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
Available online 11 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

moral reasoning, autonomous action, and communication (Malle & Scheutz, 2014; Malle, 2016). Machines perceived lack of agency, therefore, may help explain the potential aversion to machines making moral decisions.

#### 1.3.2. Experience

Discussions about moral decision-making (i.e., moral agency) often emphasize agency but seldom experience. Instead, experience is seen to be linked to questions of moral patiency—whether someone can be the recipient of good or evil and are therefore worthy of protection (Bastian et al., 2012; Gray, Jenkins, Heberlein, & Wegner, 2011; Gray et al., 2012; Haslam, 2006; Leyens et al., 2000; Rudman & Mescher, 2012; Schein & Gray, 2018; Singer, 1975; Waytz & Schroeder, 2014). For example, those who see less experience in animals are more likely to eat meat (Bastian et al., 2012) and those who see less experience in other races are more prejudiced (Leyens et al., 2000). Research also links reduced perceptions of experience to psychopathy (Gray et al., 2011), perhaps explaining why psychopaths are more willing to harm others.

Experience clearly matters for moral patiency, but it may also matter for making moral decisions. Hume argued that sentiment (i.e., experience) is also essential for making moral decisions (Hume, 1751). More recently, Damm (2010) has argued that the diminished ability to make moral decisions in autism and psychopathy is tied to deficits in emotional experience. Decades of research in psychology supports the contention that emotions are critical to moral decision-making (Greene, Sommerville, Nystrom, Darley, & Cohen, 2001; Haidt et al., 1993; Haidt, 2001; Koenigs et al., 2007; Prinz, 2007). In particular, the capacity for empathy—feeling pain on behalf of others—seems to be a core element of intact moral judgment (Aaltola, 2014; De Waal, 2010; Decety & Cowell, 2014; Kauppinen, 2017; Pinker, 2011; Rifkin, 2009; Shaw, Batson, & Todd, 1994; Zaki, 2018). It may be that laypeople intuitively appreciate this empirical link between emotional experience and moral decision-making, seeing experience as necessary for moral judgment.

Scholars have also highlighted the importance of experience for machine moral decision-making (Allen, Wallach, & Smit, 2006; Coeckelbergh, 2010; Himma, 2009; Malle & Scheutz, 2014; Malle, 2016; Wallach et al., 2010), including the ability to feel moral emotions (Malle & Scheutz, 2014) and having an "inner subjective experience like that of pain" (Himma, 2009, p. 19). If experience is indeed seen as a prerequisite of moral decision-making, this would be problematic for machine moral decision-making: although machines may be seen to have some agency, they are seen to be devoid of experience (Brink, Gray, & Wellman, 2017; Gray & Wegner, 2012). We therefore suggest that a potential aversion to machine moral decision-making likely also involves perceptions of relatively little experience.

### 1.4. The current research

In nine studies—all approved by the UNC IRB—we investigate whether people are averse to machines making moral decisions. We define this aversion as seeing moral decisions made by machines as less acceptable than those made by adult humans. We note that while there are many morally-relevant decisions faced by machines, here we examine the most paradigmatic cases of moral decisions—difficult dilemmas which directly impact human life (and death). Given the growing discussion about machines on roadways (Bonnefon et al., 2016), in the law (Angwin et al., 2016), in medicine (van den Berg et al., 2017) and in the military (Horowitz, 2016), the dilemmas we examine are in these domains. We also examine whether machines' perceived lack of mind helps to explain the potential aversion to machine moral decision-making. Although we are exploring moral decisions—because of their practical importance—we acknowledge that people may be averse to machines making decisions across many domains.

We divide the studies in this paper into three sections. Section one, "Documenting an Aversion," reveals that people would rather not have machines make life and death decisions about driving (Study 1) and parole (Study 2), and that this aversion is mediated by mind perception (Study 2). Section two, "Specifying the Outcome", reveals that the aversion to machine decision-making is not due to people assuming that machines will make worse decisions. Even when specifying the outcome—whether negative (Studies 3–5) or positive (Studies 5–6)—people think it less appropriate for machines (vs humans) to make moral decisions within the domains of medicine (Studies 3 & 6) or the military (Studies 4 & 5). Section three, "Reducing the Aversion," briefly examines three possible routes to reducing the aversion to machines making moral decisions. The aversion can be decreased by limiting machines to an advisory role (Study 7)—that is, giving humans the final decision. Increasing the perceived experience of machines (Study 8) does not reduce the aversion, but increasing the perceived expertise of machines does (Study 9), but only when this advantage in expertise is made especially salient.

In all studies, we report all conditions, data exclusions, sample size determinations, and measures.

## 2. Section 1: documenting an aversion to machine moral decision-making

We first conducted three studies to test whether people are averse to machines making moral decisions. Study 1 tested whether it is less permissible for machines (vs. humans) to make life and death driving decisions. Study 2 used a different paradigm to test whether it is less permissible for machines (vs. humans) to make parole decisions and whether this reduced permissibility is mediated by machines' perceived lack of mind.

### 2.1. Study 1: machines making life and death driving decisions

Driving often feels mundane but the lives of people frequently hang in the balance. For example, vehicle collisions are the leading cause of death of American teenagers. The rise of autonomous vehicles suggests that machines will soon be able to make decisions about human lives—already one of Uber's autonomous cars struck and killed a pedestrian. Some research has examined *how* people want autonomous vehicles to make moral decisions (Bonnefon et al., 2016), but here we examine *whether* people want these machines to make these decisions in the first place.

#### 2.1.1. Method

*2.1.1.1. Preregistration*. The study was pre-registered at https://aspredicted.org/dg88r.pdf.

*2.1.1.2. Participants*. Here, and in studies 2–6 and 8–9, we assumed a medium effect size (Cohen's $d = 0.5$). We found in a power analysis that we need 105 participants per condition to obtain a power of 0.95. We aimed for 120 participants per condition in order to account for participants who might fail the comprehension questions (Goodman, Cryder, & Cheema, 2013; Hauser & Schwarz, 2016).[2] Two hundred and forty-two participants from the United States and Canada (53.3% female; age: $M = 34.36$, $SD = 10.98$) completed the questionnaire on Amazon's Mechanical Turk for 20 cents. As specified in preregistration, participants were excluded if they failed to correctly answer the comprehension question ("Who will be the one to make life and

[2] Our sample size for Study 7 was based a study that was omitted from the paper in the review process. In that study, as in Study 7, we measured people's choice of decision maker. Based on a power analysis, we aimed for a sample of 100 participants per condition. For the sake of consistency, in Study 7, which is descriptive and also measures choice of decision maker, we aimed for 100 participants as well.

death decisions in the car you read about?"), which led to the exclusion of twenty-three participants.

#### 2.1.1.3. Procedure

*2.1.1.3.1. Descriptions*. In this between-subjects design, participants were randomly assigned to one of two conditions. In both conditions, participants read that:

> "Driving sometimes involves decisions of life and death. These decisions can affect people in the car, as well as pedestrians and people in other cars."

In the "human" condition participants read that "For a new car model, a human driver will be the one making these decisions." In the "machine" condition participants read that "For a new self-driving car model, an autonomous computer program will be the one making these decisions".

*2.1.1.3.2. Assessing permissibility*. Using a 5-point scale from 1 (strongly disagree) to 5 (strongly agree), participants rated whether "It is appropriate for a human driver/an autonomous computer program to make these decisions", "A human driver/an autonomous computer program should be the one to make these decisions" and "A human driver/an autonomous computer program should be forbidden from making these decisions" (last item is reverse scored, Cronbach's $\alpha = 0.93$). These items were used to assess permissibility in all subsequent studies that examined permissibility. Participants then answered a comprehension question and provided demographic information.

### 2.1.2. Results

Consistent with an aversion to machines making moral decisions, an independent samples *t*-test revealed that participants rated it less permissible for life and death driving decisions to be made by an autonomous computer program ($M = 2.80$, $SD = 1.13$) than a human driver ($M = 4.33$, $SD = 0.74$), $t(238) = 12.46$, $p < .001$, Cohen's $d = 1.60$. See Fig. 1.

### 2.1.3. Discussion

This study revealed preliminary evidence that people are averse to having machines make moral decisions. The next study tested whether this aversion would be observed in a different paradigm within a different moral domain—parole decisions.

## 2.2. Study 2: machines making parole decisions

Deciding whether to grant an offender parole is an important moral decision, not only determining the fate of offenders, but also involving questions of retribution, restitution, and rehabilitation. Traditionally, these moral decisions are made by a board of experts (Johnson, 1973), but some states use machines to assist with parole decisions (Kehl, Guo, & Kessler, 2017)—generating substantial controversial (Angwin et al., 2016).

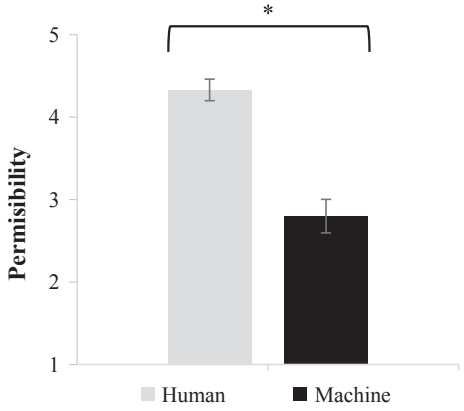


**Fig. 1.** Permissibility of human and machine as decision makers in driving life and death decisions (Study 1). Error bars reflect 95% confidence intervals. $^{*}p < .05$.

