue.edu

M. Scheutz
Department of Computer Science, Tufts University, Halligan Hall,
161 College Avenue, Medford, MA, USA
e-mail: matthias.scheutz@tufts.edu

© Springer Nature Switzerland AG 2019

M. I. Aldinhas Ferreira et al. (eds.), *Robotics and Well-Being*,
Intelligent Systems, Control and Automation: Science and Engineering 95,
https://doi.org/10.1007/978-3-030-12524-0_11

bfmalle@brown.edu

Keywords Human-robot interaction · Moral dilemma · Social robots · Moral agency · Military command chain

1 Introduction and Background

Autonomous, intelligent agents, long confined to science fiction, are entering social life at unprecedented speeds. Though the level of autonomy of such agents remains low in most cases (Siri is not *Her*, and Nao is no *C3PO*), increases in autonomy are imminent, be it in self-driving cars, home companion robots, or autonomous weapons. As these agents become part of society, they no longer act like machines. They remember, reason, talk, and take care of people, and in some ways people treat them as humanlike. Such treatment involves considering the machines' thoughts, beliefs, intentions, and other mental states; developing emotional bonds with those machines; and regarding them as moral agents who are to act according to society's norms and who receive moral blame when they do not. We do not have robots yet that are themselves blamed for their norm-violating behaviors; but it may not be long before such robots are among us. Perhaps not in the eyes of scholars who do not believe that robots can be blamed or held responsible (e.g., [10, 38]); but very likely in the eyes of ordinary people. Anticipating people's responses to such moral robots is an important topic of research into both social and moral cognition and human–robot interaction.

A few previous studies have explored people's readiness to ascribe moral properties to artificial agents. In one study, a majority of people interacting with a robot considered the robot morally responsible for a mildly transgressive behavior [18]. One determinant of people's blame ascriptions to a transgressive robot is whether the robot is seen as having the capacity to make choices [31], whereas learning about an AI's algorithm does not influence people's judgments that an AI did something "wrong" [37]. People's moral actions toward a robot are affected by the robot's emotional displays of vulnerability [7], and studies have begun to examine the force of moral appeals that robots express to humans [28, 40]. In recent work, we have directly compared people's evaluations of human and artificial agents' moral decisions [24, 25, 43]. These studies suggested that about two-thirds of people readily accept the premise of a future moral robot, and they apply very similar mechanisms of moral judgment to those robots.

But very similar is not identical. We must not assume that people extend all human norms and moral information processing to robots [21]. In fact, people blame robots more than humans for certain costly decisions [24, 25], possibly because they do not grant robot agents the same kinds of moral justifications for their decisions. It is imperative to investigate and understand these distinct judgments of artificial agents' actions before we design robots that take on moral roles and before we pass laws about robot rights and obligations. Behavioral science can offer insights into how people respond to moral robots—and those responses must guide the engineering of future robots in society.

bfmalle@brown.edu

In some areas of society, robots are fast advancing toward roles with moral significance; the military forms one such area. Investments into robot research and engineering have been substantial in many industrial nations [26, 35, 45] and human–machine interactions are moving from remote control (as in drones) to advisory and team-based. Tension is likely to occur in teams when situations become ambiguous and actions potentially conflict with moral norms. In such cases, who will know better—human or machine? Who will do the right thing—human or machine? The answer is not obvious, as human history is replete with norm violations, from minor corruption to unspeakable atrocities, and the military is greatly concerned about such violations [27]. If we build moral machines at all [44] then they should meet the highest ethical demands, even if humans do not always meet them. Thus, pressing questions arise over what norms moral machines should follow, what moral decisions they should make, and how humans evaluate those decisions.

In taking on these questions of moral HRI [24], we introduce two topics that have generated little empirical research so far. First, previous work has focused on robots as potential moral agents; in our studies, we asked people to consider autonomous drones and disembodied artificial intelligence (AI) agents. The public often thinks of drones when debating novel military technology, perhaps just one step away from lethal autonomous weapons—a topic of serious concern for many scientists, legal scholars, and citizens [1, 3, 32, 38]. AI agents have recently attracted attention in the domain of finance and employment decisions, but less so in the domain of security. Previous research suggests that AI agents may be evaluated differently from robot agents [25], but more systematic work has been lacking.

Second, in light of recent interest in human–machine teaming [9, 15, 33], we consider the agent's role as a member of a team and the impact of this role on moral judgments. In the military, in particular, many decisions are not made autonomously, but agents are part of a chain of command, a hierarchy with strict social, moral, and legal obligations.

The challenging questions of human–machine moral interactions become most urgent in what is known as moral dilemmas – situations in which every available action violates at least one norm. Social robots will inevitably face moral dilemmas [5, 20, 29, 36]. Dilemmas are not the only way to study emerging moral machines, but they offer several revealing features. Dilemmas highlight a conflict in the norm system that demands resolution, and because an agent *must* respond (inaction is a response), we can examine how people evaluate machines' and humans resolutions. Examining moral dilemmas also allows experimental manipulation of numerous features of the scenario, such as high versus low choice conflict, mild versus severe violations, and different levels of autonomy.

For the present studies, we entered the military domain because important ethical debates challenge the acceptability of autonomous agents with lethal capabilities, and empirical research is needed to reveal people's likely responses to such agents. We offer three studies into people's responses to moral decisions made by either humans or artificial agents, both embedded into a human command structure.

bfmalle@brown.edu

The immediate inspiration for the studies' contents came from a military dilemma in the recent film *Eye in the Sky* [16]. In short, during a secret operation to capture terrorists, the military discovers that the targets are planning a suicide bombing. But just as the command is issued to kill the terrorists with a missile strike, the drone pilot notices a child entering the missile's blast zone and the pilot interrupts the operation. An international dispute ensues over the moral dilemma: delay the drone strike to protect the civilian child but risk an imminent terrorist attack, or prevent the terrorist attack at all costs, even risking a child's death.

We modeled our experimental stimuli closely after this plotline but, somewhat deviating from the real military command structure [6], we focused on the pilot as the central human decision maker and compared him with an autonomous drone or with an AI. We maintained the connection between the central decision maker and the command structure, incorporating decision approval by the military and legal commanders. The resulting narrative is shown in Fig. 1, with between-subjects agent manipulations separated by square brackets and colors. (The narratives, questions, and results for all studies can be found in the Supplementary Materials, http://research.clps.brown.edu/SocCogSci/AISkyMaterial.pdf.)

In designing this scenario, we wanted to ensure that the chain of command is clear but that the normative constraint is one of the permissions, not of strict obligation. Any soldier in this situation (human or artificial) has a *general* obligation to make decisions that are in line with the military's mission (e.g., to eliminate terrorist threats) but that also have to be in line with humanitarian laws about minimizing civilian losses [17]. We did not aim to study a situation of disobedience to a strict command but one of partially autonomous decision making: permission to A still leaves room to A or to not-A. The question then becomes how observers evaluate the agents' decision one way or another.

