AI IN THE SKY*

BERTRAM F. MALLE & STUTI THAPA MAGAR
Department of Cognitive, Linguistic, and Psychological Sciences, Brown University, 190 Thayer Street, Providence, RI, USA

MATTHIAS SCHEUTZ
Department of Computer Science, Tufts University, Medford, MA, USA

Artificial intelligent agents are increasingly taking on tasks that are subject to moral judgments. Even though morally competent artificial agents have yet to emerge, we need insights from empirical science to anticipate how people will respond to such agents and how these responses should influence agent design. Three studies featuring a moral dilemma in a national security context suggest that people apply the same norms to artificial agents as they apply to humans, but they still ascribe different degrees of blame. The best supported interpretation for this asymmetry is that people grant artificial agents and human agents different justifications for their moral actions.

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 take part in society, humans begin to treat them as human-like, considering their thoughts and intentions; developing emotional bonds with them; and regarding them as moral agents who are to act according to society's norms and get criticized when they do not. We may not have robots yet that can reasonably be blamed for their norm-violating behaviors; but it will not be long before such robots are among us. Anticipating people's responses to such moral robots is the goal of this paper.

A few previous studies have documented people's readiness to ascribe moral capacities to artificial agents[1,2]. More recently, researchers have directly compared people's evaluations of moral decisions by human and artificial agents[3–5]. These studies suggest that about two thirds of people readily accept

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

the premise of a future moral robot, and they apply very similar moral judgment to those robots[5].

But very similar is not identical. We must not assume that people extend all human norms and moral information processing to robots in the same way they do to other humans. In fact, people blame robots more than humans for certain costly decisions[3,4]. It is imperative to learn about and understand these distinct judgments of artificial agents' actions *before* we design robots that take on moral roles; *before* we pass laws about robot rights and obligations.

One area in which robots are fast advancing toward previously futuristic capacities is the domain of national security and military use. Engineering and research investments are increasing worldwide, 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 despite tight legal restrictions[6]. If we build moral machines at all[7] 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 we focus on two topics that have been previously untouched.

First, previous work has focused on robots as potential moral agents; in our studies we asked people to also consider autonomous drones and disembodied artificial intelligence (AI) agents. Drones have been on the public's mind when thinking about novel military technology, and they are just one or two steps away from autonomous lethal weapons—a topic of serious ethical concern for many scientists, legal scholars, and citizens. 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 [4], but more systematic work has been lacking.

Second, in light of recent interest in human-machine teaming[8–10] we consider the agent's role as a member of a team and the impact of this role on moral judgments. In military contexts, 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"[11–13], and recently the potential dilemmas of autonomous vehicles have been salient[14,15]. For the present studies we entered the military domain because important ethical debates challenge the acceptability of

autonomous soldiers and weapons, and we need empirical research to reveal people's likely responses to such autonomous agents, especially when embedded into a human command structure. Here we offer three studies into people's responses to moral decisions made by either human or artificial agents.

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

We modeled our experimental stimuli closely after this plotline but, somewhat deviating from the real military command structure[17], we focused on the pilot as the central human decision maker and compared him with an autonomous drone or 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 experimental material can be found at http://research.clps.brown.edu/SocCogSci/AISkyMaterial.pdf.

We report here on three studies. Study 1 examined whether any asymmetry exists between a human and artificial moral decision maker in the above military dilemma. Study 2 replicated the finding and tried to distinguish between two possible interpretations of the results. Study 3 further tested the two interpretations by manipulating critical factors.

2. Study 1

2.1. *Methods*

We recruited 720 participants from Amazon Mechanical Turk who received $0.35 in compensation for completing the short task (3.5 minutes). 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 vs. cancel). After reading the narrative featuring one of the agents, participants provided two moral judgments: whether the agent's decision was morally wrong (Yes vs. No) and how much blame the agent deserved for the decision. Each time after making a judgment participants were asked to explain the basis of the judgment[5]. Any main effect of Decision across agents is a result of the specifics of the narrative (the relative attraction of the two horns of the dilemma). The critical test for a human-machine asymmetry lies in the interaction term of Agent × Decision. We defined *a priori* Helmert contrasts for Agent, comparing (1) human agent to the average of the two artificial agents and (2) the autonomous drone to the AI.