In this study, we assess whether people are averse to the idea of machines making parole decisions. We also test whether any potential aversion of machine moral decisions is mediated by reduced perceptions of mind in machines.

### 2.2.1. Method

*2.2.1.1. Preregistration*. The study was pre-registered at https://aspredicted.org/4yh36.pdf.

*2.2.1.2. Participants*. Two hundred and forty-one participants from the United States and Canada (62.7% female; age: $M = 35.89$, $SD = 11.59$) completed the questionnaire on Amazon's Mechanical Turk for 30 cents. As specified in preregistration, participants were excluded if they failed to correctly answer the comprehension question ("did you read about a computer or a committee?"), leading to the exclusion of one participant.

*2.2.1.3. Procedure*. Participants were randomly assigned to one of two conditions, reading that parole decisions were made by either an advanced machine or a panel of humans. They then rated the perceived mind of the agent (machine or human) and the permissibility of that agent making these decisions.

*2.2.1.3.1. Descriptions*. In the machine condition participants read this brief description accompanied by a picture of a supercomputer (see Fig. 2):

> "This is CompNet. CompNet is a super computer used by various government agencies for calculations, estimates, and decision-making."

In the human committee condition participants read this brief description accompanied by a picture of a committee of humans (see Fig. 2).

> "This is a picture of a state committee. The state committee consists of legal and mental health experts as well as representatives of the community."

Participants then read this brief description about how parole decisions are made (Caplan, 2007; Dawson, 1966; Johnson, 1973):

> "Parole decisions—about whether convicted criminals will be released from jail–involve many factors, including the convict's level of remorse, criminal history, rehabilitation efforts and likelihood of committing crimes. Also important is the emotional testimony of the victims."

*2.2.1.3.2. Assessing mind*. Participants rated the machine or the human on twelve different mental capacities ("To what extent do you think CompNet/the committee members can…")[3]: six agency-related, "communicate with others," "is able of thinking," "plans his actions," "is intelligent," "has foresight" and "is able to think things through" ($\alpha = 0.90$), and six experience-related, "sensitive to pain", "experience happiness", "experience fear", "experience compassion", "experience empathy" and "experience guilt" ($\alpha = 0.98$). All ratings were made on a 5 point scale from 1 (Not at all) to 5 (Extremely).

Participants then rated the permissibility of the machine or state committee to make these decisions (Cronbach's $\alpha = 0.93$), answered

[3] We designed this study later than the studies that appear after it, but it fits better here in terms of logical flow. One of our goals in this study was to use a more comprehensive measure of mind—and we obtained results consistent with Studies 3–6, which used a more concise measure.

**Fig. 2.** Images of CompNet, the super computer (left) and the state committee (right; Study 2).

comprehension questions and provided demographic information.

#### 2.2.2. Results

*2.2.2.1. Aversion to machine making moral decisions.* Consistent with an aversion to machines making moral decisions, an independent samples *t*-test revealed that participants rated it less permissible for CompNet to make parole decisions ($M = 1.80$, $SD = 1.06$) than the human committee ($M = 3.42$, $SD = 0.95$), $t(238) = 12.37$, $p < .001$, Cohen's $d = 1.61$.

*2.2.2.2. Assessing mind.* To examine potential differences in mind perception, we conducted a 2 × 2 mixed model ANOVA with decider (CompNet, human committee) as a between-subject factor and dimension of mind (agency, experience) as a within-subject factor. The ANOVA revealed a main effect for both decider, $F(1, 238) = 440.37$, $p < .001$, partial $\eta^2 = 0.65$, and dimension, $F(1, 238) = 199.50$, $p < .001$, partial $\eta^2 = 0.46$, such that the human committee was seen to have more overall mind ($M = 3.79$, $SD = 0.64$) than CompNet ($M = 1.96$, $SD = 0.71$), and that more agency ($M = 3.30$, $SD = 1.11$) was attributed overall than experience ($M = 2.45$, $SD = 1.42$). However, these were qualified by the predicted decider × dimension interaction, $F(1, 238) = 124.94$, $p < .001$, partial $\eta^2 = 0.34$. Although the human committee was perceived to have more agency ($M = 3.88$. $SD = 0.69$) than CompNet ($M = 2.73$, $SD = 1.14$), $p < .001$, the human committee was seen to have substantially more experience ($M = 3.70$, $SD = 0.71$) than CompNet ($M = 1.20$, $SD = 0.63$), $p < .001$.

The very low rating of CompNet's experience (1.20 on a 1 to 5 scale), though significantly different than 1, $p = .001$, suggests that participant naturally see machines as lacking in emotional experience and compassion—providing validation for the more explicit descriptions used in some future studies.

*2.2.2.3. Mediation of aversion with mind.* Can mind perception help explain the aversion to machine moral decision-making? A bootstrapping mediation analysis tested whether mind perception mediated the effect of decider on permissibility (Preacher & Hays, 2008; 5000 iterations, model 4). As noted earlier, decider (coding: CompNet, 1; human committee, −1) was negatively associated with permissibility, $b = -0.80$, $SE = 0.07$, $p < .001$. Additionally, CompNet was perceived to have less agency, $b = -0.57$, $SE = 0.06$, $p < .001$, and less experience, $b = -1.25$, $SE = 0.04$, $p < .001$, than the human committee. Analyses revealed that both agency, $b = -0.39$, $SE = 0.13$, $CI_{.95}[-65, -0.13]$ and experience, $b = -0.09$, $SE = 0.04$ $CI_{.95}[-0.18, -0.02]$ had significant indirect effects that mediated the link between decider and permissibility. When these two mediators were included in the regression, the effect of decider on permissibility remained significant, $b = -0.32$, $SE = 0.13$, $p = .016$, $CI_{.95}[-0.58, -0.06]$. See Fig. 3.

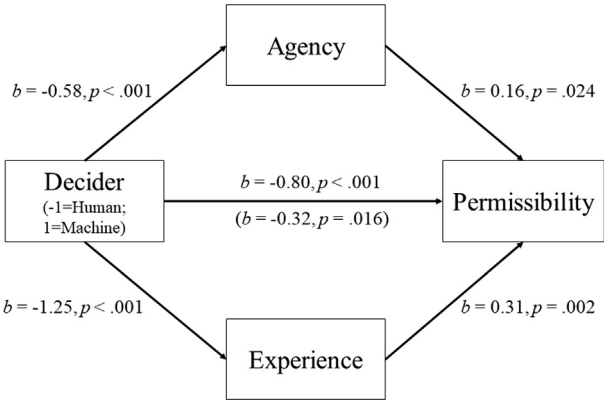


**Fig. 3.** Mediation analysis revealed that mind perception (both agency and experience) mediates the aversion to machines making parole decisions (Study 2). Both indirect effects are significant.

#### 2.2.3. Discussion

These results replicate those of Study 1, further revealing the aversion to machine moral decision-making. Importantly, this study found similar effects within a different domain of moral decision (parole decisions) and by using a different presentation of human versus machine agents. Mediation analyses revealed that, as predicted, the aversion to machine moral decision-making is mediated by mind perception.

Adding to the results of Studies 1 and 2, the supplementary materials reports an additional study in which people often choose a machine decision-maker over a human decision-maker in a medical context, despite the cost-saving benefits of choosing a machine. The results of this study are somewhat ambiguous, and the study is not included in the main paper. However, it is included in the supplementary materials for the interested readers—and to guard against the "file drawer problem"

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.cognition.2018.08.003.

One limitation of the studies conducted so far is that participants rated the permissibility of machines making moral decisions without an actual decision. It is therefore possible that the aversion found in these studies stems from concerns that machines would make decisions with worse outcomes, as compared to humans. We address this concern in the next section of studies by explicitly specifying the outcome of decisions.

## 3. Section 2: specifying the outcomes of moral decisions

Our studies have so far revealed an aversion to machines making moral decisions, which can be partially explained by perceptions of mind. However, one question is whether people are averse to machine moral decision-making because they assume that machines will make

different decisions than humans. We acknowledge this is a likely possibility (Bonnefon et al., 2016), but suggest that the aversion is robust to the outcome, such that people will be averse to machine moral decision-making even if the outcome is known—an idea we test here.

In this series of experiments (Studies 3–6), we assess reactions to moral decisions in medicine (surgery) and the military (drone strikes). Because they are the most practically consequential, we first examine negative outcomes (Studies 3 and 4) in which moral decisions result in the death of humans. Not only do negative outcomes create the most public outcry (Soroka, 2006), they are more likely to engage psychological processes related to blame (Malle, Guglielmo, & Monroe, 2014; Schein & Gray, 2018) and so provide the most likely case for revealing the aversion to machine moral decision-making. However, as a more conservative test, we also examine positive outcomes in which humans are not harmed by moral decisions (Studies 5–6). Even here—when machines make the "right" decision—we predict that people will still be averse to machine moral decision-making.

### 3.1. Study 3: machines making a medical decision with a bad outcome

Medical decisions—such as whether to perform a risky surgery—are morally laden, as they involve potential harm. This study examined whether people are averse to machines making life-and-death medical decisions, even when the outcome of that decisions is specified. Would people be more averse to a machine (versus a human doctor) that decides to perform a risky—and ultimately failed—surgery in which the patient dies?

#### 3.1.1. Method

*3.1.1.1. Preregistration.* The study was pre-registered at https://aspredicted.org/37hb9.pdf.

*3.1.1.2. Participants.* Two hundred and forty participants from the United States and Canada (56.7% female; age: $M = 38.48$, $SD = 13.14$) completed the study on Amazon's Mechanical Turk for 30 cents. As specified in preregistration, participants were excluded if they failed to correctly answer the comprehension questions ("who made the decision whether or not to perform the surgery?" and "what was the outcome of the surgery?"), leading to the exclusion of fourteen participants.

