Cognition 181 (2018) 21–34

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Original Articles

People are averse to machines making moral decisions

Yochanan E. Bigman*, Kurt Gray

Department of Psychology and Neuroscience, University of North Carolina at Chapel Hill, 235 E Cameron Ave, Chapel Hill, NC 27514, USA

ARTICLE INFO

Keywords:
Mind perception
Morality
Moral agency
Autonomous machines
Skynet
Robots

ABSTRACT

Do people want autonomous machines making moral decisions? Nine studies suggest that that the answer is 'no'—in part because machines lack a complete mind. Studies 1–6 find that people are averse to machines making morally-relevant driving, legal, medical, and military decisions, and that this aversion is mediated by the perception that machines can neither fully think nor feel. Studies 5–6 find that this aversion exists even when moral decisions have positive outcomes. Studies 7–9 briefly investigate three potential routes to increasing the acceptability of machine moral decision-making: limiting the machine to an advisory role (Study 7), increasing machines' perceived experience (Study 8), and increasing machines' perceived expertise (Study 9). Although some of these routes show promise, the aversion to machine moral decision-making is difficult to eliminate. This aversion may prove challenging for the integration of autonomous technology in moral domains including medicine, the law, the military, and self-driving vehicles.

1. Introduction

> "*Decisions about the application of violent force must not be delegated to machines*."
>
> Press release of the International Committee for Robot Arm Control[1]

Machines have long performed boring and repetitive industrial tasks, but the advance of technology is opening new vistas. Today, robotic arms are assisting with life-threatening surgeries (van den Berg, Patil, & Alterovitz, 2017), drones are surveilling and bombing enemy combatants (Horowitz, 2016), and algorithms are making recommendations for criminal sentencing (Angwin, Larson, Surya, & Lauren, 2016). Although humans make the final decision in these moral domains, machines are becoming ever more autonomous; there may soon come a time when machines can make moral decisions for themselves. The question is whether people want machines making autonomous decisions when human lives hang in the balance?

There may be good reason to delegate moral decisions to machines. Machines—and the artificial intelligence that they embody—often make more optimal decisions than human beings in domains including risk management (Heires, 2016), supply chain distribution (Validi, Bhattacharya, & Byrne, 2015), and medical diagnoses (Parkin, 2016). The sheer computational power of machines enable them to accurately compute the flight paths of thousands of planes (Bartholomew-Biggs, Parkhurst, & Wilson, 2003), the best way to manage complex inventories (Cárdenas-Barrón, Treviño-Garza, & Wee, 2012), and even predict human decisions (Wright & Leyton-Brown, 2010). Machines can also beat humans at games long exalted for requiring rationality, intelligence, and strategy, including Chess (Newborn, 2011), Go (Chouard, 2016), and Jeopardy (Markoff, 2011). The success of machine decision-making across these domains may lead people to happily cede moral decisions to them as well, but there are reasons to believe otherwise.

Morality is not like other domains. People hold strong convictions about morality (Skitka, 2010), and these convictions shape cultural identities (Haidt, Koller, & Dias, 1993; Shweder, Mahapatra, & Miller, 1987) and motivate behavior (Hertz & Krettenauer, 2016)—sometimes even irrational behavior (Fehr & Gächter, 2002). Importantly, unlike other decisions, moral decisions are deeply grounded in emotion (Gray, Schein, & Cameron, 2017; Haidt, 2001). This aspects of morality suggest that people may not be amenable to machines making moral decisions. Although machines may have great computational capacities, they seem to lack the ability to feel authentic emotion. In more psychological terms, morality is often seen to require a full human mind (Bastian, Loughnan, Haslam, & Radke, 2012; Gray, Young, & Waytz, 2012), one that can both think and feel. To the extent that machines seem to lack a human mind, they may also seem ineligible to make moral decisions.

Here we investigate whether people are averse to machines making moral decisions, whether this aversion is due—at least in part—to machines lacking a human mind. We then explore whether—and

* Corresponding author.
E-mail address: ybigman@email.unc.edu (Y.E. Bigman).

[1] https://icrac.net/wp-content/uploads/2015/05/Scientist-Call_Press-Release.pdf retrieved January 5th 2018.

https://doi.org/10.1016/j.cognition.2018.08.003
Received 27 March 2018; Received in revised form 2 August 2018; Accepted 3 August 2018

how—this aversion to machine moral decision-making might be decreased.

1.1. The rule—and rules—of machines

The idea of fully autonomous machines was long consigned to science fiction. Early automata may have moved on their own (such as Vaucanson's digesting duck), but were merely a deterministic collection of cogs. Even as technology advanced, machines were still largely deterministic, with their actions fully predictable by their human programming. However, increasing advances in statistical prediction and neural nets allows for ever more autonomous machines—machines which although programmed by humans, can at defy the expectations of their programmers. When an algorithm writes love letters (Roberts, 2017) or gains a personality from browsing the internet (Hunt, 2016) it is anyone's guess what exactly will happen. Even everyday machines are more autonomous than ever; many of us think nothing of how deep learning algorithms decide what news items we see on Facebook (DeVito, 2017), what products we see on Amazon (Chen, Mislove, & Wilson, 2016), and what route we take to work (Yamane et al., 2011).

The increasing autonomy of machines has already impacted important social events such as elections (Hern, 2017), which may influence moral outcomes such as court cases. Although machines are not yet autonomously making moral decisions *per se*, this possibility is not far away. Robotic surgery arms will soon be able to choose how exactly to operate upon a tumor, selecting the path to move through surrounding tissue (Swaney et al., 2017)—with a wrong decision resulting in the death of a patient. Self-driving cars will soon be able to choose how exactly to respond to imminent collisions, deciding whether to kill the driver or multiple bystanders.

