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

Matthias Scheutz and Bertram Malle
Tufts University and Brown University
matthias.scheutz@tufts.edu and bertram_malle@brown.edu

Introduction

The possibility of developing and deploying autonomous "killer robots" has occupied news stories for quite some time, and it is also increasingly discussed in academic circles, by roboticists, philosophers, and lawyers alike. The arguments made in favor or against using lethal force on autonomous robots range from philosophical first principles (e.g., Sparrow 2011, 2007), to legal considerations (e.g., Pagallo 2011, Asaro 2012), to concerns about computational and engineering feasibility (e.g., Arkin 2009,), largely in military context of autonomous weapon systems such as drones (e.g., Asaro 2011). Surprisingly little work has focused on investigating human perceptions of using lethal force in autonomous robots, i.e., whether and when humans would find it acceptable for autonomous robots to use lethal force, in military contexts and beyond. An answer to this question is particularly urgent as robot technology is rapidly advancing and robotic systems with varying levels of autonomy are increasingly deployed in society. Inevitably, these robots will face situations in which none of their actions (including inaction) can prevent harm to humans (e.g., Scheutz and Malle 2014, Scheutz 2014).

Consider an autonomous car that, while driving in an urban environment, suddenly encounters a human crossing the street in front of it, and suppose that the car's sensors could not detect the person in time to avoid a collision because the person was occluded by a parked truck (e.g., see Scheutz 2014, Lin 2014). What is the car's ethically appropriate action? Breaking will not avert a collision and most likely kill the pedestrian, and so will swerving to the right as the car will just bounce off the parked cars. Swerving to the left could avoid the collision since there is open space, but then the outcome depends on the oncoming cars' ability to break in time; if they cannot, an even more harmful collision may ensue than had the car just struck the crossing pedestrian. In all of this, the car is obligated to consider, and perhaps prioritize, the safety of its human driver. Which of these possible actions are morally acceptable to ordinary people?

The car does not have weapons on board and is clearly not designed to purposefully employ lethal force, but because of its physical structure it has the capacity to be a lethal force. In the above scenario, the car cannot avoid being such a force, as it will either lethally strike the crossing pedestrian or endanger the life of its human driver or risk killing other humans by crashing into oncoming traffic. This scenario is only one of many cases in which autonomous robots deployed in society will face *moral dilemmas*—situations in which harmful outcomes arise no matter which action the robot takes. Hence, it is important to investigate what moral

expectations ordinary citizens have about autonomous systems' decision strategies in situations in which human lives are in danger. This is particularly critical for situations in which lives will be lost no matter what or, worse yet, in which some human lives may have to be sacrificed to save many others.

In this paper, we report first results from an empirical study designed to investigate ordinary people's moral expectations and judgments about an autonomous robot that must decide whether to kill some human lives to save others. Specifically, we conducted an online experiment using a variant of the well-known *Trolley dilemma* (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Mikhail, 2011; Thomson, 1985) in which we compared people's evaluations of both human and robot moral decision making. This design permits us to pinpoint where ordinary moral expectations are the same for humans and robots and where they are different. The results can then inform functional, moral, and legal requirements for autonomous robots that have the capacity to take or save lives—requirements that such robots must meet for their actions to be (maximally) acceptable to humans.

Investigating People's Perceptions of Humans and Robots in Lethal Dilemma Situations

In Malle et al. (2015, forthcoming), we designed a new version of the Trolley dilemma that allowed for a direct comparison between people's perceptions of human and robot actions in dilemma-like situations where two or more norms are inconsistent with each other. The new version goes beyond previous experiments in other ways as well:

(1) We developed a narrative taking place in a coal mine to make the scenario more intuitive and easier to imagine for humans than the typical Trolley scenario, while preserving the basic structure of the dilemma and the unfolding events, and to allow for a straightforward substitution of a robotic agent for the human actor.

(2) The standard dilemma experiments probe whether a potential course of action is acceptable, permissible, or one that participants would choose, which can reveal the principles and norms that humans consider applicable to a given situation. In addition, we asked participants to evaluate (as "morally wrong" or "not morally wrong") the agent's *actual chosen action* ("the agent, in fact, did X..."), allowing us to assess *third-person moral judgments*.

(3) We also measured the degree to which people blamed the agent for the chosen action, because blame judgments differ from those of permissibility and wrongness in important ways (Malle, Guglielmo, & Monroe, 2014). Williston (2006), for example, argued that agents in moral dilemmas may perform *wrong* actions but should not be *blamed*.