*3.1.1.3. Procedure.* Participants were randomly assigned to one of two conditions, reading that a medical scenario was decided upon by either an advanced machine or a human doctor—a decision that resulted in the patient's death.

*3.1.1.3.1. Descriptions.* All participants read the same opening paragraph:

> "Jason is a child who was just hit by a car. He is in stable condition at the hospital, but his spinal cord is damaged, and he will likely be permanently paralyzed. There is a new surgery that can fix his spinal cord, but it has a 5% chance of killing the patient. The surgery is time-sensitive and Jason's parents cannot be reached to make a decision"

In the machine condition participants read that a machine decided about the surgery and that it was a failure:

> "HealthComp is charged with making the decision. HealthComp is an autonomous statistics-based computer system with a great capacity for rational thinking, but totally lacking in emotional compassion. The computer system decides to perform the surgery. The surgery is a failure and Jason dies."

The human condition was similar but described Dr. Jones, a human doctor, making the decision: "Dr. Jones is charged with making the decision. Dr. Jones is a doctor with a great capacity for both rational thinking and for emotional compassion. Dr. Jones decides to perform the surgery. The surgery is a failure and Jason dies."

Participants then rated the permissibility of the machine or human doctor to make these decisions (Cronbach's $\alpha = 0.90$), the perceived mind of the agent (see "assessing mind" below), answered comprehension questions and provided demographic information.[4]

*3.1.1.3.2. Assessing mind.* Participants rated the machine or the human on six different mental capacities ("To what extent do you think HealthComp/Dr. Jones can…"): three agency-related, "communicate with others," "is able of thinking," "plans his actions," ($\alpha = 0.84$), and three experience-related, "sensitive to pain", "experience happiness", "experience fear" ($\alpha = 0.96$). All ratings were made on a 5-point scale from 1 (Not at all) to 5 (Extremely).

#### 3.1.2. Results

*3.1.2.1. Aversion to machine making moral decisions.* Consistent with an aversion to machines making moral decisions, an independent samples *t*-test revealed that participants rated it as less permissible for HealthComp ($M = 1.91$, $SD = 1.04$) than for Dr. Jones ($M = 2.81$, $SD = 1.24$) to make a medical decision that resulted in the death of a patient, $t(224) = 5.88$, $p < .001$, Cohen's $d = 0.79$.

*3.1.2.2. Assessing mind.* To examine potential differences in mind perception, we conducted a 2 × 2 mixed model ANOVA with decider (HealthComp, Dr. Jones) as a between-subject factor and dimension of mind (agency, experience) as a within-subject factor. The ANOVA revealed a main effect for both decider, $F(1, 224) = 677.25$, $p < .001$, partial $\eta^2 = 0.75$, and dimension, $F(1, 224) = 121.60$, $p < .001$, partial $\eta^2 = 0.35$, such that Dr. Jones ($M = 4.03$, $SD = 0.76$) was perceived as having more overall mind than HealthComp ($M = 1.72$, $SD = 0.55$), and that more agency ($M = 3.20$, $SD = 1.27$) was attributed overall than experience ($M = 2.54$, $SD = 1.57$). However, these were qualified by the predicted significant decider × dimension interaction, $F(1, 224) = 74.69$, $p < .001$, partial $\eta^2 = 0.25$. Although Dr. Jones was perceived to have more agency ($M = 4.10$, $SD = 0.84$) than HealthComp ($M = 2.31$, $SD = 0.99$), $p < .001$, he was seen as having substantially more experience ($M = 3.96$, $SD = 0.86$) than HealthComp ($M = 1.12$, $SD = 0.41$), $p < .001$.

*3.1.2.3. Mediation of aversion with mind.* Can mind perception help explain the aversion to machine moral decision-making? A bootstrapping mediation analysis tested whether mind perception mediated the effect of decider on permissibility (Preacher & Hays, 2008; 5000 iterations, model 4). As noted earlier, decider (coding: HealthComp, 1; Dr. Jones, −1) was negatively associated with permissibility, $b = -0.45$, $SE = 0.08$, $p < .001$. Additionally, HealthComp was perceived to have less agency, $b = -0.90$, $SE = 0.06$, $p < .001$, and less experience, $b = -1.42$, $SE = 0.05$, $p < .001$ than Dr. Jones. Analyses revealed that both agency, $b = -0.23$, $SE = 0.08$, $CI_{.95}[-0.40, -0.09]$ and experience, $b = -0.37$, $SE = 0.17$, $CI_{.95}[-0.68, -0.03]$ were significant indirect effects that mediated the link between decider and permissibility. When these mediators were included in the regression, the effect of decider on permissibility was no longer significant, $b = 0.16$, $SE = 0.17$, $p = .361$, $CI_{.95}[-0.18, 0.49]$. See Fig. 4.

#### 3.1.3. Discussion

These results again suggest that people are averse to machines making moral decisions, even when they make the same decision—with the same outcome—as a human agent. These results also replicated the

[4] In Studies 3–5 participants also provided additional judgments of the decider and the situation, including how much compensation the victim's family deserves. These measures and results are detailed in the supplementary materials. We had planned to use these results as pilot data for a separate manuscript, however, the effects were less clear than we hoped.

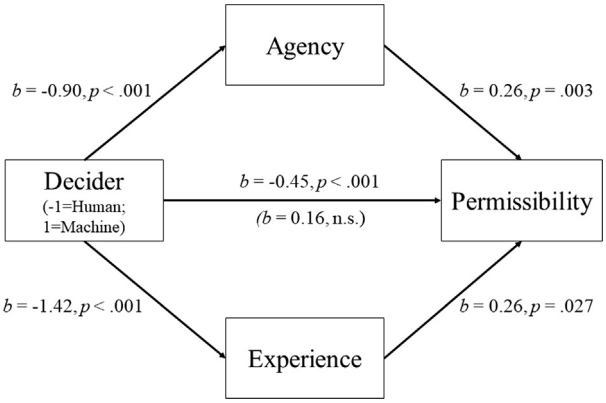


**Fig. 4.** Mediation analysis revealed that mind perception (both agency and experience) mediates the aversion to machines making a medical decision (Study 3). Both indirect effects are significant.

mediation pattern from Study 2, in which reduced perceptions of both agency and experience mediated the aversion to machine moral decision-making.

### 3.2. Study 4: military drones

This study sought to replicate the findings of Study 3 in a different domain—military drones. Participants read about a paradigmatic drone dilemma in which a missile strike could kill a terrorist but also harm innocent civilians (Arkin, 2008; Horowitz, 2016; Johnson & Axinn, 2013). We examined whether people would be averse to a machine making this decision when it resulted in failure, and whether this aversion is mediated by mind perception.

#### 3.2.1. Method

*3.2.1.1. Preregistration.* The study was pre-registered at https://aspredicted.org/qt53u.pdf.

*3.2.1.2. Participants.* Two hundred and forty-two participants from the United States and Canada (57.4% female; age: $M = 35.83$, $SD = 11.46$) completed the questionnaire on Amazon's Mechanical Turk for 30 cents. As specified in preregistration, participants were excluded if they failed to correctly answer the comprehension questions ("who made the decision in the story you read?" and "what was the outcome of the missile strike in the story you read?"), leading to the exclusion of five participants.

*3.2.1.3. Procedure.* Participants were randomly assigned to one of two conditions, and rated the permissibility of either a machine or human making a decision that lead to a failed drone strike.

*3.2.1.3.1. Descriptions.* In the machine condition participants read the following scenario:

> "A US military drone flying over Baghdad detects a man who matches the description of a known terrorist. He is hiding out in a crowded suburb. The drone feeds the information back to CompNet. CompNet is an autonomous statistics-based computer system with a great capacity for rational thinking but is totally lacking in emotional compassion. It makes the decision to fire a missile at the terrorist's location, despite the possibility of collateral damage.
> The missile strike is a failure. It does not kill the terrorist but does kill 4 nearby children who are on their way to school."

The human condition was similar but described Colonel Jones, a human officer, making the decision: "Colonel Jones is an officer with a great capacity for both rational thinking and for emotional compassion."

Participants then rated the permissibility of the machine or human doctor to make such decisions (Cronbach's $\alpha = 0.90$), the perceived mind of the agent (assessed in the same way as in Study 3, Cronbach's $\alpha$: agency = 0.84; experience = 0.92), answered the comprehension questions and provided demographic information.

#### 3.2.2. Results

*3.2.2.1. Aversion to machine making moral decisions.* Consistent with an aversion to machines making moral decisions, an independent samples $t$-test revealed that participants rated it as less permissible for CompNet ($M = 1.69$, $SD = 0.90$) than for Colonel Jones to make the decision ($M = 2.97$, $SD = 1.26$), $t(234) = 9.02$, $p < .001$, Cohen's $d = 1.26$.