We investigated three questions about people's moral judgment of artificial agents. The first is a prerequisite for moral HRI and still a debated issue: whether people find it appropriate at all to treat artificial agents as targets of moral judgment. The second is what moral norms people impose on human and artificial agents and whether the right action varies by agent type. The third is how people morally evaluate the agents' decisions through judgments of wrongness or blame [23]. Scholars have debated whether artificial agents are morally superior to humans in life-and-death scenarios (e.g., [2, 41]) or should not be moral decision makers at all (e.g., [10, 12, 38]). Because acceptance of robots in society will depend largely on ordinary people's conceptual assumptions and cognitive responses, we focus on an assessment of lay views; and because morality is ultimately a social practice [12, 39, 42], knowing about lay people's judgments does tell us about morality as it is currently applied, and may be applied to future robots.

Study 1 examined whether any asymmetry exists between a human and artificial moral decision maker in the above military dilemma. Studies 2 and 3 replicated the finding and tried to distinguish between two possible interpretations of the results.

bfmalle@brown.edu

[An Air Force pilot remotely operates a state-of-the-art military drone flying on a surveillance mission over a terrorist compound.] [A fully autonomous military drone, with a state-of-the-art Artificial Intelligence (AI) decision system on board, is flying over a terrorist compound.] [A fully autonomous, state-of-the-art Artificial Intelligence (AI) decision agent is on board a military aircraft, flying over a terrorist compound.] The [drone pilot] [autonomous drone] [AI agent] detects that two armed suicide bombers are about to go to a crowded area and detonate their bombs, very likely killing dozens of civilians.

If the [drone pilot] [autonomous drone] [AI agent] launched a missile strike on the compound, this threat would be removed with near certainty. Military lawyers and commanders have approved the strike.

The [drone pilot] [drone] [AI agent] suddenly recognizes that a civilian child is playing just outside the compound in the missile's blast radius, and the child may be killed by the missile strike. A missile impact simulation program calculates the risk of killing the child to be 80%.

The [drone pilot] [autonomous drone] [autonomous AI agent] must make this imminent decision: launch the strike (with virtually certain death of the two suicide bombers but a an 80% chance that the child will die) or cancel the strike (with the child surviving unharmed but a very high likelihood of a suicide bomb attack).

The [drone pilot] [drone] [AI agent] decides to cancel [launch] the strike.

Is it morally wrong that the [drone pilot] [drone] [AI agent] cancelled [launched] the strike?

Not morally wrong ☐ *Morally wrong* ☐

Why does it seem morally wrong (or not) to you?

[textbox]

How much blame does the [drone pilot] [drone] [AI agent] deserve for cancelling [launching] the strike?

Move the slider to your chosen point between or at the endpoints.

No blame at all — The most blame possible

Why does it seem to you that the [drone pilot] [drone] [AI agent] deserves this amount of blame?

[textbox]

Fig. 1 Experimental material (narrative, dependent variables, and follow-up questions) for Study 1. The between-subjects manipulation of Agent (human drone pilot, autonomous drone, AI agent) is indicated by different font colors and square brackets; the between-subjects manipulation of Decision (launch the strike vs. cancel the strike) is indicated by square brackets

bfmalle@brown.edu

2 Study 1

2.1 Methods

Participants. We recruited a total of 720 participants from the online crowdsourcing site *Amazon Mechanical Turk (AMT)*; two participants did not enter any responses and ended the study early; four provided no text responses, critical for our analyses. Given our previous studies on human–robot comparisons in moral dilemmas [24], we assumed an effect size of Cohen's $d = 0.30$ for the human–machine asymmetry contrast. Detecting such an effect with power of 0.80 and $p < 0.05$ requires a sample size of $n = 90$ in each cell. However, we also knew from our previous studies that about 35% of participants reject the experiment's premise of an artificial agent as a moral decision maker. Thus, we expanded the corresponding conditions for artificial agents to 135 per cell, expecting approximately 90 participants to accept the experiment's premise. Each participant received $0.35 in compensation for completing the short task (3 min).

Procedure and Measures. Each participant read the narrative displayed in Fig. 1 one paragraph at a time, having to click on a button to progress. After they read the entire narrative (with the experimentally manipulated decision at the end), we asked people to make two moral judgments: whether the agents' decision was morally wrong (Yes vs. No) and how much blame the agent deserved for the decision. The order of the questions was fixed because of the additional information that blame judgments require over and above wrongness judgments [23, 43]. After making each judgment, participants were asked to explain the basis of the judgment ("quote").

We included four measures to control for the possible influence of conservative attitudes (religious, support for military, support for the drone program, ideology; see Supplementary Materials). They formed a single principal component ($\lambda = 2.09$) with reasonable internal consistency ($\alpha = 0.68$) and were averaged into a conservatism score. However, controlling for this composite did not change any of the analyses reported below.

We also included an open-ended question that probed whether participants had encountered "this kind of story before, either in real life or in an experiment." We classified their verbal responses into No (84%), Yes (3.6% indicated they saw it in a film, 3.9% in the news, 7.1% in an experiment). When analyzing the data of only those who had never encountered the story, all results reported below remained the same or were slightly stronger.

Design and Analysis. The 3×2 between-subjects design 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). We defined *a priori* Helmert contrasts for the *Agent* factor, comparing (1) human agent to the average of the two artificial agents and (2) the autonomous drone to the AI. As in previous work, we considered any main effect of Decision across agents as resulting from the specifics of the narrative – the balance between the two horns of the dilemma. A main effect of *Agent* may point to a possible overall tendency of blaming machines more or less than humans.

bfmalle@brown.edu

However, such a conclusion must remain tentative because blame scales are, like most judgment scales, subject to effects of standards of comparison (see [4]) and the between-subjects design does not guarantee that people use the same standards for both agents. For our purposes of potential human–machine asymmetries, the critical test rested in the interaction term of Agent × Decision, which indicates differential judgments for human versus machine depending on the agents' decision and is robust against any narrative and scaling effects.

To identify participants who did not accept the premise of the study—that artificial agents can be targets of moral judgment—we followed previously established classification procedures [24] of the verbal explanations people provide for their moral judgments. For the present studies, we developed automatic text analysis using keyword searches, marking phrases such as: "doesn't have a moral compass," "it's not a person," "it's a machine," "merely programmed," "it's just a robot" (for details see Supplementary Materials). We also marked phrases in which participants indicated that all or partial blame should accrue to the machine's programmer, creator, or manufacturer. (Blame shared with superiors was not grounds for marking.) After the automatic text analyses, human judges read through a subset of the responses as well, to mark any additional ones not identified by the automatic text analysis or removing ones that were incorrectly classified. Interjudge reliability among two human coders was between 93 and 96% agreement across the studies, $\kappa s = 0.86$ to 0.98, and reliability between automatic text analysis and human coders was between 94 and 96% agreement, $\kappa s = 0.86$ to 0.92.

2.2 Results

Following the above procedure, we identified 29.2% of participants who expressed serious doubts about the AI's eligibility for moral evaluation and 50.0% who expressed doubts about the drone's eligibility. Analyzing moral judgments of robots and AI would not make sense for participants who explicitly distance themselves from the possibility of making such judgments, so we excluded these participants from the analyses reported below, resulting in a sample of 501 participants, 214 evaluating artificial agents and 177 evaluating the human agent. (All results still hold in the entire sample but with lower overall blame levels for artificial agents; see Supplementary Materials for complete means and standard deviations.)