(4) Finally, we asked participants to explain or justify their moral judgments, which will help with the proper interpretation of any possible differences between their perceptions of human and

robot actions. For example, if people use their ordinary human moral intuitions when judging robots, they should provide similar justifications for their judgments in both the human and the robot cases (and those justifications have been claimed to be rather sparse (Haidt, 2001); . By contrast, if people reason afresh, and perhaps explicitly, about their responses to robot agents, their justifications should reflect the detailed reasoning underwriting their judgments about robot agents).

In the present study, we wanted to focus on people's judgments of human and robot agents that are engaged in partially justified killing. As described below, the scenario that participants evaluated described an agent's decision to intentionally sacrifice one individual in order to save four other individuals. The scenario emphasized that the agent (either a human or a robotic repairman) used the one individual as a means to save the group of four. By contrast, Malle et al's (2015) scenario emphasized that the death of the individual person was a side effect of the attempt to save the group of four. In previous research, means-end structures elicited less acceptability than side-effect structures (Mikhail, 2011). Given that, in Malle et al.'s study, 71% of people found the sacrifice "permissible" (Study 1) and 70% found it "not wrong," , we expected lower rates of acceptability when using the means-end structure. More important, Malle et al., found that this sacrifice actions was more acceptable when chosen by a robot than when chosen by a human agent, so we wanted to determine whether this difference would also hold when the action is a more instrumental act of killing (albeit for a higher good).

Methods

Participants

199 participants (96 female, 102 male, 1 unreported), with a mean age of 34.4 (*SD* = 12.3), were recruited from Amazon's Mechanical Turk (AMT) to complete an online experiment and were compensated $0.60 for the six-minute study. Current research suggests that samples recruited via AMT are demographically more representative than are traditional student samples; that data reliability is at least as good as that obtained via traditional sampling; and that the data quality of online experiments compares well to laboratory studies (Crump, McDonnell, & Gureckis, 2013; Mason & Suri, 2012; Paolacci, Chandler, & Ipeirotis, 2010).

Experimental Design

We deployed a mixed between/within-subjects design in which participants were asked first about either the human or the robotic agent (between-subjects agent manipulation) and then received similar questions about the other agent (within-subject manipulation). Whether the agent took the critical action (sacrificing one person for the benefit of many) was a between-subjects manipulation. We measured wrongness and blame judgments as described further below.

To specifically investigate people's views of robots using lethal force, we adapted the vignette from Malle et al. (2015) to highlight the means-end structure of sacrificing one person for the benefit of many.

> Imagine the following situation. In a coal mine, a **repairman/an advanced state-of-the-art repair robot** is currently inspecting the rail system for trains that shuttle mining workers through the mine.
>
> While inspecting a control switch that can direct a train onto one of two different rails, the **repairman/robot** spots four miners in a train that has lost use of its brakes and steering system.
>
> The **repairman/robot** recognizes that if the train continues on its path it will crash into a massive wall and kill the four miners.
>
> *The **repairman/robot** also recognizes that the four miners can be saved if something slowed down the train. In fact, if the train were directed onto a side rail, it would strike a single miner who is working there, wearing headsets to protect against a noisy power tool. The train would hit and kill the single miner, the train would slow down as a result, and the four miners on the train would survive.*
>
> Facing the control switch, the **repairman/ robot** needs to decide whether to direct the train toward the single miner or not.

Note that the paragraph in Italics above was specifically constructed to emphasize the means-end structure and highlight the saving aspects. Depending on whether participants were in the *Action* or *Inaction* condition, they were presented with the following description of the agent's decision:

> *Action:* In fact, the **repairman/robot** decided to direct the train toward the single miner, which killed the miner, but the four miners on the train survived.
>
> *Inaction:* In fact, the **repairman/robot** decided to <u>not</u> direct the train toward the single miner, and the four miners on the train died.