*3.2.2.2. Assessing mind.* To examine potential differences in mind perception, we conducted a 2 × 2 mixed model ANOVA with decider (CompNet, Colonel Jones) as a between-subject factor and dimension of mind (agency, experience) as a within-subject factor. The ANOVA revealed a main effect for both decider, $F(1, 234) = 252.62$, $p < .001$, partial $\eta^2 = 0.52$, and dimension, $F(1, 234) = 164.67$, $p < .001$, partial $\eta^2 = 0.41$, such that Colonel Jones ($M = 3.45$, $SD = 0.88$) was perceived as having more overall mind than CompNet ($M = 1.90$, $SD = 0.60$), and that more agency ($M = 3.11$, $SD = 1.18$) was attributed overall than experience ($M = 2.19$, $SD = 1.30$). However, these were qualified by the predicted significant decider × dimension interaction, $F(1, 234) = 65.65$, $p < .001$, partial $\eta^2 = 0.22$. Although Colonel Jones was perceived to have more agency ($M = 3.62$. $SD = 1.02$) than CompNet ($M = 2.63$, $SD = 1.11$), $p < .001$, he was seen as having substantially more experience ($M = 3.28$, $SD = 0.96$) than CompNet ($M = 1.16$, $SD = 0.48$), $p < .001$.

*3.2.2.3. Mediation of aversion with mind.* Can mind perception help explain the aversion to machine moral decision-making? A bootstrapping mediation analysis tested whether mind perception mediated the effect of decider on permissibility (Preacher & Hays, 2008; 5000 iterations, model 4). As noted earlier, decider (coding: CompNet, 1; Colonel Jones, −1) was negatively associated with permissibility, $b = -0.64$, $SE = 0.07$, $p < .001$. Additionally, CompNet was perceived as having less agency, $b = -0.49$, $SE = 0.07$, $p < .001$, and less experience, $b = -1.05$, $SE = 0.05$, $p < .001$, than Colonel Jones. Analyses revealed that both agency, $b = -0.10$, $SE = 0.04$, $CI_{95}[-0.18, -0.04]$, and experience, $b = -0.43$, $SE = 0.11$ $CI_{95}[-0.64, -0.21]$, were significant indirect effects that mediated the link between decider and permissibility. When these mediators were included in the regression, the effect of decider on permissibility was no longer significant $b = -0.11$, $SE = 0.11$, $p = .349$, $CI_{95}[-0.33, 0.12]$. See Fig. 5.

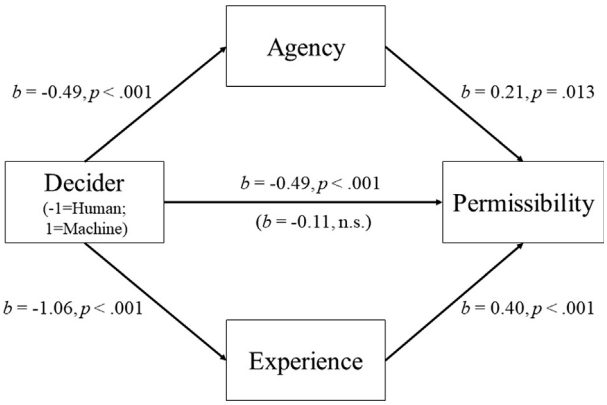


**Fig. 5.** Mediation analysis for the effect of the agent on permissibility by mind perception (agency and experience) for the military decision (Study 4). Both indirect effects are significant.

#### 3.2.3. Discussion

These results again reveal an aversion to machine moral decision-making, within both another domain and controlling for outcome. One concern is again whether the results hinge on explicitly mentioning differences in mind between the human decision maker and the machine. We address this concern in a replication.

### 3.3. Study 4 Replication: no explicit mention of compassion

This study was identical to Study 4 except it did not explicitly mention that the human decision maker Colonel Jones had a "great capacity for both rational thinking and for emotional compassion" whereas CompNet was "totally lacking in emotional compassion." Although these differences reflect naturalistic differences in mind perception (see Study 2)—and help confer experimental control—we want to make sure the effects are robust to mentioning these differences. Participants (N = 243) were recruited from MTurk (59.7% female; age: $M = 36.25$, $SD = 11.37$), with 21 excluded for failing the comprehension questions.

As in Study 4, participants rated it as less permissible for CompNet to make the moral decision ($M = 1.65$, $SD = 0.92$) than Colonel Jones ($M = 2.77$, $SD = 1.24$), $t(217) = 7.62$, $p < .001$, Cohen's $d = 1.03$. Compared to Colonel Jones, participants also saw CompNet as having less agency ($M = 3.53$ vs. $M = 2.39$) and—especially—less experience ($M = 3.05$ vs. $M = 1.08$), $ps < .001$. See supplementary materials for all analyses. As in Study 2, these results reveal that people naturalistically see machines as completely lacking in experience.

The mediation analysis revealed that both perceived agency, $b = -0.21$, $SE = 0.05$, $CI_{.95}[-0.33, -0.13]$ and experience, $b = -0.27$, $SE = 0.11$ $CI_{.95}[-0.48, -0.06]$, mediated the aversion to machine moral decision-making. When these two mediators were included in the regression, the effect of decider on permissibility was no longer significant, $b = -0.08$, $SE = 0.10$, $p = .447$, $CI_{.95}[-0.28, 0.13]$.

These results replicate those of Study 4 and reveal that the aversion to machine moral decision-making (and its mediation by mind perception) do not hinge upon the explicit descriptions used in Studies 3 and 4—descriptions which reflect participants' naturalistic views about the minds of machines. In the next study, we tested whether the aversion to machine moral decision-making remains with positive outcomes.

### 3.4. Study 5: both good and bad outcomes

Negative outcomes are most likely to induce blame and maybe especially likely to turn opinions against machines (Dietvorst, Simmons, & Massey, 2015). Here we test whether the aversion to machine moral decision-making changes when the outcome is positive—a missile strike is successful and does not kill civilians.

#### 3.4.1. Method

*3.4.1.1. Preregistration.* The study was pre-registered at https://aspredicted.org/g4i2q.pdf.

*3.4.1.2. Participants.* Four hundred and eighty-five participants from the United States and Canada (60.4% female; age: $M = 37.27$, $SD = 11.63$) completed the questionnaire on Amazon's Mechanical Turk for 30 cents. As specified in preregistration, participants were excluded if they failed to correctly answer the comprehension questions ("who made the decision in the story you read?" and "what was the outcome of the missile strike in the story you read?"), leading to the exclusion of twenty-six participants.

*3.4.1.3. Procedure.* The procedure was identical to Study 4 except rather than 2 conditions (decider: CompNet, Colonel Jones) this study used a 2 (decider: CompNet, Colonel Jones) × 2 (outcome: negative, positive) design. In the negative outcome condition (identical to Study 4), participants read that "The missile strike is a failure. It does not kill the terrorist but does kill 4 nearby children who are on their way to school." In the positive outcome condition, participants read that "The missile strike is successful. It kills the terrorist, and causes only minor injuries to a few civilians who are standing nearby."

After reading the scenario participants rated the permissibility of the decider in making the decision (Cronbach's $\alpha = 0.92$) and the perceived mind perception of the decider (agency, $\alpha = 0.82$, experience, $\alpha = 0.94$) in the same scales used in Studies 3 and 4, and a few exploratory items as specified in the preregistration. Participants then answered the comprehension questions and provided demographic information.

#### 3.4.2. Results

*3.4.2.1. Aversion to machines making moral decisions.* Consistent with an aversion to machines making moral decisions, a 2 (decider: human, machine) × 2 (outcome: positive, negative) between-subject ANOVA of permissibility ratings revealed a main effect for decider, $F(1, 455) = 228.21$, $p < .001$, partial $\eta^2 = 0.39$, such that across outcomes, people rated CompNet ($M = 1.90$, $SD = 0.99$) as less permissible than Colonel Jones ($M = 3.53$, $SD = 1.17$) in making the decision. In addition, we found a main effect for outcome, $F(1, 455) = 53.55$, $p < .001$, partial $\eta^2 = 0.11$, such that overall deciders were seen as more permissible when the outcome was positive ($M = 3.08$, $SD = 1.36$) than negative ($M = 2.37$, $SD = 1.26$). The decider × outcome interaction was not significant, $F(1, 455) = 2.54$, $p = .112$, revealing that the aversion to machine moral decision-making is not restricted to negative outcomes. See Fig. 6.

*3.4.2.2. Assessing mind.* To examine potential differences in mind perception, we conducted we conducted a 2 × 2 × 2 mixed model ANOVA with decider (CompNet, Colonel Jones) and outcome (positive, negative) as between-subject factors and dimension of mind (agency, experience) as a within-subject factor. The ANOVA revealed a main effect for decider, $F(1, 455) = 931.65$, $p < .001$, partial $\eta^2 = 0.67$, dimension, $F(1, 455) = 558.61.67$, $p < .001$, partial $\eta^2 = 0.55$, and outcome, $F(1, 455) = 28.71$, $p < .001$, $\eta^2 = 0.06$, such Colonel Jones ($M = 3.65$, $SD = 0.53$) was seen as having more mind than CompNet ($M = 1.80$, $SD = 0$ .79), that more agency ($M = 3.25$, $SD = 1.19$) was attributed overall than experience ($M = 2.21$, $SD = 1.31$), and that when the outcome was positive ($M = 2.89$, $SD = 1.17$) more mind was attributed overall than when the outcome was negative ($M = 2.56$, $SD = 1.10$).

However, these were qualified by the predicted significant decider × dimension interaction, $F(1, 238) = 124.94$, $p < .001$, partial

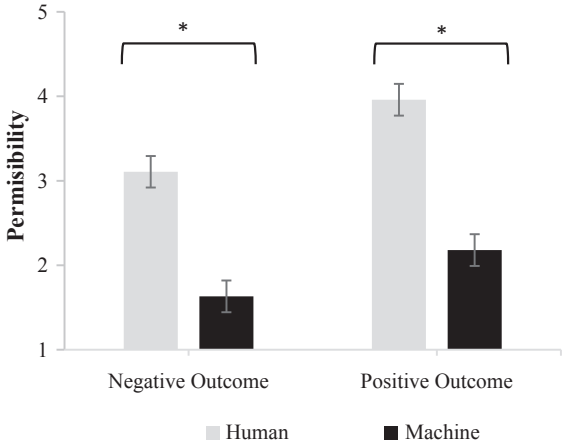


**Fig. 6.** Permissibility of human and machine deciders for negative and positive outcomes (Study 5). Error bars reflect 95% confidence intervals. $^{*}p < .05$.

