Holding Robots Responsible: The Elements of Machine Morality

Yochanan E. Bigman[1], Adam Waytz[2], Ron Alterovitz[3], and Kurt Gray[1]

In press, Trends in Cognitive Sciences

[1]Department of Psychology and Neuroscience, University of North Carolina at Chapel-Hill.

[2]Kellogg School of Management, Northwestern University.

[3]Department of Computer Science, University of North Carolina at Chapel-Hill.

*Correspondence: ybigman@gmail.com (Y.E. Bigman)

Keywords: Autonomous Machines, Autonomy, Responsibility, Morality

Abstract

As robots become more autonomous, people will see them as more responsible for wrongdoing. Moral psychology suggests that judgments of robot responsibility will hinge on perceived situational awareness, intentionality, and free will—plus anthropomorphism and the robot's capacity for harm. We also consider questions of robot rights and moral decision-making.

Advances in robotics mean that humans already share roads, skies, and hospitals with autonomous machines. Soon, it will become commonplace for cars to autonomously maneuver across highways, military drones to autonomously select missile trajectories, and medical robots to autonomously seek out and remove tumors. The actions of these autonomous machines can spell life and death for humans [1], such as when self-driving vehicles kill pedestrians. When robots harm humans, how will we understand their moral responsibility?

Morality and Autonomy

Philosophy, law, and modern cognitive science all reveal that judgments of human moral responsibility hinge on autonomy [2,3]. This explains why children, who seem to have less autonomy than adults, are held less responsible for wrongdoing. Autonomy is also likely crucial in judgments of robot moral responsibility [4,5]. The reason people ponder and debate the ethical implications of drones and self-driving cars (but not tractors or blenders) is because these machines can act autonomously.

Admittedly, today's robots have limited autonomy, but it is an expressed goal of roboticists to develop fully autonomous robots—machine systems that can act without human input [6]. As robots become more autonomous their potential for moral responsibility will only grow. Even as roboticists create robots with more "objective" autonomy, we note that "subjective" autonomy may be more important: work in cognitive science suggest that autonomy and moral responsibility are more matters of perception than objective truths [3].

Perceiving the Minds of Robots

For programmers and developers, autonomy is understood as a robot's ability to operate in dynamic real-world environments for extended periods of time without external human control [6]. However, for everyday people, autonomy is more likely tied a robot's mental capacities. Some may balk at the idea that robots have (or will have) any human-like mental capacities, but people also long balked at the idea that animals had minds, and now think of them as having rich inner lives.

Of course, animals are flesh and blood whereas machines are silicon and circuits, but research emphasizes that minds are always matters of perception [3,7]. The "problem of other minds" means that the thoughts and feelings of others are ultimately inaccessible, and so we are left to perceive them based upon context, cues, and cultural assumptions. Importantly, people *do* ascribe to machines at least some ability to think, plan, remember, and exert self-control [7,8]—and as when judging humans, people make sense of the morality of robots based upon these ascriptions of mind [8].

How people see mind—i.e., "mind perception"—predicts moral judgments [3], but mind perception is not monolithic: there are many mental abilities [8], some of which (e.g., the ability to plan ahead) are more relevant to autonomy and moral judgment than others (e.g., the ability to feel thirsty). Cognitive science has outlined these autonomy-relevant abilities as they concern humans, but only a subset of these are likely important for making sense of morality in autonomous machines. Here we outline one subset of robot "mental" abilities that likely seem relevant to autonomy (and therefore moral judgment).

Autonomous Elements Tied to Robot Morality

Situation Awareness

For someone to be perceived as morally responsible for wrongdoing, that person must seem to be aware of the moral concerns inherent in the situation [9]. For example, a young child with no understanding about the danger of guns will not be held responsible for shooting someone. For a robot to be held responsible for causing harm, it will likely need to be seen as aware that its actions are indeed harmful. Although today's robots cannot appreciate the depths of others' suffering, they can at least understand some situational aspects. For example, robots can understand whether stimuli belong to protected categories, such as civilians for military drones, pedestrians for autonomous cars, and healthy-organs for medical robots. People already ascribe some of this "meaning-lite" understanding to machines [7], and we suggest that greater ascriptions of situational awareness will increase perceptions of robot responsibility.