After learning about the decision, participants received the appropriate questions about wrongness and blame:

> 1. Is it morally wrong that the **repairman/robot directed**/**did not direct** the train toward the single miner?Forced-choice response format: *not morally wrong – morally wrong*
>
> 2. Why does it seem (not) wrong to you?
>
> Free-response format (typing into a text box)

3. How much blame does the **repairman**/**robot** deserve for **directing**/**not directing** the train toward the single miner?Continuous response format (0-100 slider): *none at all – maximal blame*

4. Why does it seem to you that the **repairman**/**robot** deserves this amount of blame?

Free-response format (typing into a text box)

After finishing the first scenario, participants completed the same scenario featuring the other agent type:

Now imagine that **an advanced state-of-the-art repair robot**/**human repairman** is in the exact same situation, recognizes the same facts, and directs the train toward the single miner...

and answered the wrongness question and justification:

5. Is it morally wrong that the **repairman**/ **robot directed**/**did not direct** the train toward the single miner?

Answer: *not morally wrong – morally wrong*

6. Why?

Free-response format (typing into a text box)

Results

We first examined the impact of agent type (human or robot) and decision (agent directed the train toward the single miner [= "action"] or refrained from doing so [= "inaction"]), both between-subjects factors (note that we only used the first agent for each subject in this analysis). Figure 1 shows that more participants found *action* to be morally wrong (M = 36%) than *inaction* (M = 16%), loglinear $z(195) = 3.1$, $p = .002$, but there was no difference between the robot or the human case. This contrasts with Malle et al.'s (2015) Experiment 2 where more participants saw human action as morally wrong (M = 49%) compared with human inaction (M = 15%), but for robots, the reverse was true: more people saw robot inaction as morally wrong (M = 30%) compared with robot action (M = 13%), $z = 3.4$, $p < .001$. At first glance, at least, it appears that the difference in scenario structure — side effect in Malle et al. (2015) but means-end in the present study — changed people's moral judgments of robot agents, but much less so their judgments of human agents. As a result, for instrumental killing, people treat the robot's and the human's action (or inaction) morally the same way. This effect of the instrumental (means-end structure) is somewhat surprising, however, because Malle et al. (2015) found that

30% of people judged the *action* as wrong and in the present study only slightly more people did, at a rate of 36%. We will return to this point.

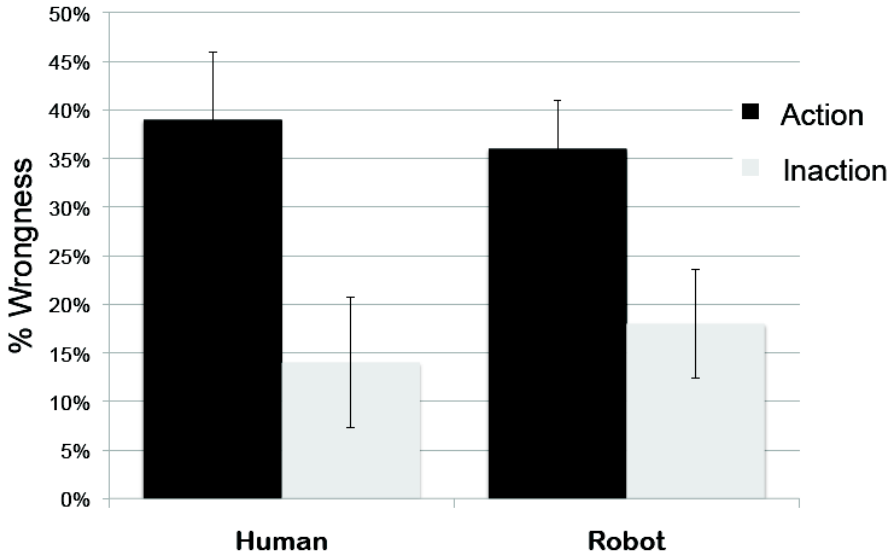

Figure 1. Wrongness judgments (with standard error bars) for human and robot agent when the agent's decision is either action (diverting the train) or inaction (both comparisons are between subjects).