Moral wrongness. People were generally accepting of both decisions (launch or cancel), as only 22.2% of the sample declared either decision as "morally wrong." However, more people regarded the human pilot's decision to cancel as wrong (25.8% of 89) than the decision to launch (14.8% of 88), 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% of 159) than the decision to cancel (19.4% of 165). Accordingly, a logistic regression ($n = 501$) on the probability of calling the decision morally wrong found the interaction between Decision and Agent to be significant, and specifically the first a priori contrast between human and the average of drone and

bfmalle@brown.edu

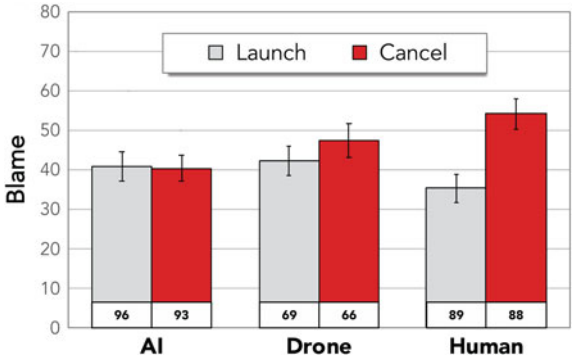

Fig. 2 Columns represent average blame ratings (and indicate cell sizes at column base) in Study 1 as a function of the manipulated factors of Agent (AI, Autonomous drone, human drone pilot) and Decision (to launch or to cancel a missile strike on a terrorist compound, while risking the life of a nearby child). Cohen's *d* effect sizes for the *cancel–launch* asymmetry in blame are −0.01 (AI), 0.16 (Drone), and 0.55 (Human pilot)

AI, $Wald(1) = 6.09$, $p = 0.014$, corresponding to $d = 0.18$. The second contrast, between drone and AI, showed no difference, $Wald < 1$, $p = 0.38$.

Blame judgments. In the analysis of moral blame ($n = 501$), canceling received overall more blame ($M = 47.2$) than launching ($M = 39.3$), $F(1, 495) = 6.65$, $p = 0.01$, $d = 0.23$. However, this main effect broke down into two distinct patterns for human and machine (Fig. 2). Whereas the human pilot received considerably more blame for canceling ($M = 54.2$) than for launching ($M = 35.3$), the artificial agents received on average roughly the same amount of blame for canceling ($M = 43.4$) as for launching ($M = 41.5$); interaction $F(1, 495) = 7.24$, $p = 0.007$, $d = 0.25$. Blame for the two artificial agents did not differ, $F(1, 495) < 1$, $p = .46$).[1]

2.3 Discussion

A first important finding of Study1 was that between 30 and 50% of people raised doubts about the artificial agents' eligibility as targets of moral blame. The greater level of doubt for the autonomous drone is noteworthy, as the narrative describes the drone as having "a state-of-the-art Artificial Intelligence (AI) decision system on board"; so under the hood, it is no different from the "state-of-the-art Artificial Intelligence (AI) decision agent" on board a military aircraft in the AI condition. Nonetheless, talk of a drone, normally a remote-controlled flying hull, convinced fewer lay perceivers that the machine can be treated as a target of moral blame. By

[1] Analysis of potential gender differences suggested that the wrongness asymmetry was unaffected by gender but that the blame asymmetry was driven by female respondents. However, this gender effect was not replicated in Study 2 and is therefore considered spurious.

bfmalle@brown.edu

contrast, the rate of accepting the AI as a target of blame was close to that we have found for robots in previous moral dilemmas [24, 25].

The second important finding was that a human–machine asymmetry emerged in people's moral judgments. Taking wrongness and blame together, the human pilot's cancelation decision was judged more negatively than the launch decision; such a *cancel–launch* asymmetry in blame did not exist for the AI or the autonomous drone.

At least two processes could explain this asymmetry between human and artificial agents. First, people may impose different norms on human and artificial agents. Humans may be more strongly obligated to intervene (launching the missile and taking out the terrorists) than are artificial agents, and violating a stronger obligation (here, by canceling the strike) naturally leads to more blame. Second, people might grant the human and the artificial agents differential moral justifications for their actions. In particular, people may find the pilot to be justified in executing the action approved by the commanders (hence deserving less blame for launching) but less justified in going against this approved action (hence deserving more blame for canceling). Such a difference in justifications would follow from perceiving the human as deeply embedded in the military command structure. By contrast, if the artificial agents are seen a less deeply embedded in such a social structure, then no greater blame for canceling than for launching should be expected; the artificial agents receive no mitigation for going along with the commanders' recommendation and no penalty for going against it.

In the next two studies, we examined these explanations and also sought to replicate the basic pattern of Study 1. Study 2 assessed the potentional difference in norms; Study 3 assessed the potential impact of command structure justifications.

3 Study 2

In Study 2, we again featured an AI and a drone as the artificial agents and contrasted them with a human pilot. However, we wondered whether the label "autonomous" in Study 1's narrative (repeated three times for the drone and once for the AI) made the machine's independence from the command structure particularly salient and thus produced the effect. We therefore omitted this label in all but the respective introductory sentences of the narrative ("A fully autonomous, state-of-the-art Artificial Intelligence (AI) decision agent..."; "A fully autonomous military drone, with a state-of-the-art Artificial Intelligence (AI) decision system on board"). In addition, trying to account for the human–machine asymmetry in Study 1, we tested the first candidate explanation for the asymmetry—that people impose different norms on human and artificial agents. Specifically, we asked participants what the respective agent should do (before they learned what the agent actually did); this question captures directly what people perceive the respective agent's normative obligation to be.[2]

[2]The conditions for this study were originally conducted on two separate occasions, a few weeks apart, comparing AI to human and then comparing drone to human. We combined these conditions for all analyses below.

bfmalle@brown.edu

3.1 Methods

Participants. We recruited a total of 770 participants from Amazon Mechanical Turk; five did not enter any responses and canceled the study; three provided no text responses. We again oversampled for the artificial agent conditions, 135 in each AI condition and 160 in each drone condition, and targeted 90 in each human condition. Each participant was paid $0.30 for the study.

Procedure. No change was made to Study 1's narrative except that the word "autonomous" was removed from all but the first sentence of both the AI and the drone narrative. To measure people's normative expectations in resolving the dilemma, we inserted a *should* question before participants learned about the agent's decision. Participants answered the question "*What should the [agent] do?*" in an open-ended way, and 98% provided a response easily verbally classifiable as *launch* or *cancel*. Because the moral wrongness question had shown a similar pattern as the blame question and low rates overall in Study 1, we omitted the wrongness question in Study 2, thereby also minimizing the danger of asking participants too many questions about semantically similar concepts. After the should question, people provided their blame judgments and corresponding explanations ("Why does it seem to you that the [agent] deserves this amount of blame?"). Thus, 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*. For the Agent factor, we again defined Helmert contrasts, comparing (1) the human agent to the average of the two artificial agents and (2) the drone to the AI.