$\eta^2 = 0.34$. Although Colonel Jones was perceived to have more agency ($M = 3.99$. $SD = 0.89$) than CompNet ($M = 2.50$, $SD = 0.95$), $p < .001$, he was seen to have substantially more experience ($M = 3.30$, $SD = 0.91$) than the machine ($M = 1.10$, $SD = 0.22$), $p < .001$. See supplementary materials for the full analysis.

*3.4.2.3. Mediation of aversion with mind.* Can mind perception help explain the aversion to machine moral decision-making? Since the difference in permissibility of Colonel Jones and CompNet did not vary across outcome conditions, we tested this while collapsing across outcomes. A boot-strapping mediation analysis tested whether mind perception mediated the effect of decider on permissibility (Preacher & Hays, 2008; 5000 iterations, model 4). As noted earlier, decider (coding: CompNet, 1; Colonel Jones, −1) was negatively associated with permissibility, $b = -0.81$, $SE = 0.05$, $p < .001$. Additionally, CompNet was perceived to have less agency, $b = -0.75$, $SE = 0.04$, $p < .001$, and less experience, $b = -1.10$, $SE = 0.03$, $p < .001$, than Colonel Jones. Analyses revealed that both agency, $b = -0.25$, $SE = 0.05$, $CI._{95}[-0.34, -0.16]$, and experience, $b = -0.38$, $SE = 0.07$ $CI._{95}[-0.53, -0.23]$, had significant indirect effects that mediated the link between decider and permissibility. When these two mediators were included in the regression, the effect of decider on permissibility remained significant, $b = -0.19$, $SE = 0.09$, $p = .028$, $CI._{95}[-0.36, -0.02]$.

### 3.4.3. Discussion

These results suggest that the aversion to machine moral decision-making does not require unknown (Studies 1 and 2) or negative outcomes (Studies 3 and 4). Instead, people are averse to machines making moral decisions even when they result in generally positive outcomes. However, we note that even though the death of a terrorist is more positive than the death of four innocent children, the "positive outcome" used here still involved minor injuries to civilians and therefore might not be perceived as really positive. In Study 6 we address this concern.

## 3.5. Study 6: good outcome in a medical decision

This study examines whether people are averse to machines making medical decisions that result in good outcomes. We used the medical scenario from Study 3 (a child at risk of paralysis dies in surgery) with two important changes. First, the surgery was described as having a positive outcome–the child not only lives but regains control of his body, an unequivocal positive outcome. Second, we mentioned neither the doctor's nor machine's (in)ability to experience emotions.

### 3.5.1. Method

*3.5.1.1. Preregistration.* The study was pre-registered at https://aspredicted.org/c5df7.pdf.

*3.5.1.2. Participants.* Two hundred and thirty nine participants from the United States and Canada (50.6% female; age: $M = 35.27$, $SD = 11.01$) completed the study on Amazon's Mechanical Turk for 30 cents. As in Study 3 and as specified in preregistration, participants were excluded if they failed to correctly answer the comprehension questions ("who made the decision whether or not to perform the surgery?" and "what was the outcome of the surgery?"), leading to the exclusion of twenty five participants.

*3.5.1.3. Procedure.* The procedure was similar to that of Study 3. Participants read about the same medical dilemma as in Study 3 about whether or not to perform a risky surgery that can save a child from paralysis. Participants were randomly assigned to one of two conditions. In the machine condition participants read that:

"HealthComp is charged with making the decision. HealthComp is an autonomous statistics-based computer system. HealthComp decides to perform the surgery."

In the human condition participants read that:

"Dr. Jones is charged with making the decision. Dr. Jones decides to perform the surgery."

In both conditions participants then read that "The surgery is a success. Jason lives and regains control over his body." After reading the scenario participants rated the permissibility of the decider in making the decision (Cronbach's $\alpha = 0.88$) and the perceived mind perception of the decider (agency, $\alpha = 0.89$, experience, $\alpha = 0.96$) in the same scales used in Studies 3–5. Participants then answered the comprehension questions and provided demographic information.

### 3.5.2. Results

*3.5.2.1. Aversion to machine making moral decisions.* Consistent with an aversion to machines making moral decisions, an independent samples *t*-test revealed that participants rated it as less permissible for HealthComp ($M = 2.71$, $SD = 1.20$) than for Dr. Jones ($M = 3.43$, $SD = 1.11$) to make a medical decision that had a positive outcome, $t(212) = 4.57$, $p < .001$, Cohen's $d = 0.62$.

*3.5.2.2. Assessing mind.* To examine potential differences in mind perception, we conducted a 2 × 2 mixed model ANOVA with decider (HealthComp, Dr. Jones) as a between-subject factor and dimension of mind (agency, experience) as a within-subject factor. The ANOVA revealed a main effect for both decider, $F(1, 212) = 576.79$, $p < .001$, partial $\eta^2 = 0.73$, and dimension, $F(1, 212) = 165.52$, $p < .001$, partial $\eta^2 = 0.44$, such that Dr. Jones ($M = 4.20$, $SD = 0.62$) was perceived as having more overall mind than HealthComp ($M = 1.88$, $SD = 0.78$), and that more agency ($M = 3.46$, $SD = 1.34$) was attributed overall than experience ($M = 2.68$, $SD = 1.55$). However, these were qualified by the predicted significant decider × dimension interaction, $F(1, 212) = 56.37$, $p < .001$, partial $\eta^2 = 0.21$. Although Dr. Jones was perceived to have more agency ($M = 4.36$. $SD = 0.66$) than HealthComp ($M = 2.51$, $SD = 1.21$), $p < .001$, he was seen as having substantially more experience ($M = 4.03$, $SD = 0.76$) than HealthComp ($M = 1.26$, $SD = 0.58$), $p < .001$.

*3.5.2.3. Mediation of aversion with mind.* Can mind perception help explain the aversion to machine moral decision-making? A boot-strapping mediation analysis tested whether mind perception mediated the effect of decider on permissibility (Preacher & Hays, 2008; 5000 iterations, model 4). As noted earlier, decider (coding: HealthComp, 1; Dr. Jones, −1) was negatively associated with permissibility, $b = -0.36$, $SE = 0.08$, $p < .001$. Additionally, HealthComp was perceived to have less agency, $b = -0.93$, $SE = 0.07$, $p < .001$, and less experience, $b = -1.39$, $SE = 0.05$, $p < .001$ than Dr. Jones. Analyses revealed that agency, $b = -0.33$, $SE = 0.08$, $CI._{95}[-0.49, -0.17]$, but not experience, $b = -0.18$, $SE = 0.15$, $CI._{95}[-0.47, 0.12]$, had a significant indirect effect that mediated the link between decider and permissibility. When these mediators were included in the regression, the effect of decider on permissibility was no longer significant, $b = 0.15$, $SE = 0.17$, $p = .383$, $CI._{95}[-0.19, 0.49]$. See Fig. 7.

### 3.5.3. Discussion

These results replicate those of the "positive outcome" conditions in Study 5, and support the idea that people are averse to machines making moral decisions even when the outcome is positive. In addition, these results join those of "Study 4: replication" and further demonstrate that the aversion to machine moral decision-making (and its mediation by mind perception) do not hinge upon the explicit descriptions of machine versus human mind that were used in Studies 3–5.

Together, the results of Studies 5 and 6 suggest that the aversion to

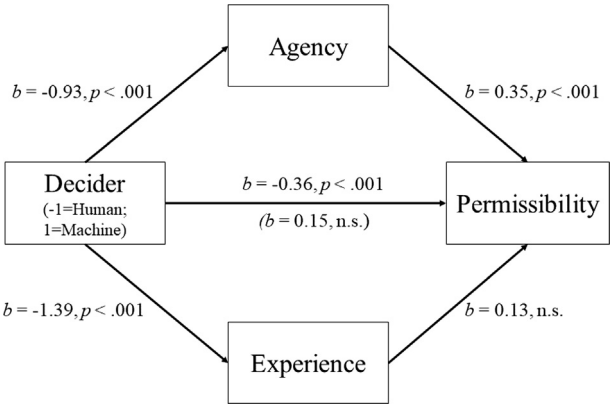


**Fig. 7.** Mediation analysis revealed that mind perception mediates the aversion to machines making a medical decision with a positive outcome (Study 6). The indirect effect of agency, but not experience, was significant.

machine moral decision-making does not require unknown outcomes (Studies 1–2) or negative outcomes (Studies 3 and 4). Instead, people are averse to machines making moral decisions even when they result in generally positive outcomes. In the next section, we try to reduce this aversion.

## 4. Section 3: reducing the aversion

The studies of sections one and two reveal that people are averse to machines making moral decisions, whether or not the outcomes of those decisions are known. This section examines possible ways to decrease this aversion: limiting machines to an advisory role (Study 7), increasing machines' perceived experience (Study 8), and finally increasing machines' perceived expertise (Study 9). Studies 7 and 8 are relatively inconclusive and should be interpreted with caution, but are reported for full transparency and to guard against the "file drawer," consistent with open science practices (Nosek et al., 2015). We describe them only briefly here and provide a more detailed account of these studies in the supplementary materials.