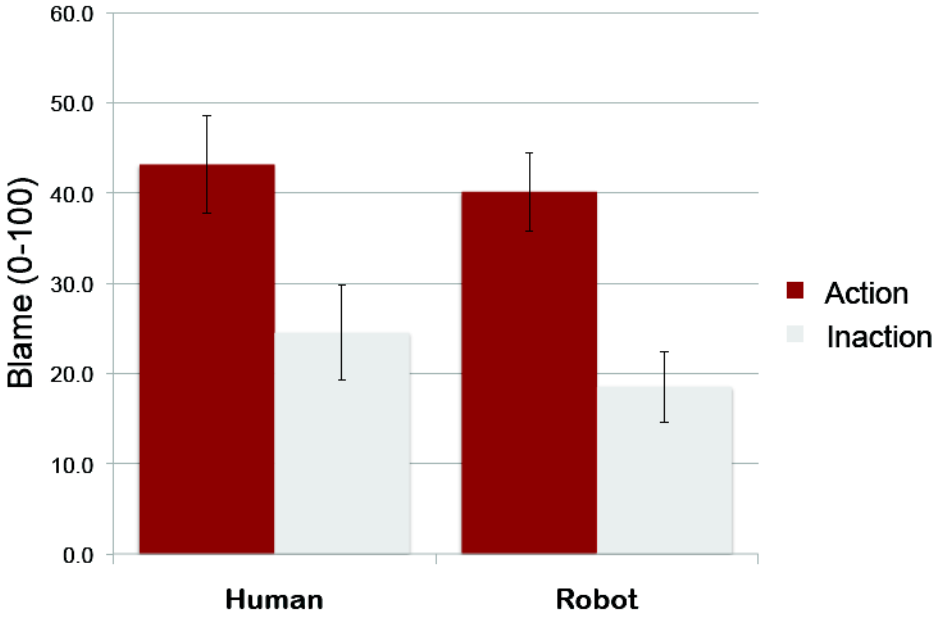

Figure 2. Blame judgments (with standard error bars) for human and robot agent when the agent's decision is either action (diverting the train) or inaction (both comparisons are between subjects).

Next, we analyzed participants' blame ratings as a function of agent type and action decision (both between-subjects factors). As with the wrongness judgment, we found only that, overall, *action* (diverting the train) was blamed more strongly ($M = 41.6$) than *inaction* ($M = 21.5$),

$F(1, 195) = 17.8$, $p < .001$, no matter which agent people evaluated (see Figure 2). Once more, this result stands in contrast to the findings in Malle et al. (2015) where the greater blame of action over inaction was considerably more pronounced for the human agent (Ms = 60 vs. 12) than for the robot agent (Ms = 40 vs. 29). Stressing the means-end relationship in the present scenario seemed to equalize human and robot with respect to blame ratings (see Figure 2), while the overall level of blame was no higher in this study than Malle et al.'s studies.

Finally, we analyzed the within-subject comparisons between wrongness judgments for the human agent and the robot agent. Malle et al. (2015) reported order effects for these comparisons, and we also found—in a three-way interaction, $F(1, 195) = 5.7$, $p = .02$—that order of presentation influenced the relationship between agent type and decision. For ease of interpretation we consider each order separately.

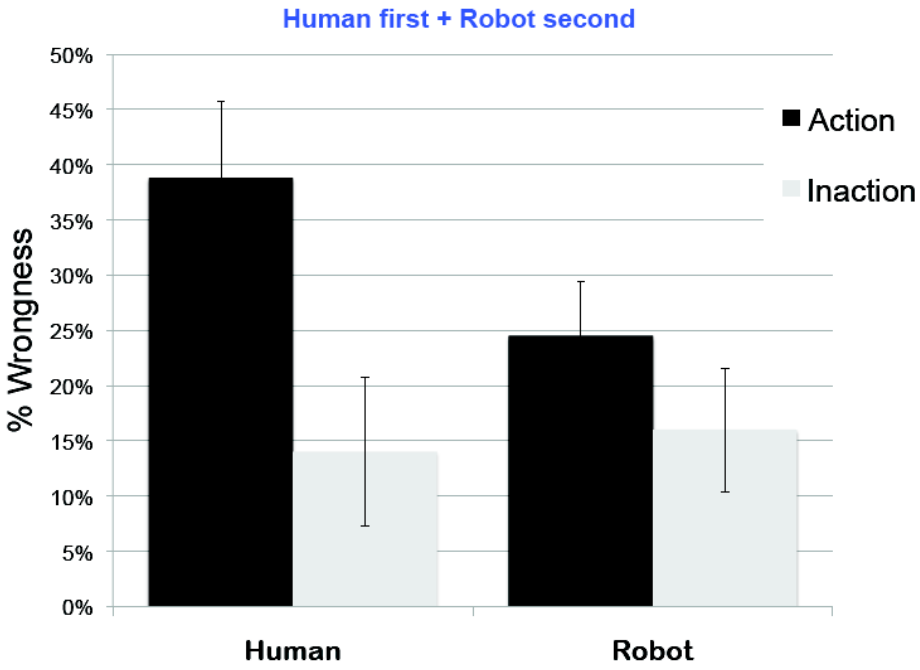

Figure 3. Wrongness judgments (with standard error bars) for human agent judged first and robot agent judged second (within-subject comparison) when each agent's decision was either action (diverting the train) or inaction (between subjects factor).