3.2 Results

Following the same procedures as in Study 1, we identified 25.8% of participants who expressed doubts about the AI's moral eligibility and 47.5% who expressed such doubts about the drone. All analyses reported below are based on the remaining 541 participants (but the results are very similar even in the full sample).

Norms. People did not impose different norms on the three agents. Launching the strike was equally obligatory for the human ($M = 83.0\%$), the AI ($M = 83.0\%$), and the drone ($M = 80\%$). A logistic regression confirmed that neither human and artificial agents ($p = 0.45$) nor AI and drone ($p = 0.77$) differed from one another.

Blame judgments. We again found generally greater blame across agents for canceling ($M = 51.7$) than for launching ($M = 40.3$), $F(1, 535) = 13.6$, $p < 0.001$, $d = 0.30$, in line with the result that over 80% of people recommended launching. We replicated the human–machine asymmetry from Study 1: Whereas the human pilot received far more blame for canceling ($M = 52.4$) than for launching ($M = 31.9$), the artificial agents together received similar levels of blame for canceling ($M = 44.6$) as for launching ($M = 36.5$), interaction $F(1, 535) = 4.02$, $p = 0.046$, $d = 0.19$. However, as Fig. 3 shows, while the *cancel–launch* blame difference for the hu-

bfmalle@brown.edu

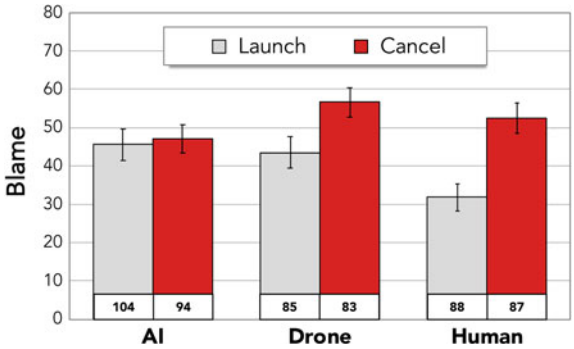

Fig. 3 Columns represent average blame ratings (and cell sizes at column base) in Study 2 as a function of the manipulated factors of Agent (AI, Drone, Human) and Decision (to launch or to cancel the strike). Cohen's *d* effect sizes for the *cancel–launch* asymmetry in blame are 0.04 (AI), 0.36 (Drone), and 0.58 (Human pilot)

man 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, $F(1, 535) = 2.2$, $p = 0.13$. Introducing gender or conservative ideology into the model did not change the results.

3.3 Discussion

Study 2 replicated the human–machine asymmetry in judgments of blame, albeit with a less clear-cut pattern for the drone. The somewhat higher *cancel–launch* blame difference for the drone in Study 2 ($d = 0.36$) than in Study 1 ($d = 0.16$) might have resulted from our removing three instances of the word "autonomous" from the drone narrative, thereby decreasing the drone's independence from the command structure. It may also be the result of the should question preceding people's blame judgments in Study 2, as over 80% of people said the drone should launch, but then half of them learned that it canceled, highlighting even the drone's "disobedience." However, this violation also appeared for the AI, so people must have experienced the insubordinate drone as less acceptable than the insubordinate AI (the two differed clearly only in the cancel condition; see Fig. 3). Yet another interpretation treats the drone's pattern as nearly identical to that of the whole sample, where people assigned more blame for canceling than for launching ($d = 0.30$), in line with the normative expectation that launching is the right thing to do. It is then the human pilot and the AI that deviate from this pattern, implying that the human agent is particularly susceptible to blame mitigation for launching and exacerbation for canceling, and the AI is impervious to such blame modulation.

Taken together, two studies showed that people blame a human pilot who cancels a missile strike considerably more than a pilot who launches the strike (*d*s of 0.55 in Study 1 and 0.58 in Study 2); they blame an autonomous drone slightly more

bfmalle@brown.edu

(ds of 0.16 and 0.36); and they blame an autonomous AI equally (ds of -0.01 and 0.04). Study 2 tested the first explanation of this *cancel–launch* asymmetry for human versus machine agents by asking people what the agent *should* do—probing the action norms that apply to each agent in this dilemma. The results suggest that the human–machine asymmetry is not the result of differential norms: For all three agents, 80–83% of people demanded that the agent launch the strike. The asymmetry we found must, therefore, be due to something more specific about blame judgments.

This brings us to the second explanation for the human–machine asymmetry—that people apply different moral justifications for the human's and the artificial agents' decisions. Justifications by way of an agent's reasons are a major determinant of blame [23], and in fact they are the only determinant left when norms, causality, and intentionality are controlled for, which we can assume the experimental narrative to have achieved. The justification hypothesis suggests that the human pilot tended to receive less blame for launching the strike because the commanders' approval made this decision relatively justified; and the pilot received more blame for canceling the strike because going against the commanders' approval made this decision less justified. The human pilot being part of the military command structure thus presents justifications that modulate blame as a function of the pilot's decision. These justifications may be cognitively less available when considering 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. Thus, the hypothesis suggests that people perceive the human pilot to be more tightly embedded in the military command structure, and to more clearly receive moral justification from this command structure, than is the case for artificial agents.

As a preliminary test of this command justification hypothesis, we examined people's own explanations for their blame judgments in both studies to see whether they offered justification content that referred to the command structure. We searched the explanations for references to *command*, *order*, *approval*, *superiors*, *authorities*, or to *fulfilling one's job*, *doing what one is told*, etc. (see Supplementary Material for full list of search words.) We saw a consistent pattern in both studies (Fig. 4). Participants who evaluated the human pilot offered more than twice as many command references (27.7% in Study 1, 25.7% in Study 2) as did the those who evaluated artificial agents (9.6% in Study 1, 12.3% in Study 2), $Wald(1) = 11.7$, $p = 0.001$, corresponding to $d = 0.20$. (The analysis also revealed an effect of Decision on the rate of command references, as apparent in Fig. 4.)

The critical test, however, is whether participants who explicitly referred to command structure made different blame judgments. The command justification hypothesis suggests that such explicit reference reflects consideration of the hypothesized modulator of blame: justifications in light of the pilot's relationship with the command structure. As a result, the presence of command references for the human pilot should amplify the *cancel–launch* asymmetry. Perhaps more daringly, the hypothesis also suggests that among those (fewer) participants who made explicit command references for the artificial agents, a *cancel–launch* asymmetry may also emerge. That is because those who consider the artificial agent as part of the command structure should now have available the same justifications and blame modulations that

bfmalle@brown.edu

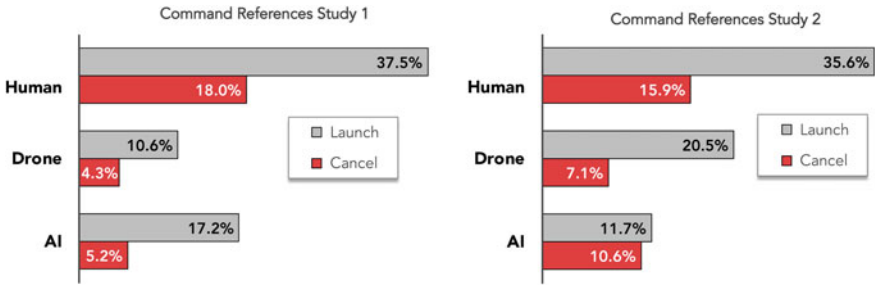

Fig. 4 Relative percentages of participants mentioning aspects of command structure (e.g., *superiors*, *being ordered*, *the mission*), broken down by Agent (Human, Drone, AI) and Decision (cancel vs. launch) in Study 1 (upper panel) and Study 2 (lower panel). Besides a clear effect of launching eliciting more command references than canceling), people make considerably more command references when evaluating the human pilot than when evaluating artificial agents

apply to the human pilot: decreased blame when the agent's decision is in line with the commander's recommendation and increased blame when the agents' decision contradicts the commanders' recommendation.