In this and the subsequent studies, we identified participants who did not accept the premise of the study—that artificial agents can be moral decision makers. To this end we used automatic text analysis of people's explanations for both moral judgments, identifying phrases such as: "doesn't have a moral compass," "it's not a person," "it's a machine," "merely programmed," etc. 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. Reliability among two human coders was κ = 0.82, 93% agreement; reliability between automatic text analysis and human coders was as high or higher.

2.2. *Results*

Following the above procedures we identified 29% of participants who denied moral agency to the AI and 51% who denied it to the drone. These participants were excluded from analyses. (The majority of excluded participants assigned little or no blame to the artificial agent, so including their data only lowers the overall average of blame for artificial agents and does not alter possible human-machine differences in the evaluation of *cancel* vs. *launch*.)

Moral wrongness. People were generally accepting of either decision (to cancel or to launch the strike), as only 22% of the sample declared either decision as "morally wrong." Nonetheless, more people regarded the human pilot's decision as wrong when he canceled (26%) than when he launched the strike (15%), whereas for the two artificial agents, the trend went in the opposite direction: 20% saw it as wrong that the drone or AI canceled the strike and 28% of people saw it as wrong that it launched the strike. In an ANOVA model, the first *a priori* contrast of human vs. machine (average of drone and AI) was statistically significant, $F(1, 498) = 5.23$, $p = 0.02$.[†] The second contrast showed no difference between drone and AI, $F(1, 498) < 1$.

Blame. A similar human-machine asymmetry emerged as in the wrongness judgments: Statistically controlling for main effects, the human pilot received 7.2 points more blame for canceling than for launching whereas artificial agents (taken together) received 7.2 points *less* blame for canceling than for launching, interaction $F(1, 498) = 5.69$, $p = .017$), $d = 0.22$. There was no difference in overall blame between the two artificial agents, $F(1, 498) < 1$, $p = .40$).

Inspecting directly the cell means in Figure 1 (uncorrected for main effects) suggests that the pilot is blamed both more for canceling ($d = 0.28$) and less for launching ($d = -0.16$) than the average of the artificial agents.

[†] Logistic regression analyses showed the same results, but we report ANOVAs for simplicity.

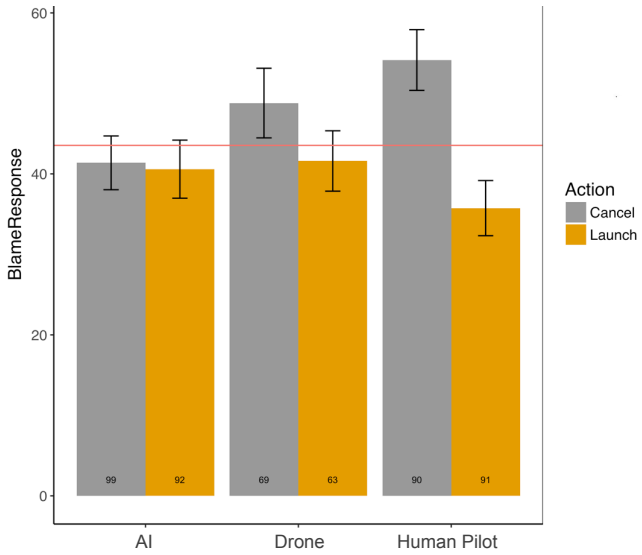

Figure 1. Degree of blame (0-100 scale) for three agents (AI, Drone, Human) deciding to either cancel the strike or launch it, Study 1.