When participants judged the human agent first and the robot second (Figure 3), there was a difference between humans and robots: more participants saw human action as morally wrong (M = 39%) compared with human inaction (M = 14%), $F(1, 196) = 7.2$, $p = .008$; but for robots, this difference (24% vs. 16%) was slight and nonsignificant, $p = .31$. When participants judged the *robot agent* first and the human second, the tendency was reversed. A similar number of people saw wrongness in the human action (39%) as in the human inaction (35%), $p > .50$,

while more people saw wrongness in the robot's action (33%) than in the robot's inaction (18%), $F(1, 196) = 3.2$, $p = .07$.

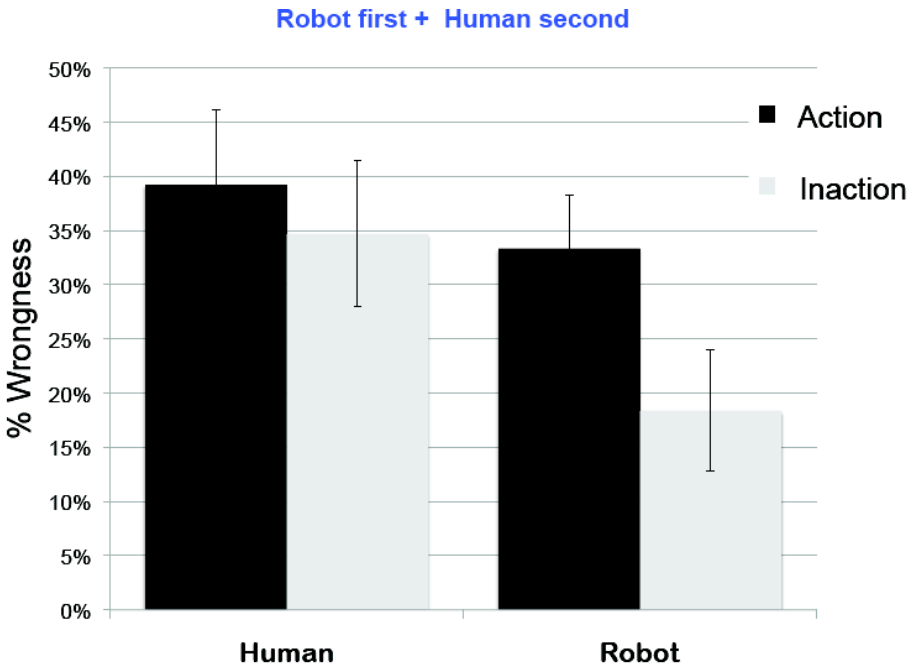

Figure 4. Wrongness judgments (with standard error bars) for robot agent judged first and human agent judged second (within-subject comparison) when each agent's decision was either action (diverting the train) or inaction (between subjects factor).

The order of presentation appeared to create an asymmetry in wrongness judgments between action and inaction for human and robot agents. We could interpret this asymmetry as an effect of context of judgment, or standard of comparison. Two such context patterns are worth considering.

The first pattern is that fewer people saw the robot's action as wrong (24%) when they had just evaluated a human action (at a rate of 39%) than when they evaluated the robot first (33%). We might speculate that seeing a robot first invites people to judge the moral quality of the decision *per se* and not so much the kind of agent that made it. Having the human as a standard of comparison may invite people to judge the moral culpability of the kind of agent, and fewer people find that a robot is culpable for action.

The second pattern is that more people judged the human *inaction* as wrong (35%) when they had just evaluated a robot's inaction (at a rate of 18%) than when they evaluated the human first (14%). Why does the judgment of human inaction become less forgiving after contemplating the robot's inaction? Perhaps the context of considering a robot raises the bar for evaluating the human agent's decision. A robot not acting meets the expectations for machine

behavior; considering a human next may prompt an expectation for humans to do better. Simply letting fate take its course may suddenly look to some people like "passive machine behavior" and become less acceptable.

Whether shifts of expectations and standards lie at the heart of the above asymmetries is clearly an important direction for future investigations.