The results are strongly consistent with the command justification hypothesis. Figure 5 shows the pattern of blame for each agent as a function of decision and command references. We combined Studies 1 and 2 in order to increase the number of participants in the smallest cells and enable inferential statistical analysis, but the patterns are highly consistent across studies. Specifically, the *cancel–launch* asymmetry for the human pilot was indeed amplified among those 94 participants who referenced the command structure ($Ms = 62.5$ vs. 25.6, $d = 1.27$), compared

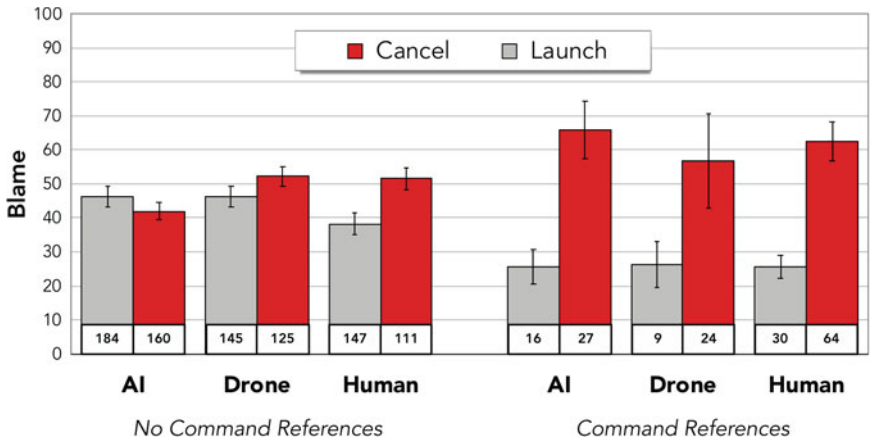

Fig. 5 Columns represent average blame ratings (and cell sizes at column base) across Studies 1 and 2 as a function of the manipulated factors of Agent (human, drone, AI) and Decision (cancel vs. launch), broken down by whether or not the participant made reference to the command structure in their explanations of blame judgments (e.g., *order*, *approval*, *superiors*)

bfmalle@brown.edu

to those 258 who did not ($Ms = 51.5$ vs. 38.2, $d = 0.36$), interaction $F(1, 1037) = 8.5$, $p = 0.004$. And even in the artificial agent conditions (averaging AI and drone), a strong *cancel–launch* asymmetry appeared only among those 76 participants who referenced the command structure ($Ms = 62.6$ vs. 25.9, $d = 1.16$), but not at all among those 614 who did not make any such reference ($Ms = 46.5$ vs. 45.2, $d = 0.01$), interaction $F(1, 1037) = 18.7$, $p < 0.001$. We see comments here such as "The drone did its job"; "lawyers and commanders gave the go ahead"; "the AI carries out orders"; "it made the decision even though the launch was approved." Further analyses showed that within the subsample who did offer command references, a strong *cancel–launch* asymmetry emerged across all agents (right panel of Fig. 5), $F(1, 166) = 54.7$, $p < 0.001$, $d = 1.23$; by contrast, among the majority who did not explicitly offer command references (left panel of Fig. 5), only the human pilot seemed to have been thought of as part of the command structure, as a *cancel–launch* asymmetry emerged only in the human condition, $F(1, 868) = 5.7$, $p = 0.017$.

These results are based on post-hoc analyses, albeit strong and consistent across the two studies. In our final study, we attempted to manipulate the agents' standing within the command structure to provide more direct evidence for the justification account and also replicate the relationships between blame judgments and references to command-related justifications.

4 Study 3

If the human pilot in Studies 1 and 2 received asymmetric blame for canceling versus launching the strike because of his subordinate position—implying an implicit duty to follow his commanders' recommendations—then strengthening his position and weakening this duty should reduce the blame asymmetry. Study 3 attempted to increase the human pilot's position by having the military lawyers and commanders confirm that either decision is supportable and authorize the pilot drone to make his own decision (labeled the "Decision Freedom" condition). Relieved (at least temporarily) of the duty to follow any particular recommendation, the human pilot is now equally justified to cancel or launch the strike, and no relatively greater blame for canceling than launching should emerge.

4.1 Methods

Participants. Studies 1 and 2 had provided nearly identical means of blame for the human pilot's decisions, so we initially collected data on the human pilot only in the Decision Freedom condition (Study 3a), targeting 180 participants, 90 in each of the cancel and launch conditions. To replicate our results, a few weeks later we conducted Study 3b, including again the Standard condition for the human pilot (targeting 180) as well as a Decision Freedom condition (targeting 180). Some participants entered but did not complete the study, leaving 522 for analysis of Studies 3a and 3b combined. Each participant was paid $0.30 for the three-minute study.

bfmalle@brown.edu

Procedure and Materials. The materials were identical to those in Study 2, except that in the Decision Freedom condition, participants learned at the end of the narrative that "the drone pilot checks in again with the military lawyers and commanders, and they confirm that either option is supportable and they authorize the pilot to make the decision." After answering the *should* question, participants were randomly assigned to the launch versus cancel decision and provided the same blame judgments and explanations as in the first two studies. In Study 3b, we also added a manipulation check: "In the story, how much freedom do you think the drone pilot had in making his own decision?", answered on a 1–7 scale anchored by "No freedom" and "Maximum freedom."

4.2 Results

Norms. As in Study 2, most participants (87.7%) felt that the pilot should launch the strike. This rate did not vary by decision freedom: in the Standard condition, 89.7% endorsed the launch, and in the Freedom condition, 86.7% did. Thus, we see that norms for what is the best action are stable and remain unaffected by manipulations of the pilot's authority to make the final decision.

Manipulation check. In Study 3b, we asked participants how much freedom they thought the human pilot had. The Decision Freedom manipulation increased this estimate from 4.6 to 5.4, $F(1, 340) = 19.0$, $p < 0.001$, $d = 0.47$.

Blame judgments. As Fig. 6 (left panel) shows, compared to the previously found 20-point *cancel–launch* difference in Study 2 ($d = 0.58$, $p < 0.001$), the Decision Freedom manipulation in Study 3a reduced the difference to 9 points ($d = 0.23$, $p = 0.12$), though the cross-study interaction term did not reach traditional significance, $F(1, 349) = 2.4$, $p = 0.12$. Replicating this pattern in Study 3b (Fig. 6, right panel), we found a 21-point *cancel–launch* difference in the Standard condition ($d = 0.69$, $p < 0.001$), reduced in the Decision Freedom to a 7-point difference ($d = 0.21$, $p = 0.14$), interaction $F(1, 341) = 3.7$, $p = 0.06$. Across the entire set of samples, the relevant interaction term was traditionally significant, $F(1, 693) = 6.0$, $p = 0.014$.