Intentionality

Harm-doers are seen as more responsible for intentional actions than for unintentional actions, often because people infer a desire or a reason behind intentional acts [10]. Although people are unlikely to perceive robots as capable of desire, they do see robots as capable of intentionality—

holding a belief that an action will have a certain outcome [7]. This perception is consistent with robots' ability to evaluate multiple response options in the service of achieving a goal [11]. We suggest that the more people see robots as intentional agents—being able to understand and select their own goals—the more they will be ascribed moral responsibility.

Free Will

The ability to freely act—or to "do otherwise" [2]—is a cornerstone of lay judgments of moral responsibility [2]. Although robots are not seen as possessing a rich humanlike free will, they are ascribed the ability to independently implement actions [7]. Consistent with this ascription, today's robots can independently execute action programs [11], however this independence is relatively constrained. The behavior of robots is predictable given the transparency of their (human-given) programming, and predictability undermines perceptions of free will [2]. Technological advances (e.g., deep neural networks) will likely render the minds of machines less transparent to both programmers and perceivers, thereby elevating perceptions of unpredictability. We suggest that as robotic minds become more opaque, people will see robots as possessing more free will—and ascribe them more moral responsibility.

Anthropomorphism

People perceive the mind of machines based on their abilities and behaviors, but also on their appearance. The more humanlike a machine looks, the more people perceive it as having a mind, a phenomenon called anthropomorphism [12]. Individuals vary in their tendency to anthropomorphize, but people consistently perceive more mind—and therefore more moral responsibility—in machines that look and act like humans [13]. We suggest that having humanlike bodies, humanlike voices, and humanlike faces will all cause people to attribute more moral responsibility to machines.

Potential Harm

Even with powerful computational abilities, today's robots are limited in their physical ability to act upon the world. As technology advances, these increased capacities (e.g., the ability to walk, shoot, operate, and drive) will allow robots to cause more damage to humans. Studies reveal that observing damage and suffering lead people to search for an intentional agent to hold responsible for that damage [14]. If people cannot find another person to hold responsible, they will seek other agents—including corporations and gods [14]—and infer the capacity for intention. This link between suffering and intention means that the more robots cause damage, the more they

will seem to possess intentionality, which (as we outline above) will then lead to increased perceptions of moral responsibility. We therefore suggest that causing harm can amplify both perceptions of mind and judgments of moral responsibility.

Future Implications

The future of robotics holds considerable promise, but it is also important to consider what today's semi-autonomous machines might mean for moral judgment. As Box 1 explores, even robots with some perceived mind can help shield their human creators and owners (e.g., corporations and governments) from responsibility. Today's machines are also capable of making some kind of moral decisions, and Box 2 explores whether people actually *want* machines to make these basic decisions.

Although we focus here on moral responsibility, we note that people might also see sophisticated machines as worthy of moral rights. While some might find the idea of robots rights to be ridiculous, the American Society for the Prevention of Cruelty to Robots and a 2017 European Union report both argue for extending some moral protections to machines. Debates about whether to recognize the personhood of robots often revolve around its impact on humanity (i.e., expanding the moral circle to machines may better protect other people), but also involves questions about whether robots possess the *appropriate* mind required for rights. Although autonomy is important for judgments of moral responsibility, discussions of moral rights typically focus on the ability to *feel*. It is an open question whether robots will ever be capable of feeling love or pain—and relatedly, whether people will ever *perceive* these abilities in machines.

Whether we are considering questions of moral responsibility or rights, issues of robot morality may currently seem like science fiction. However, we suggest that *now*—while machines and our intuitions about them are still in flux—is the best time to systematically explore questions of robot morality. By understanding how human minds make sense of morality, and how we perceive the mind of machines, we can help society think more clearly about the impending rise of robots, and help roboticists understand how their creations are likely to be received.

Box 1 Machines can shield humans from responsibility

When people harm others, they often try to avoid responsibility by pointing fingers elsewhere. Soldiers who commit heinous acts invoke the mantra that they were "just following orders" from superior officers. Conversely, superior officers shirk responsibility by claiming that they did not actually pull the trigger. These excuses can work because responsibility is often a zero-sum game. The more we assign responsibility to the proximate agent (the entity who physically perpetrated the harm) the less we assign responsibility to the distal agent (the entity who directed the harm)—and vice versa [3].