2.3. *Discussion*

We found an asymmetry in moral judgments such that, taking wrongness and blame together, the human agent's decision to launch was judged less negatively and the decision to cancel more negatively, than the corresponding artificial agents' decisions. The patterns for AI and autonomous drone were indistinguishable.

At least two processes could explain this asymmetry. For one, people may apply different norms to human and artificial agents. Intervening (launching the missile and taking out the terrorists) may be a greater obligation for a human than an artificial agent; and violating a stronger obligation naturally leads to more blame. The second process that could explain the asymmetry is a difference in justifications that people grant the human and the artificial agents. People may find the pilot more justified in executing the action approved by the commanders (hence receive less blame for launching) and less justified in going against this approved action (hence receive more blame for canceling). The artificial agents, by contrast, may be seen a less deeply embedded in the military command structure and are therefore blamed equally in the two cases.

In Studies 2 and 3 we sought to differentiate these two interpretations. Because of space constraints we report the results of the two studies together.

3. Studies 2 and 3

To test the first candidate explanation for Study 1's human-machine asymmetry we asked participants in both studies 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 is. In both studies, no difference in obligation between human and artificial agents emerged. People found it equally obligatory for the AI to launch the strike (M = 83.1%) as for the drone to launch the strike (M = 80.0%) as for the human to launch the strike (M = 83.0%), $F(2, 1078) = 1.47$, $p = .23$.

Study 2 (n = 549) featured the AI as the only artificial agent and replicated the blame asymmetry. Controlling for main effects, the human pilot received 9.6 points more blame for canceling than for launching whereas the AI agent received 9.6 points *less* blame for canceling than for launching, interaction $F(1, 545) = 8.61$, $p = .003$, $d = 0.26$. Study 3 (n = 556) featured the drone, but we attempted to decrease its autonomy by removing the label "autonomous" from all but the first time the agent was mentioned. This change was enough to reduce the asymmetry between blame for the human pilot and the drone to nonsignificance, $F(1, 513) = 1.39$, $p = .24$, $d = 0.11$. Thus, the original "autonomous" drone in Study 1 may have exuded greater independence from the command structure and therefore received less blame for canceling (and more for launching), whereas a "mere" drone may be seen as more integrated into the command structure and therefore be blamed similarly to the way humans are.

Conversely, Study 3 attempted to increase the autonomy of the human decision maker by letting the pilot check in with the commanders and receive full authority to make the decision. If the human's obligation to the military command structure increased blame for canceling over launching in Studies 1 and 2, then reducing this obligation by giving the person complete decision authority should eliminate the greater blame for canceling. Indeed, whereas the regular human pilot was blamed over 20 points more for canceling than launching, the authorized pilot was blamed only 8.5 points more. The interaction pattern approached significance, $F(1, 524) = 3.24$, $p = .07$, but was relatively small, $d = 17$. Perhaps more compellingly, whereas the standard human agent was blamed more for canceling than the AI agent in Study 2 (as reported above), the decision-authorized human no longer differed from the AI, $F(1, 721) < 1$.

4. General Discussion

When considering how people perceive human and machine agents that take morally significant actions, a plausible hypothesis is that people prefer "utilitarian" machines: sacrificing a person for the greater good is acceptable for machines but less so for humans. This is not what we found in our studies.

People demanded the same moral actions of human and machine agents, but they blamed human and machines differently for those actions.

Overwhelmingly, participants in our studies wanted to see the missile strike launched and the terrorists killed, even at the risk of killing a child. Naturally, then, people blamed agents who canceled the strike more than agents who launched it; but human agents were blamed even more for canceling (and less for launching) than were artificial agents. Given that agents' obligations were judged as similar, such differences in blame are likely to stem from the justifications people ascribed to each agent[18]. The human pilot appeared to be seen as more strongly embedded in the military command structure and therefore as less justified in going against the "approved" decision to launch and more justified in launching the missile (even if it meant killing a child) because he was following orders. For machines, by contrast, such justifications by way of command structure may not have been as salient, leading to a human-machine blame asymmetry in Studies 1 and 2. It stands at least as an intriguing hypothesis that artificial agents are by default seen as more independent and possibly autonomous (if one accepts them as moral agents in the first place) whereas humans are by default seen as embedded into the social roles and relationships they participate in. Study 3 provided at least tentative evidence that decreasing the machine's autonomy *or* increasing the human's autonomy succeeded in eliminating this asymmetry and equalizing blame for human and machine. If this finding replicates in other contexts as well, it suggests a new demand for robot design: artificial moral agents that are to be treated similarly to human moral agents must be explicitly embedded in a structure of social relations and social norms.