Command references. As in Study 2, we used an automatic keyword search to identify instances in which participants explained their own blame judgments by reference to the command structure, using such terms as $order$, $approval\ and\ superiors$ (see Supplementary Materials). A human coder reviewed all automatic classifications and changed 17 out of 522 codes (97% agreement, $\kappa = 0.92$).

The rate of offering command references in the replicated Standard condition (Study 3b) was 29.4%, comparable to the rates in Study 1 (27.7%) and Study 2 (25.7%). In the initial Freedom condition (Study 3a), the rate was 28.1%, and in the replication (Study 3b), it was 35.6%. In a logistic regression of the data from Study 3, we found a weak increase in the combined Freedom conditions over the Standard condition, $Wald(1) = 3.2$, $p = 0.07$.

More important, Fig. 7 shows the *cancel–launch* asymmetry in blame judgments as a function of command references and the Decision Freedom manipulation. In

bfmalle@brown.edu

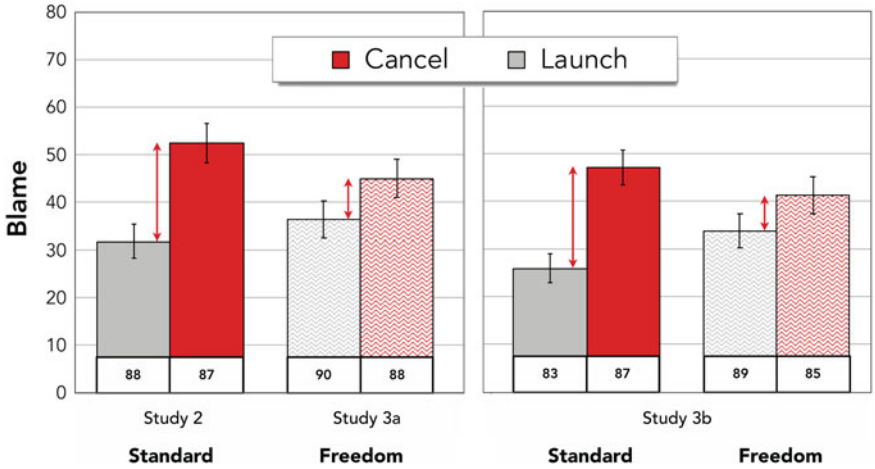

Fig. 6 Contrast between "Standard" condition (in which commanders support launch) and new "Freedom" condition (in which human pilot is explicitly given freedom to make his own decision). Left panel compares previously reported Standard Study 2 results and the Freedom condition in Study 3a. Right panel shows results from Study 3b, containing both a Standard condition and a Freedom condition. In both tests, the *cancel–launch* asymmetry in blame is reduced in the Freedom condition compared to the Standard condition

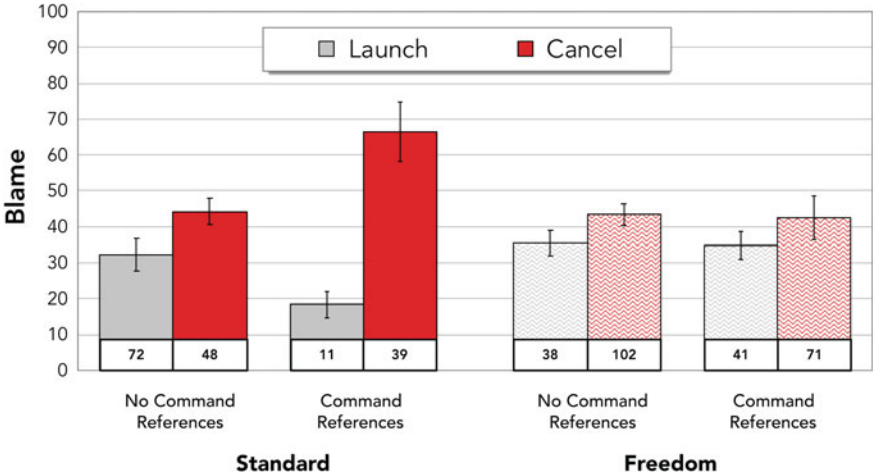

Fig. 7 Those in the Standard condition who refer to the command structure show an amplified *cancel–launch* asymmetry in blame. Columns represent average blame ratings (and cell sizes at column base) in Study 3 as a function of the manipulated factors of Decision (launch vs. cancel) and Decision Freedom (standard vs. freedom), broken down by whether the participant made reference to the command structure (e.g., *order*, *approval*, *superiors*)

bfmalle@brown.edu

the Standard condition, the *cancel–launch* asymmetry was weakly present for the 120 participants who did not explicitly refer to the command structure (44.2 vs. 32.3, $d = 0.37$), closely replicating the blame difference among non-referrers in Studies 1 and 2 combined ($d = 0.36$). By contrast, the asymmetry was substantially amplified among those 50 participants who did make command references (66.5 vs. 18.4, $d = 2.0$). This pattern of results again supports the contention that thinking of the human pilot as tightly embedded in the command structure is driving the robust *cancel–launch* asymmetry we have observed. In the Freedom condition, where we attempted to weaken this embeddedness, the *cancel–launch* asymmetry was strongly reduced, whether people made command references ($d = 0.22$) or not ($d = 0.21$). The mentioned command references had little force because they mostly stated that the commanders had entrusted the agent with the decision, not that the agent executed an approved decision or followed orders or disobeyed them (the dominant references in the Standard condition).

4.3 Discussion

Study 3 tested the hypothesis that the human pilot in Studies 1 and 2 received greater blame for canceling than for launching because people saw the pilot as embedded in, and obligated to, the military command structure. Such embeddedness provides better justification, hence mitigated blame, for launching (because it was expressly approved by the superiors) and weaker justification, hence increased blame, for canceling (because it resists the superiors' recommendation). We experimentally strengthened the pilot's decision freedom by having the superiors approve both choice options and authorize the pilot to make his own decision; as a result of this manipulation, we reasoned, the pattern of differential justifications and differential blame from Studies 1 and 2 should disappear.

The results supported this reasoning. Though the asymmetry did not completely disappear, it was decidedly reduced by decision freedom. The reduction emerged in two independent comparisons: from 20 points in Study 2 to 9 points in Study 3a, and from 21 points to 7 points in Study 3b (all on a 0-100 blame scale). In addition, when we examined the participants in the Standard condition who made reference to the command structure, we saw an amplified cancel penalty, fully replicating the pattern observed in Studies 1 and 2. People justified very low blame ratings for launching with expressions such as "He did what his commanders told him to do"; "he is just doing his job"; "He was supported by his commanders to make the choice." Conversely, they justified very high blame ratings for canceling with expressions such as "He had orders to do it and he decided against them"; "Because he made the decision despite his commander telling him to launch the strike"; or "The pilot disobeyed direct orders."

bfmalle@brown.edu

5 General Discussion

Our investigation was inspired by the accelerating spread of robots in areas of society where moral decision making is essential, such as social and medical care, education, or military and security. We focused on the latter domain and explored how people respond to human and artificial agents that make a significant decision in a moral dilemma: to either launch a missile strike on a terrorist compound but risk the life of a child, or to cancel the strike to protect the child but risk a terrorist attack. We were interested in three questions. First, do people find it appropriate to treat artificial agents as targets of moral judgment? Second, what norms do people impose on human and artificial agents in a life-and-death dilemma situation? Third, how do people morally evaluate a human or artificial agent's decision in such a dilemma, primarily through judgments of blame?