### 4.1. Study 7: humans acting on the advice of machines

If people averse to machines making moral decisions, then perhaps people would less averse to limiting machines to an advisory role—in which humans make the final decision. As long as machines are subordinate to humans, the computational power of machines might even lead people to prefer a machine/human team to a human without a machine—demonstrating some value to machines within the moral domain. We tested this idea in the medical domain using the same medical scenario as in past studies. Participants (N = 100, 64% female; age: $M = 35.65$, $SD = 11.72$) read the same basic scenario as in Studies 3 and 6 about a risky surgery that could save a child from paralysis, but also potentially kill him if it fails. Participants were given three options about who should make the decision: (1) HealthComp, (2) Dr. Jones, or (3) Dr. Jones advised by HealthComp.

Out of 100 participants, 4 chose HealthComp to make the decision, 32 chose Dr. Jones, and 64 chose Dr. Jones after receiving a recommendation from HealthComp. A chi-square test revealed a significant difference from an even distribution, $\chi^2(2) = 54.08$, $p < .001$.

These results suggest that most people are willing to have machines involved in moral decisions, as long as they are not the ones to make the actual decisions. However, it bears mentioning that a substantial percentage of people (32%) chose only a human doctor, demonstrating the tenacious aversion to machine moral decision-making.

### 4.2. Study 8: making machines compassionate

One reason that people are averse to machines making moral decisions is that machines are perceived to lack experience. In this study, we tested whether increasing the experience of a machine might decrease the aversion to machine moral decision-making. Participants (N = 240, 60.8% female; age: $M = 34.25$, $SD = 10.03$, see preregistration at https://aspredicted.org/73qi6.pdf) read that HealthComp would make a decision in the medical scenario used in studies 3 and 6–7, and then listened to an audio recording of HealthComp speaking. In the "low experience" condition HealthComp spoke with an expressionless computer voice, and described itself as devoid of emotions. In the "high experience" condition HealthComp used an emotional and expressive voice, and described itself as having the ability to experience emotions. See supplementary materials for full details. Participants then rated the permissibility of HealthComp to make the decision ($\alpha = 0.90$) and the perceived mind of HealthComp (agency: $\alpha = 0.87$; experience: $\alpha = 0.95$).

Although the manipulation did impact ratings of perceived experience (High: $M = 1.95$, $SD = 1.04$; Low: $M = 1.14$, $SD = 0.41$), $F(1, 237) = 62.44$, $p < .001$, there was not a significant difference in permissibility ratings between the high ($M = 2.43$, $SD = 2.43$) and the low ($M = 2.29$, $SD = 1.18$) experience conditions, $t(237) = 0.99$, $p = .321$. A bootstrapping mediation analysis revealed a significant indirect effects such that condition impacted permissibility through both perceived agency, $b = 0.16$, $SE = 0.04$, $CI._{95}[0.09, 0.25]$ and experience, $b = 0.12$, $SE = 0.04$ $CI._{95}[0.06, 0.20]$. Interestingly, when the mind perception mediators were included in the regression, the effect of condition on permissibility became significant but negative, $b = -0.21$, $SE = 0.07$, $p = .003$, $CI._{95}[-0.35, -0.07]$, suggesting that our manipulation had two effects on permissibility which cancelled each other out. This permissibility-reducing effect may be the "Uncanny Valley," as research reveals that seeing experience in a machine can be unnerving (Gray & Wegner, 2012; Mori, 1970). It seems likely that potential feelings of uncanniness could have canceled out any gains in permissibility given by the high experience condition. Whatever the explanation, these results offer only mixed support for the idea that increasing experience could reduce the aversion to machine moral decision-making.

### 4.3. Study 9: are expert machines more acceptable?

In this study, we examined another possible way to reduce the aversion from machines making moral decisions: expertise. In our previous studies, we did not provide any information about the machine's or the human decider's level of expertise. It is possible that people will be less averse to machines moral decision-making if machines have high levels of expertise. We tested this idea using both a within-subject and a between-subject design, to test the robustness of any potential effect.

### 4.4. Within-subject design

Participants (N = 201, 48.3% female; age: $M = 34.42$, $SD = 11.01$, MTurk, see preregistration at https://aspredicted.org/ej6nh.pdf) read about the medical scenario used in studies 3 and 6–8 and then were randomly assigned to one of two conditions. In both conditions, they had to choose whether Dr. Jones or HealthComp should make the surgery decision, but in the equal expertise condition, both had 75% success rates, and in the machine advantage condition, Dr. Jones had a 75% success rate whereas HealthComp that had a 95% success rate.

A chi-squared test revealed that while in the equal expertise condition people were less likely to choose HealthComp over Dr. Jones (7%)—again revealing the averse to machine moral decision-making—in the machine advantage condition, people were more likely to choose HealthComp over Dr. Jones (72.28%), $\chi^2$ (1, N = 201) = 89.37,

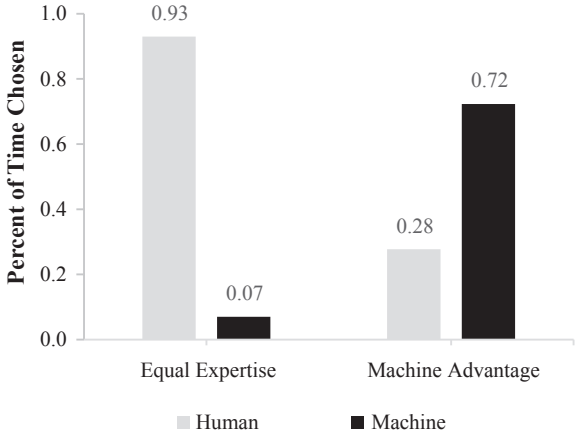


**Fig. 8.** Selection of who should make the medication decision for equal expertise and when the machine has an advantage (Study 9: Within-Subject Design).

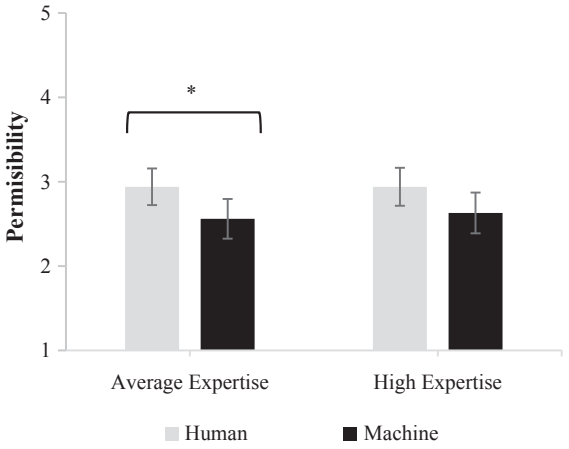


**Fig. 9.** Permissibility of human and machine deciders for negative and positive outcomes (Study 9: Between-Subject Design). Error bars reflect 95% confidence intervals. $^{*}p < .05$.

$p < .001$, $\varphi = 0.67$ (see Kramer, Borg, Conitzer, & Sinnott-Armstrong, 2018 for a similar finding). Although again, choosing the machine was far from ceiling. See Fig. 8.

### 4.5. Between-subject design

The within-subjects decision revealed that people are more likely to choose the machine to make a moral decision when the difference in expertise is made salient through a pair-wise choice (Hsee, Loewenstein, Blount, & Bazerman, 1999). Would the results replicate when the relative difference is less salient, such as when people simply evaluate the permissibility of average/expert HealthComp or average/expert Dr. Jones in a between-subjects design?

In this 2 × 2 between-subjects study, participants (N = 482, 58.9% female; age: $M = 35.80$, $SD = 11.90$, Mturk, see preregistration at https://aspredicted.org/cb742.pdf, 73 excluded for failing comprehension questions as specified in the preregistration) were randomly assigned to a decider condition (HealthComp or Dr. Jones) and an expertise condition (average or high). After reading the medical scenario used in studies 3 and 6–8, all participants read that "On average, doctors have a success rate (their decisions have positive outcomes) of 75%". In the average expertise condition participants read that either HealthComp or Dr. Jones has been making such decisions in the hospital for 3 years and has a success rate of 75%." In the high expertise condition they read that HealthComp's or Dr. Jones's success rate is 95%. After reading the scenario participants rated the permissibility of the decider in making the decision (Cronbach's $\alpha = 0.88$) answered the comprehension questions and provided demographic information.

Consistent with an aversion to machines making moral decisions, a 2 (decider: human, machine) × 2 (expertise: average, high) between-subject ANOVA of permissibility ratings revealed a main effect for decider, $F(1, 405) = 14.36$, $p < .001$, partial $\eta^2 = 0.34$, such that across levels of expertise, people rated HealthComp ($M = 2.49$, $SD = 1.11$) as less permissible than Dr. Jones ($M = 2.94$, $SD = 1.24$) in making the decision. There was no main effect of expertise nor was there an interaction between expertise and decider, all $F$s < 2, $p$s > .3. See Fig. 9.

These results again reveal the tenacity of the aversion to machine moral decision-making. In fact, a planned contrast revealed marginally significant higher permissibility ratings for Dr. Jones when he had average expertise ($M = 2.94$, $SD = 1.25$) than for HealthComp when he had high expertise ($M = 2.63$, $SD = 1.13$), $t(405) = 1.84$, $p = .067$. In other words, people would (almost) rather have an average doctor than an expertise machine—unless these differences in expertise are made explicit through pair-wise comparisons, as revealed by the within-subjects design.