References

1. Kahn, Jr., P. H. *et al.* Do people hold a humanoid robot morally accountable for the harm it causes? in *Proceedings of the Seventh Annual ACM/IEEE International Conference on Human-Robot Interaction* 33–40 (ACM, 2012). doi:10.1145/2157689.2157696
2. Monroe, A. E., Dillon, K. D. & Malle, B. F. 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 (2014).
3. Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J. & Cusimano, C. Sacrifice one for the good of many? People apply different moral norms to human and robot agents. in *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction, HRI'15* 117–124 (ACM, 2015).
4. Malle, B. F., Scheutz, M., Forlizzi, J. & Voiklis, J. 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'16 125–132 (IEEE Press, 2016).
5. Voiklis, J., Kim, B., Cusimano, C. & Malle, B. F. Moral judgments of human vs. robot agents. in *Proceedings of the 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)* 486–491 (IEEE, 2016).
6. MHAT-IV. *Mental Health Advisory Team (MHAT) IV: Operation Iraqi Freedom 05-07 Final report*. (Office of the Surgeon, Multinational Force-Iraq; Office of the Surgeon General, United States Army Medical Command, 2006).
7. Wallach, W. & Allen, C. *Moral machines: Teaching robots right from wrong*. (Oxford University Press, 2008).
8. Cooke, N. J. Team cognition as interaction. *Current Directions in Psychological Science* **24,** 415–419 (2015).
9. Harriott, C. E. & Adams, J. A. Modeling human performance for human–robot systems. *Reviews of Human Factors and Ergonomics* **9,** 94–130 (2013).
10. Pellerin, C. Work: Human-machine teaming represents defense technology future. *U.S. Department of Defense* (2015). Available at: https://www.defense.gov/News/Article/Article/628154/work-human-machine-teaming-represents-defense-technology-future/. (Accessed: 30th June 2017)
11. Lin, P. The ethics of autonomous cars. *The Atlantic* (2013). Available at: http://www.theatlantic.com/technology/archive/2013/10/the-ethics-of-autonomous-cars/280360/. (Accessed: 30th September 2014)
12. Millar, J. An ethical dilemma: When robot cars must kill, who should pick the victim? | Robohub. *Robohub.org* (2014). Available at: http://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim/. (Accessed: 28th September 2014)
13. Scheutz, M. & Malle, B. F. "Think and do the right thing": A plea for morally competent autonomous robots. in (2014).
14. Bonnefon, J.-F., Shariff, A. & Rahwan, I. The social dilemma of autonomous vehicles. *Science* **352,** 1573–1576 (2016).
15. Li, J., Zhao, X., Cho, M.-J., Ju, W. & Malle, B. F. From trolley to autonomous vehicle: Perceptions of responsibility and moral norms in traffic accidents with self-driving cars. in *SAE Technical Paper 2016-01-0164* (2016). doi:10.4271/2016-01-0164
16. Hood, G. *Eye in the sky*. (Bleecker Street Media, New York, NY, 2016).
17. Bowen, P. The kill chain. *Bleecker Street* (2016). Available at: http://bleeckerstreetmedia.com/editorial/eyeinthesky-chain-of-command. (Accessed: 30th June 2017)
18. Malle, B. F., Guglielmo, S. & Monroe, A. E. A theory of blame. *Psychological Inquiry* **25,** 147–186 (2014).