As robots spread through society, they will more frequently become the proximal agent in harm-doing: collateral damage will be caused by drones and accidents will caused by self-driving cars. Although humans will remain the distal agents who program and direct these machines, the more that people can point fingers at their autonomous robots, the less they will be held accountable for wrongdoing—a fact that corporations and governments could leverage to escape responsibility for misdeeds. Increasing autonomy for robots could mean increasing absolution for their owners.

Box 2 Do we want machines making moral decisions?

Much discussion in robotics concerns *how* robots should make moral decisions [1], but it is worth asking *whether* they should make moral decisions in the first place. For example, some argue that autonomous military robots (e.g., drones) should never independently make decisions about human life and death. However, others argue in favor of these autonomous military robots, suggesting that they could be programmed to follow the rules of war better than humans.

Putting these ethical debates in perspective is new research revealing that people are reluctant to have machines make any moral decisions—whether in the military, the law, driving, or medicine [8]. One reason for people's aversion to machines making moral decisions is that they see robots as lacking a full human mind [7,8]. Without the full human ability to think and feel, we do not see robots as qualified to make decisions about human lives.

This aversion to machine moral decision-making has seem quite robust [8], but may fade as the perceived mental capacities of machines advance [15]. As the autonomy of machines rises, people may become more comfortable with robots making moral decisions, although people may eventually wonder whether the goals of machines align with their own.

Acknowledgments

We thank Bertram Malle, Ilan Finkelstein, Michael Clamann and an anonymous reviewer for their comments on a draft of this paper. This work has been supported by the National Science Foundation SBE Postdoctoral Research fellowship (1714298) to YEB, by the National Science Foundation awards IIS-1149965 and CCF-1533844 to RA, and a grant from the Charles Koch Foundation to KG.

References

[1] Awad E. et al. (2018) The moral machine experiment. Nature. 563,59–64

[2] Shariff A.F. et al. (2014) Free will and punishment: A mechanistic view of human nature reduces retribution. Psychol Sci. 25,1563–70

[3] Wegner D.M. and Gray K. (2017) *The mind club*, Viking

[4] Kim T. and Hinds P. (2006) Who should I blame? Effects of autonomy and transparency on attributions in human-robot interaction. In *ROMAN 2006 - The 15th IEEE International Symposium on Robot and Human Interactive Communication*, pp. 80–85, IEEE

[5] van der Woerdt S. and Haselager P. (in press) When robots appear to have a mind: The human perception of machine agency and responsibility. New Ideas Psychol

[6] Bekey G.A. (2005). *Autonomous robots : from biological inspiration to implementation and control*, The MIT Press

[7] Weisman K. et al. (2017) Rethinking people's conceptions of mental life. Proc Natl Acad Sci. 114, 11374-11379

[8] Bigman Y.E. and Gray K. (2018) People are averse to machines making moral decisions. Cognition. 181, 21–34

[9] Kissinger-Knox A. et al. (2018). Does non-moral ignorance exculpate? Situational awareness and attributions of blame and forgiveness. Acta Anal. 33, 161–179

[10] Monroe A.E. and Malle B.F. (2017) Two paths to blame: Intentionality directs moral information processing along two distinct tracks. J Exp Psychol Gen. 146, 23–33

[11] Dudek G. and Jenkin M. (2010) *Computational principles of mobile robotics*, Cambridge University Press

[12] de Visser E.J. et al. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. J Exp Psychol Appl. 22, 331–349

[13] Waytz A. et al., (2014) The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. J Exp Soc Psychol. 52, 113–117

[14] Gray K. et al., (2014) The myth of harmless wrongs in moral cognition: Automatic dyadic completion from sin to suffering. J Exp Psychol Gen. 143, 1600–1615

[15] Malle B.F. et al (in press) AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In Robots and well-being (Aldinhas Ferreira I., Silva Sequeira J., Virk G.S., Kadar E.E., and Tokhi O., eds), Springer