Together, the results of Section 3 studies suggest that reducing the aversion to machine moral decision-making is not easy, and depends upon making very salient the expertise of machines (Study 9) and the over-riding authority of humans (Study 7)—and even then, it still lingers.

## 5. General discussion

Nine studies investigated the potential aversion to machines making moral decisions. People prefer humans over machines for decisions of life and death in driving (Study 1), law (Study 2), medicine (Studies 3 and 6–9), and the military (Studies 4–5). This aversion is partially explained by reduced perceptions of minds in machines (Studies 2–6), and persists when the outcome of the moral decision is specified—whether negative (Studies 3–5) or positive (Study 5–6). This aversion is not impacted by manipulations of experience (Study 8), but is somewhat lessened when machines are limited to an advisory role (Study 7), and when the greater expertise of machines is made extremely salient (Study 9).

Despite the robustness of these effects, we acknowledge that they must be understood within context. First, for maximum power, we examined the most paradigmatic of moral scenarios—dilemmas in which life and death hang in the balance. Although these scenarios capture potential applications of autonomous machines in driving, medicine, the law, and the military, there are undoubtedly many more domains in which machines can make decisions. It is an open question how much the aversion generalizes to other moral and non-moral domains, and how much perceptions of mind matter. Indeed, in our studies we did not test for specificity to the moral domain, and it is possible that this aversion might exist in non-moral domains as well.

Second, the sample we used was from an online sample (Amazon's Mechanical Turk) from the US and Canada. While we have no reason to believe that this population is systematically different than other potential samples (Buhrmester, Kwang, & Gosling, 2011), future research should test for generalizability. It is especially worth investigating whether people from other cultures share North American concerns about machines making moral decisions. For example, people from Japan might be more familiar with robots in everyday life and this familiarity may lead to more acceptance of machine moral decision-making.

Third, our research focused on only one aspect of morality—the permissibility of making moral decisions. We acknowledge that are many other important elements that could show intriguing effects with machines including blame (Malle, Scheutz, Forlizzi, & Voiklis, 2016),

punishment (Lokhorst & van den Hoven, 2011), and moral value (Bartneck, Verbunt, Mubin, & Al Mahmud, 2007). Fourth, our studies all involved dispassionate third-person decisions. It is possible that this aversion could be weaker—or perhaps stronger—in cases where people are personally involved in the outcome. If the life of your own child hangs in the balance, would you want a robot making a moral decision? Fifth, it is possible that with stronger manipulations, our attempts to reduce the aversion would be more successful. For example, although we had only mixed results in trying to imbue machines with perceived experience—perhaps because of the uncanny valley—future attempts might be more successful (Malle et al., 2016; Waytz, Heafner, & Epley, 2014).

Our results are consistent with other recent research on whether people want machines to make decisions that impact humans (Gogoll & Uhl, 2018; Kramer et al., 2018). For example, both our studies and those of Kramer and colleagues' (2018) highlight the importance of expertise in people's willingness to accept machine decisions. Gogoll & Uhl (2018) also found people preferred to delegate decisions to human rather to machines within economic games. Our research extends these initial findings to a wider variety of moral contexts, and most importantly, demonstrates the role of mind perception in the aversion from machines making moral decisions.

### 5.1. Implications

Machines play a large role in industry and a growing role in social domains. For example, robots are assisting with mental health interventions (Rabbitt, Kazdin, & Scassellati, 2015) and are helping children with autism practice their social skills (Kim, Paul, Shic, & Scassellati, 2012). However, these data reveal that they are not yet accepted as autonomous moral deciders. To the extent that scientists and policymakers are concerned with public opinion, they might carefully consider how much machines should be given autonomy in moral decision-making. Importantly, this doesn't mean that scholars should stop their important on revealing how to design moral machines (Conitzer et al., 2017; Kuipers, 2016; Malle, 2016; Tonkens, 2012; Wiltshire, 2015), but only that we might first consider what kind of decisions humans want machines to make.

This work also highlights the importance of mind perception within morality. Past research has revealed that people use perceptions of agency and experience when making decisions about what is right or wrong (Gray et al., 2012, Schein & Gray, 2018). This work reveals that people use the same perceptions when making *meta*-moral decisions—who gets to make decisions about right or wrong. Our research therefore supplements normative philosophical discussions about the role of mental qualities in questions about who is a legitimate moral agent (Damm, 2010; Hume, 1751; Kant, 1785). Laypeople believe that also experience, and not only agency, are essential to being a moral agent.

Although the studies here revealed the importance of mind for moral agency by comparing machines (who are generally seen to lack mind) to humans (who are generally seen to possess mind), the results should apply more generally. For example, it should seem more permissible for people to make moral decisions when they are perceived to possess more agency and experience. To test this idea, we ran a study modeled after the military scenarios in Study 5, in which participants (N = 485, 57.9% female, age: $M = 35.99$, $SD = 11.57$, 12 exclusions) read about a human agent, Colonel Jones, who made a decision about a risky drone missile strike. In a 2 × 2 between-subjects design Colonel Jones was described as having high/low experience and high/low agency. Moral permissibility judgments (Cronbach's $\alpha = 0.91$) were higher when Colonel Jones had high agency ($M = 3.27$, $SD = 1.17$) versus low agency ($M = 2.71$, $SD = 1.22$), $F(1, 469) = 25.54$, $p < .001$, partial $\eta^2 = 0.05$, and were higher when he had high experience ($M = 3.24$, $SD = 1.19$) versus low experience ($M = 2.76$, $SD = 1.22$), $F(1, 469) = 18.83$, $p < .001$, partial $\eta^2 = 0.04$. The interaction between agency and experience was only marginally significant, $p = .050$; see supplementary materials for the full study. These results further support the idea that both agency and experience are important for judgments about who can make legitimate moral judgments.

### 5.2. Conclusion

Machines are becoming ubiquitous in modern society, with algorithms making decisions about navigation (Google Maps), advertising (Amazon), and even dating (OK Cupid). Although people are often indifferent about the relentless creep of artificial intelligence, they appear to less accepting of machines making moral decisions. When human life and death hang in the balance, it seems that we want another human—with a fully human mind—to make the call.

## Authors' note

For funding, the first author acknowledges the National Science Foundation SBE Postdoctoral Research Fellowship (1714298), and the second author acknowledges the Charles Koch Foundation. This work was presented in seminars in UNC Chapel-Hill, Wake Forest University and at a workshop at the Rock Ethics Institute at Penn State. The authors thank the participants in those forums for their invaluable feedback.

## References

Aaltola, E. (2014). Affective empathy as core moral agency: Psychopathy, autism and reason revisited. *Philosophical Explorations, 17*(1), 76–92. https://doi.org/10.1080/13869795.2013.825004.

Allen, C., Wallach, W., & Smit, I. (2006). Why machine ethics? *IEEE Intelligent Systems, 21*(4), 12–17. https://doi.org/10.1109/MIS.2006.83.

Angwin, J., Larson, J., Surya, M., & Lauren, K. (2016). Machine Bias. Retrieved February 21, 2018, from < https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing > .

Aristotle (350BC). The Nicomachean ethics. (W. D. Ross, Trans.). New York, NY: World Library Classics.

Arkin, R. C. (2009). Ethical robots in warfare. *IEEE Technology and Society Magazine, 28*(1), 30–33. https://doi.org/10.1109/MTS.2009.931858.

Arkin, R. C. (2008). Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture Part I: Motivation and philosophy. *Proceedings of the 3rd international conference on human robot interaction - HRI '08* (pp. 121). . https://doi.org/10.1145/1349822.1349839.

Asimov, I. (1950). *I, Robot.* New York, NY: Doubleday and Company.

Bartholomew-Biggs, M. C., Parkhurst, S. C., & Wilson, S. P. (2003). Global optimization approaches to an aircraft routing problem. *European Journal of Operational Research, 146*(2), 417–431. https://doi.org/10.1016/S0377-2217(02)00229-1.

Bartneck, C., Verbunt, M., Mubin, O., & Al Mahmud, A. (2007). To kill a mockingbird robot. *Proceeding of the ACM/IEEE international conference on human-robot interaction - HRI '07* (pp. 81). . https://doi.org/10.1145/1228716.1228728.

Bastian, B., Loughnan, S., Haslam, N., & Radke, H. R. M. (2012). Don't mind meat? The denial of mind to animals used for human consumption. *Personality and Social Psychology Bulletin, 38*(2), 247–256. https://doi.org/10.1177/0146167211424291.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science, 352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654.

Bostrom, N. (2014). *Superintelligence.* Oxford: Oxford University Press.

Brink, K. A., Gray, K., & Wellman, H. M. (2017). Creepiness creeps in: Uncanny valley feelings are acquired in childhood. *Child Development, 0*(0), https://doi.org/10.1111/cdev.12999.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's mechanical Turk: A new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science, 6*(1), 3–5. https://doi.org/10.1177/1745691610393980.

Cameron, C. D., Lindquist, K. A., & Gray, K. (2015). A constructionist review of morality and emotions: no evidence for specific links between moral content and discrete emotions. *Personality and Social Psychology Review, 19*(4), 371–394. https://doi.org/10.1177/1088868314566683.

Cameron, J. (1984). The terminator. USA.

Caplan, J. M. (2007). What Factors Affect Parole-A Review of Empirical Research. *Fed. Probation, 71*, 16.