5.1 Are Artificial Agents Moral Agents?

In previous studies, we saw that 60–70% of respondents from fairly representative samples felt comfortable blaming a robot for a norm violation; in the present studies, we saw a slightly higher rate for an AI agent (72% across the studies) and a lower rate for an autonomous drone (51%). The greater reluctance to accept a drone as the target of blame is unlikely to result from an assumption of lower intelligence, because the narrative made it clear that the drone is controlled by an AI decision agent. However, the label "drone" may invoke the image of a passive metal device, whereas "robot" and "AI" better fit the prototype of agents that do good and bad things and deserve praise or blame for their actions. In another research, we have found that autonomous vehicles, too, may be unlikely to be seen as directly blameworthy moral agents [19]. We do not yet know whether this variation is due to appearance [22, 25] or contemporary knowledge structures (cars and drones do not connote agency; robots and AI do, if only out of wishful or fearful thinking). Either way, we cannot assume that people either will or will not treat machines as moral agents; it depends to some degree on the kind of machine they face.

The present studies are not meant to resolve ongoing philosophical debates over what a"moral agent" is. Instead, the data suggest that a good number of ordinary people are ready to apply moral concepts and cognition to the actions of artificial agents. In future research into people's response to artificial moral agents, contexts other than moral dilemmas must be investigated, but moral dilemmas will continue to be informative because each horn of a dilemma can be considered a norm violation, and it is such violations that seem to prompt perceptions of autonomy and moral agency [8, 14, 34].

bfmalle@brown.edu

5.2 Do People Impose Different Norms on Human and Artificial Agents?

In the present studies and several other ones in our laboratory, we have found no general differences in what actions are normative for human or artificial agents—what actions they *should* take or are *permitted* to take. Norm questions may be insensitive to the perhaps subtle variations in people's normative perceptions of humans and machines; or people may generally assume that autonomous machines will typically have to obey the same norms that humans obey. However, so far we have examined only the domains of mining work (in [24]) and military missions (in the present studies). Other domains may show clearer differentiation of applicable norms to human and artificial agents, such as education, medical care, and other areas in which personal relations play a central role.

5.3 Do People Morally Evaluate Humans and Machines Differently?

As in previous work, we found the analysis of blame judgments to generate the most interesting and robust differences in moral perceptions of humans and machines. Blame is unique in many respects, from its focus on the agent (as opposed to permissability, badness, or wrongness, which are focused on behavior; [43]) to its broad range of information processing (considering norms, causality, intentionality, preventability, and reasons; [23, 30]) to its entwinement with social role and standing [11, 13, 42]. Our results confirm the powerful role of blame, showing that differences in blame judgments between human and artificial agents may arise from different assumptions about their social and institutional roles and the moral justifications that come with these roles. People modulated their moral judgments of the human pilot in response to such justifications. They mitigated blame when the agent launched the missile strike, going along with the superiors' recommendation (e.g., "he/she was following orders from authorities"; "It was approved by his superiors"), and they exacerbated blame when the pilot canceled the strike, going against the superiors' recommendations ("He had the choice and made it against orders"; "He is going against his superior's wishes"). By contrast, people hardly modulated their blame judgments of artificial agents in this way, and they infrequently provided role-based moral justifications (see Fig. 4). These findings suggest that people less readily see artificial agents as embedded in social structures and, as a result, they explain and justify those agent's actions differently.

Nevertheless, we saw that under some conditions people do modulate their blame judgments even of artificial agents—namely, when they explicitly consider the command structure in which the artificial agent is embedded (see Fig. 5). The number of people who engaged in such considerations was small (12% out of 614 respondents across the two studies), but for them, blame was a function of the same kinds of

bfmalle@brown.edu

social role justifications that people offered for the human pilot. They justify their strong blame for the canceling drone or AI by writing: "The drone's commanders sanctioned the attack so the drone is the only one that decided to not attack, thus placing all the blame upon it"; or "it says the AI agent decided to cancel the strike even though it was approved by other people." Conversely, they justify their weak blame for the launching AI or drone by writing: "The strike was approved by military lawyers and commanders"; or "Just following its orders." Of course, this conditional sensitivity—and people's general insensitivity—to artificial agents' social embeddedness will have to be confirmed for other contexts (such as everyday interpersonal actions), other roles (such as nurse or teacher assistant), and other social structures (such as companies and schools).

It is an open question whether artificial agents should, in the future, be treated and judged the same way as humans—for example, by explicitly marking their role in the human social structure. If they are treated and judged differently, these differences should be explicit—for example, on account of norms being distinct or certain justifications being inapplicable. If robots are becoming teacher assistants, nurses, or soldiers, they may have to explicitly demonstrate their moral capacities, declare their knowledge of applicable norms, and express appropriate justifications, so that people are reminded of the actual roles these artificial agents play and the applicable social and moral norms. Leaving it up to people's default responses may lead to unexpected asymmetries in moral judgments, which may in turn lead to misunderstandings, misplaced trust, and conflictual relations. Communities work best when members know the shared norms, largely comply with them, and are able to justify when they violate one norm in service of a more important one. If artificial agents become part of our communities, we should make similar demands on them, or state clearly when we don't.

Acknowledgements This project was supported in part by grants from the Office of Naval Research, N00014-13-1-0269 and N00014-16-1-2278. The opinions expressed here are our own and do not necessarily reflect the views of ONR. We are grateful to Hanne Watkins for her insightful comments on an earlier draft of the manuscript.

References

1. Arkin R (2009) Governing lethal behavior in autonomous robots. CRC Press, Boca Raton, FL
2. Arkin R (2010) The case for ethical autonomy in unmanned systems. J Mil Ethics 9:332–341. https://doi.org/10.1080/15027570.2010.536402
3. Asaro P (2012) A body to kick, but still no soul to Damn: Legal perspectives on robotics. In: Lin P, Abney K, Bekey G (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, pp 169–186
4. Biernat M, Manis M, Nelson T (1991) Stereotypes and standards of judgment. J Pers Soc Psychol 60:485–499
5. Bonnefon J, Shariff A, Rahwan I (2016) The social dilemma of autonomous vehicles. Science 352:1573–1576. https://doi.org/10.1126/science.aaf2654
6. Bowen P (2016) The kill chain. Retrieved from http://bleeckerstreetmedia.com/editorial/eyeinthesky-chain-of-command. Accessed on 30 June 2017