Mirroring the increasing autonomy of machines in moral situations, research in psychology and cognitive science has investigated people's perceptions about machine morality. In one popular paper, researchers revealed that people want a self-driving car to save the most number of people—unless they are the driver, in which case they want self-driving cars to save them (Bonnefon, Shariff, & Rahwan, 2016). A burgeoning literature strives to identify an acceptable set of rules, algorithms, or architecture that governs (or at least limits) machine moral behavior (e.g., Arkin, 2009; Conitzer, Sinnott-Armstrong, Borg, Deng, & Kramer, 2017; Kuipers, 2016; van Wynsberghe, 2013; Wiltshire, 2015). Dovetailing with this work are studies examining what kind of decision rules people want machines to follow (Bonnefon et al., 2016; Malle, Scheutz, Arnold, Voiklis, & Cusimano, 2015).

Uncovering rules for machine morality has a distinguished past—starting from Isaac Asimov's (Asimov, 1950) three laws of robotics—and is essential to our technological future. But despite the importance of uncovering *how* machines should make moral decisions, it also important to investigate a basic question: do people think that machines *should* make moral decisions in the first place.

1.2. An aversion to machines making moral decisions?

Autonomous machines can do many things, but people may not want them making moral decisions. If the arc of science fiction is any guide, humans fear machines making decisions when human lives hang in the balance: in 2001: A Space Odyssey (Kubrick, 1968), HAL sends out an astronaut into the void of space, and in The Terminator (Cameron, 1984), SkyNet launches a pre-emptive nuclear strike against humanity. Modern academic works are no less pessimistic, with one popular philosophical treatise arguing that machines making decisions on behalf of humanity might lead to disaster (Bostrom, 2014). Even Elon Musk—an ardent pro-technologist—called the rise of autonomous machines humanity's "biggest existential threat" (McFarland, 2014).

Whether this fear of autonomous machines is misplaced is open to debate—machines may not care enough to rise up and destroy humanity (Pinker, 2016)—but even misplaced aversions have societal impacts. Aversions to vaccines (Hornsey, Harris, & Fielding, 2018), to science (Osborne, Simon, & Collins, 2003), and to change (Pardo del Val & Martínez Fuentes, 2003) all drive behavior and shape policy, and so it is important to explore whether people are averse to machines making moral decisions—and why. We suggest that the potential aversion to machine moral decision-making can be explained (at least in part) by the machines perceived lack of mind.

1.3. Mind (perception) and morality

In law, philosophy, and lay judgments, a complete human mind is seen as a prerequisite for morality (Aristotle, 350BC; Monroe, Dillon, & Malle, 2014; Nahmias, Shepard, & Reuter, 2014; O'Connor, 2000; Robinson, 1996; Rosati, 2016). From the time of the ancient Greeks and Romans, people who "lost their mind" were not considered fully morally responsible (Robinson, 1996). Psychological research reveals that judgments of moral status are tied to a suite of mental capacities—including the ability to freely choose actions (Fischer, 2005; Harris, 2012; Monroe, Brady, & Malle, 2017; Nahmias et al., 2014) and the ability to appreciate the consequences of one's actions (Cushman, 2008). Further revealing the mind-morality link are arguments about who has (and lacks) moral standing; people have denied full moral status to animals (Bastian et al., 2012; Gray, Gray, & Wegner, 2007), children (Cameron, Lindquist, & Gray, 2015), and even other races (Haslam, 2006; Jahoda, 1999; Warren, 1997; Waytz & Schroeder, 2014) on the basis of perceived differences in mind.

Mind may be important for morality, but it is difficult to know for certain whether someone else has a mind (Chalmers, 1997). Questions of mind are often, therefore, matters of perception (Wegner & Gray, 2017), especially in the case of machines (Gray & Wegner, 2012). Research on mind perception reveals that minds are perceived along two dimensions, agency and experience (Gray et al., 2007). Agency refers to the capacity to think, to reason, to plan, and to carry out one's intentions (Gray et al., 2012), whereas experience refers to the capacity to feel emotions and sensations, including pain and fear (Gray et al., 2012). Both these dimensions may be important for making moral decisions—and for explaining a potential aversion to machine moral decision-making.

1.3.1. Agency

Agency is often seen as necessary for making moral decisions. Historically, Kant (1788) and Hume (1751) both argued that moral decisions required reason and Locke argued that people must be "active thinking beings" (Locke, 1836) in order to be allowed to make moral judgments. More modern legal scholars and philosophers also emphasize agency-related abilities in making moral judgments, including intelligence (Vanderblit, 1956), being able to choose rationally between alternatives (Clarke, 1992; Frankfurt, 1969), and understanding the consequence of actions (Mele & Sverdlik, 1996). When children and those with mental disabilities are given less blame for their moral decisions, it is because they are seen to have less agency than adults (Gray & Wegner, 2009).

Machines are often seen to have some agency (Gray & Wegner, 2012; Gray et al., 2007)—they can play chess and perform complex calculations—but their ability to think is often quite domain specific. Moreover, agency includes aspects beyond the ability to make raw calculations, including self-control, planning, communication and thought (Gray et al., 2007). In this full sense of agency, machines are perceived as having less agency than adult humans (Gray et al., 2007)—suggesting that they may seem as less able to make legitimate moral decisions. Consistent with this idea, many argue that—normatively speaking—machines need agency in order to make moral decisions (Floridi & Sanders, 2004; Hellström, 2013; Malle & Scheutz, 2014; Steinert, 2014; Wallach & Allen, 2009; Wallach, Franklin, & Allen, 2010). These agency-related abilities include interactivity, autonomy and adaptability (Floridi & Sanders, 2004), and also the ability for