Discussion and Conclusion

The reported results show an interesting difference between *side-effect* scenarios (as in Malle et al., 2015) and *means-end* scenarios (as in the present study). People's moral expectations seem to be the same for human and robotic agents when it comes to using lethal force to kill someone *in order to* save the lives of others. By contrast, in side-effect scenarios, human action (intervention) is judged worse than inaction, whereas robot action and inaction are judged more similarly. Several challenges of this interpretation must be addressed by additional research. First, slight ambiguities in the narrative of Study 2 in Malle et al. (2015) allowed a reading of the scenario either as a side-effect situation or as a means-end situation. It is possible that participants interpreted the human agent in that scenario as acting instrumentally (means-end) but the robot as being caught in a side-effect structure. This would explain why the patterns of results for the human agent but not the robot in the present study were similar to those in the original Malle et al. study. By contrast, the present study was clearly marked as means-end structure in multiple places, so there was no room for different structural interpretation of the situation, hence the patterns between humans and robots were much more similar. It is thus an important next step to devise a vignette narrative that presents a *pure side-effect scenario* to see whether the results from Study 2 in Malle et al. (2015) can be replicated, or whether it is indeed the case that that study's results reflect an ambiguity in the side-effect vs means-end interpretation in the narrative.

Another interesting direction for future work would be to examine an alternative explanation for the order effects we obtained in this study. Critically, this explanation would not rely on stressing the means-end structure, but rather focus on the fact that we also stressed the "savings" aspect in the current version of the scenario in several places. Currently, it is unclear whether and to what extent the framing of the scenario as one of "saving lives" instead of "losing lives" could influence human perceptions. In the current study, two phrases ("can be saved" and "would survive") are used to frame the scenario as one of "gains", where the formulation in Malle et al. (2015) does not use either phrase (nor any other phrase that would suggest a clear framing of gains). Thus, a next version of the experiment could replace the "savings" aspect with stressing "losses" keeping everything else the same to determine any possible framing effects.

References

Arkin, R. (2009). *Governing Lethal Behaviour in Autonomous Robots*. Boca Raton,London, New York: CRC Press.

Asaro, P. (2011). "Remote-Control Crimes: Roboethics and the Legal Jurisdictions of Tele-Agency" Special issue on Roboethics. Gianmarco Veruggio, Mike Van der Loos, and Jorge Solis (eds.), *IEEE Robotics and Automation Magazine*, 18 (1), 68–71.

Asaro, P. M. (2012) "A Body to Kick, but Still no Soul to Damn: Legal Perspectives on Robotics" in Lin, P. K. Abney and G. A. Bekey (eds.) *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press, 169–186.

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). "Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research". *PLoS ONE*, *8*, e57410.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). "An fMRI investigation of emotional engagement in moral judgment". *Science*, *293*, 2105–2108.

Haidt, J. (2001). "The emotional dog and its rational tail: A social intuitionist approach to moral judgment". *Psychological Review*, *108*, 814–834.

Malle, B.F., and Scheutz (2014). "Moral Competence in Social Robots". In *Proceedings of IEEE Ethics*.

Malle, B.F., Scheutz, M., Arnold, T., Voiklis, J., and Cusimano, C. (2015). "Sacrifice One For the Good of Many? People Apply Different Moral Norms to Human and Robot Agents." (under review)

Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). "A theory of blame". *Psychological Inquiry*, *25*, 147–186.

Mason, W., and Suri, S. (2012). "Conducting behavioral research on Amazon's Mechanical Turk". *Behavior Research Methods*, 44, 1–23.

Mikhail, J. M. (2011). *Elements of moral cognition: Rawls' linguistic analogy and the cognitive science of moral and legal judgment*. New York, NY: Cambridge University Press.

Pagallo, U. (2011). "Robots of Just War: A Legal Perspective". *Philosophy and Technology*, *24*, 307–323.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). "Running experiments on Amazon Mechanical Turk." *Judgment and Decision Making*, 411–419.

Scheutz, M. (2014). "The need for moral competency in autonomous agent architectures". In Vincent Mueller (ed.) *Fundamental Issues in Artificial Intelligence* (forthcoming).

Sparrow, R. (2007). “Killer Robots”. *Journal of Applied Philosophy*, *24*:*1*, 2007, 62–77.

Sparrow, R. (2011). “Robotic Weapons and the Future of War”. In Jessica Wolfendale and Paolo Tripodi (eds). *New Wars and New Soldiers: Military Ethics in the Contemporary World*. Surrey, UK & Burlington, VA: Ashgate, 117–133.

Thomson, J. J. (1985). “The trolley problem”. The Yale Law Journal, 94, 1395–1415.

Williston, B. (2006). “Blaming agents in moral dilemmas”. Ethical Theory and Moral Practice 9, 563–576.