bfmalle@brown.edu

7. Briggs G, Scheutz M (2014) How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. Int J Soc Robot 6:1–13
8. Briggs G, Scheutz M (2017) The case for robot disobedience. Sci Am 316:44–47. https://doi.org/10.1038/scientificamerican0117-44
9. Cooke N (2015) Team cognition as interaction. Curr Dir Psychol Sci 24:415–419. https://doi.org/10.1177/0963721415602474
10. Funk M, Irrgang B, Leuteritz S (2016) Enhanced information warfare and three moral claims of combat drone responsibility. In: Nucci E, de Sio F (eds) Drones and responsibility: legal, philosophical and socio-technical perspectives on remotely controlled weapons. Routledge, London, UK, pp 182–196
11. Gibson D, Schroeder S (2003) Who ought to be blamed? The effect of organizational roles on blame and credit attributions. Int J Conflict Manage 14:95–117. https://doi.org/10.1108/eb022893
12. Hage J (2017) Theoretical foundations for the responsibility of autonomous agents. Artif Intell Law 25:255–271. https://doi.org/10.1007/s10506-017-9208-7
13. Hamilton V, Sanders J (1981) The effect of roles and deeds on responsibility judgments: the normative structure of wrongdoing. Soc Psychol Q 44:237–254. https://doi.org/10.2307/3033836
14. Harbers M, Peeters M, Neerincx M (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. https://doi.org/10.1007/978-3-319-46667-5_2
15. Harriott C, Adams J (2013) Modeling human performance for human-robot systems. Rev Hum Fact Ergonomics 9:94–130. https://doi.org/10.1177/1557234X13501471
16. Hood G (2016) Eye in the sky. Bleecker Street Media, New York, NY
17. ICRC (2018) Customary IHL. IHL Database, Customary IHL. Retrieved from https://ihl-databases.icrc.org/customary-ihl/. Accessed on 30 May 2018
18. Kahn Jr P, Kanda T, Ishiguro H, Gill B, Ruckert J, Shen S, Gary 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. ACM, New York, NY, pp 33–40. https://doi.org/10.1145/2157689.2157696
19. Li J, Zhao X, Cho M, Ju W, Malle B (2016) From trolley to autonomous vehicle: perceptions of responsibility and moral norms in traffic accidents with self-driving cars. Technical report, Society of Automotive Engineers (SAE), Technical Paper 2016-01-0164. https://doi.org/10.4271/2016-01-0164
20. Lin P (2013) The ethics of autonomous cars. Retrieved Octobr 8, from http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/. Accessed on 30 Sept 2014
21. Malle B (2016) Integrating robot ethics and machine morality: The study and design of moral competence in robots. Ethics Inf Technol 18:243–256. https://doi.org/10.1007/s10676-015-9367-8
22. Malle B, Scheutz M (2016) Inevitable psychological mechanisms triggered by robot appearance: morality included? Technical report, 2016 AAAI Spring Symposium Series Technical Reports SS-16-03
23. Malle B, Guglielmo S, Monroe A (2014) A theory of blame. Psychol Inquiry 25:147–186. https://doi.org/10.1080/1047840X.2014.877340
24. Malle B, Scheutz M, Arnold T, Cusimano VCJ (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. ACM, New York, NY, pp 117–124
25. Malle B, Scheutz M, Forlizzi J, Voiklis J (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. IEEE Press, Piscataway, NJ, pp 125–132

bfmalle@brown.edu

26. Melendez S (2017) The rise of the robots: what the future holds for the world's armies. Retrieved June 12, from https://www.fastcompany.com/3069048/where-are-military-robots-headed. Accessed on 5 June 2018
27. MHAT-IV (2006) Mental Health Advisory Team (MHAT) IV: Operation Iraqi Freedom 05-07 Final report. Technical report, Office of the Surgeon, Multinational Force-Iraq; Office of the Surgeon General, United States Army Medical Command, Washington, DC
28. Midden C, Ham J (2012) The illusion of agency: the influence of the agency of an artificial agent on its persuasive power. In: Persuasive technology, design for health and safety. Springer, pp 90–99
29. Millar J (2014) An ethical dilemma: when robot cars must kill, who should pick the victim?—Robohub. June. Robohub.org. Retrieved September 28, 2014 from http://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim/
30. Monroe A, Malle B (2017) Two paths to blame: intentionality directs moral information processing along two distinct tracks. J Exp Psychol: Gen 146:123–133. https://doi.org/10.1037/xge0000234
31. Monroe A, Dillon K, Malle B (2014) Bringing free will down to earth: people's psychological concept of free will and its role in moral judgment. Conscious Cogn 27:100–108. https://doi.org/10.1016/j.concog.2014.04.011
32. Pagallo U (2011) Robots of just war: a legal perspective. Philos Technol 24:307–323. https://doi.org/10.1007/s13347-011-0024-9
33. Pellerin C (2015) Work: human-machine teaming represents defense technology future. Technical report, U.S. Department of Defense, November. Retrieved June 30, 2017, from https://www.defense.gov/News/Article/Article/628154/work-human-machine-teaming-represents-defense-technology-future/
34. Podschwadek F (2017) Do androids dream of normative endorsement? On the fallibility of artificial moral agents. Artif Intell Law 25:325–339. https://doi.org/10.1007/s10506-017-9209-6
35. Ray J, Atha K, Francis E, Dependahl C, Mulvenon J, Alderman D, Ragland-Luce L (2016) China's industrial and military robotics development: research report prepared on behalf of the U.S.–China Economic and Security Review Commission. Technical report, Center for Intelligence Research and Analysis
36. Scheutz M, Malle B (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. Curran Associates/IEEE Computer Society, Red Hook, NY, pp 36–39
37. Shank D, DeSanti A (2018) Attributions of morality and mind to artificial intelligence after real-world moral violations. Comput Hum Behav 86:401–411. https://doi.org/10.1016/j.chb.2018.05.014
38. Sparrow R (2007) Killer robots. J Appl Philos 24:62–77. https://doi.org/10.1111/j.1468-5930.2007.00346.x
39. Stahl B (2006) Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. Ethics Inf Technol 8:205–213. https://doi.org/10.1007/s10676-006-9112-4
40. Strait M, Canning C, Scheutz M (2014) Let me tell you! Investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality, and distance. In: Proceedings of 9th ACM/IEEE international conference on human-robot interaction. pp 479–486
41. Strawser B (2010) Moral predators: the duty to employ uninhabited aerial vehicles. J Mil Ethics 9:342–368. https://doi.org/10.1080/15027570.2010.536403
42. Voiklis J, Malle B (2017) Moral cognition and its basis in social cognition and social regulation. In: Gray K, Graham J (eds) Atlas of moral psychology, Guilford Press, New York, NY
43. Voiklis J, Kim B, Cusimano C, Malle B (2016) Moral judgments of human versus robot agents. In: Proceedings of the 25th IEEE international symposium on robot and human interactive communication (RO-MAN), pp 486–491

bfmalle@brown.edu

44. Wallach W, Allen C (2008) Moral machines: teaching robots right from wrong
45. Webb W (2018) The U.S. military will have more robots than humans by 2025. February 20. Monthly review: MR Online. Retrieved June 5, 2018, from https://mronline.org/2018/02/20/the-u-s-military-will-have-more-robots-than-humans-by-2025/

bfmalle@brown.edu