Cárdenas-Barrón, L. E., Treviño-Garza, G., & Wee, H. M. (2012). A simple and better algorithm to solve the vendor managed inventory control system of multi-product multi-constraint economic order quantity model. *Expert Systems with Applications, 39*(3), 3888–3895. https://doi.org/10.1016/j.eswa.2011.09.057.

Chalmers, D. J. (1997). *The conscious mind: In search of a fundamental theory.* New York, NY: Oxford University Press.

Chen, L., Mislove, A., & Wilson, C. (2016). An empirical analysis of algorithmic pricing on

amazon marketplace. *Proceedings of the 25th international conference on World Wide Web - WWW '16* (pp. 1339–1349). . https://doi.org/10.1145/2872427.2883089.
Chouard, T. (2016). The Go Files: AI computer clinches victory against Go champion. *Nature*. https://doi.org/10.1038/nature.2016.19553.
Clarke, R. (1992). Free will and the conditions of moral responsibility. Retrieved from *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 66*(1), 53–74. <http://www.jstor.org/stable/4320296>.
Coeckelbergh, M. (2010). Moral appearances: Emotions, robots, and human morality. *Ethics and Information Technology, 12*(3), 235–241. https://doi.org/10.1007/s10676-010-9221-y.
Conitzer, V., Sinnott-Armstrong, W., Borg, J. S., Deng, Y., & Kramer, M. (2017). Moral decision making frameworks for artificial intelligence. In Association for the advancement of artificial intelligence, (Moor 2006) (pp. 4831–4835). Retrieved from < https://pdfs.semanticscholar.org/a3bb/ffdcc1c7c4cae66d6af373651389d94b7090.pdf > .
Cushman, F. (2008). Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition, 108*(2), 353–380. https://doi.org/10.1016/j.cognition.2008.03.006.
Damm, L. (2010). Emotions and moral agency. *Philosophical Explorations, 13*(3), 275–292. https://doi.org/10.1080/13869795.2010.501898.
Dawson, R. O. (1966). The decision to grant or deny parole: A study of parole criteria in law and practice. *Washington University Law Quarterly, 1966*(3), 234–303.
De Waal, F. (2010). *The age of empathy: Nature's lessons for a kinder society.* New York, NY: Random House.
Decety, J., & Cowell, J. M. (2014). The complex relation between morality and empathy. *Trends in Cognitive Sciences, 18*(7), 337–339. https://doi.org/10.1016/j.tics.2014.04.008.
DeVito, M. A. (2017). From editors to algorithms: A values-based approach to understanding story selection in the Facebook news feed. *Digital Journalism, 5*(6), 753–773. https://doi.org/10.1080/21670811.2016.1178592.
Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033.
Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature, 415*, 137–140. https://doi.org/10.1038/415137a.
Fischer, J. M. (2005). Free Will and Moral Responsibility. In D. Copp (Ed.). *The Oxford handbook of ethical theory* (pp. 321–354). Oxford University Press. https://doi.org/10.1093/0195147790.003.0013.
Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines, 14*(3), 349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d.
Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy, 66*(23), 829. https://doi.org/10.2307/2023833.
Gogoll, J., & Uhl, M. (2018). Rage against the machine: Automation in the moral domain. *Journal of Behavioral and Experimental Economics, 74*, 97–103. https://doi.org/10.1016/j.socec.2018.04.003.
Goodman, J. K., Cryder, C. E., & Cheema, A. (2013). Data collection in a flat world: The strengths and weaknesses of mechanical Turk samples. *Journal of Behavioral Decision Making, 26*(3), 213–224. https://doi.org/10.1002/bdm.1753.
Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science (New York, New York), 315*(5812), 619. https://doi.org/10.1126/science.1134475.
Gray, K., Jenkins, A. C., Heberlein, A. S., & Wegner, D. M. (2011). Distortions of mind perception in psychopathology. *Proceedings of the National Academy of Sciences of the United States of America, 108*(2), 477–479. https://doi.org/10.1073/pnas.1015493108.
Gray, K., Schein, C., & Cameron, C. D. (2017). How to think about emotions and morality: Circles, not arrows. *Current Opinion in Psychology, 17*, 41–46.
Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology, 96*(3), 505–520. https://doi.org/10.1037/a0013748.
Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition, 125*(1), 125–130. https://doi.org/10.1016/j.cognition.2012.06.007.
Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry, 23*(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387.
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*(5537), 2105–2108. https://doi.org/10.1126/science.1062872.
Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgment. *Psychological Review, 108*(4), 814–834. https://doi.org/10.1037//0033-295X.
Haidt, J., Koller, S. H., & Dias, M. G. (1993). Affect, culture, and morality, or is it wrong to eat your dog? Retrieved from *Journal of Personality and Social Psychology, 65*(4), 613–628. <http://www.ncbi.nlm.nih.gov/pubmed/8229648>.
Harris, S. (2012). *Free will.* New York, NY: Free Press.
Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10*(3), 252–264. https://doi.org/10.1207/s15327957pspr1003_4.
Hauser, D. J., & Schwarz, N. (2016). Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior Research Methods, 48*(1), 400–407. https://doi.org/10.3758/s13428-015-0578-z.
Heires, K. (2016). Rise of the Robots. Retrieved March 25, 2018, from <http://www.rmmagazine.com/2016/09/01/rise-of-the-robots/>.
Hellström, T. (2013). On the moral responsibility of military robots. *Ethics and Information Technology, 15*(2), 99–107. https://doi.org/10.1007/s10676-012-9301-2.
Hern, A. (2017). How social media filter bubbles and algorithms influence the election. Retrieved March 25, 2018, from < https://www.theguardian.com/technology/2017/may/22/social-media-election-facebook-filter-bubbles > .
Hertz, S. G., & Krettenauer, T. (2016). Does moral identity effectively predict moral behavior? A meta-analysis. *Review of General Psychology, 20*(2), 129–140. https://doi.org/10.1037/gpr0000062.
Himma, K. E. (2009). Artificial agency, consciousness, and the criteria for moral agency: What properties must an artificial agent have to be a moral agent? *Ethics and Information Technology, 11*(1), 19–29. https://doi.org/10.1007/s10676-008-9167-5.
Hornsey, M. J., Harris, E. A., & Fielding, K. S. (2018). The psychological roots of anti-vaccination attitudes: A 24-nation investigation. *Health Psychology, 37*(4), 307–315. https://doi.org/10.1037/hea0000586.
Horowitz, M. (2016). The ethics and morality of robotic warfare: Assessing the debate over autonomous weapons. *American Academy of Arts & Sciences Project on New Dilemmas in Ethics, Technology, & War for Special Issue of Daedalus, 145*(4), 25–36. https://doi.org/10.1017/CBO9781107415324.004.
Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin, 125*(5), 576–590. https://doi.org/10.1037/0033-2909.125.5.576.
Hume, D. (1751). *An enquiry concerning the principles of morals.* Oxford: Clarendon Press.
Hunt, E. (2016). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. Retrieved February 25, 2018, from < https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter > .
Jahoda, G. (1999). *Images of savages: Ancients [sic] roots of modern prejudice in Western culture.* London, UK: Routledge.
Johnson, P. E. (1973). Federal parole procedures. *Administrative Law Review,* 459–529.
Johnson, A. M., & Axinn, S. (2013). The morality of autonomous robots. *Journal of Military Ethics, 12*(2), 129–141. https://doi.org/10.1080/15027570.2013.818399.
Kant, I. (1785). *Groundwork of the metaphysics of morals.* New York, NY: Harperand Row Publishers.
Kant, I. (1788). *Critique of practical reason.* Cambridge: Cambridge University Press.
Kauppinen, A. (2017). Empathy and moral judgment. In The Routledge handbook of philosophy of empathy, (September 2015) (pp. 215–226). Chapter xiii, 396pages. Retrieved from < https://search.proquest.com/docview/1960502860?accountid=13904 > .
Kehl, D., Guo, P., & Kessler, S. (2017). *Algorithms in the criminal justice system: Assessing the use of risk assessments in sentencing. Responsive communities initiative, Berkman Klein.* Center for Internet & Society.
Kim, E., Paul, R., Shic, F., & Scassellati, B. (2012). Bridging the research gap: Making HRI useful to individuals with autism. *Journal of Human-Robot Interaction, 1*(1), 26–54. https://doi.org/10.5898/JHRI.1.1.Kim.
Koenigs, M., Young, L., Adolphs, R., Tranel, D., Cushman, F., Hauser, M., & Damasio, A. (2007). Damage to the prefrontal cortex increases utilitarian moral judgements. *Nature, 446*(7138), 908–911. https://doi.org/10.1038/nature05631.
Kramer, M. F., Borg, J. S., Conitzer, V., & Sinnott-Armstrong, W. (2018). When do people want AI to make decisions ? In Proceedings of first annual AAAI/ACM conference on artificial intelligence, ethics, and society (AIES-18).
Kubrick, S. (1968). 2001: a space odyssey. USA.
Kuipers, B. (2016). Human-like morality and ethics for robots. Retrieved from <https://pdfs.semanticscholar.org/e76d/c10c84342296ecf06debeedf284edb897704.pdf>.
Leyens, J.-P., Paladino, P. M., Rodriguez-Torres, R., Vaes, J., Demoulin, S., Rodriguez-Perez, A., & Gaunt, R. (2000). The emotional side of prejudice: The attribution of secondary emotions to ingroups and outgroups. *Personality and Social Psychology Review, 4*(2), 186–197. https://doi.org/10.1207/S15327957PSPR0402_06.
Locke, J. (1836). An essay concerning human understanding book II : Ideas. T. Tegg and Son. Retrieved from < http://www.earlymoderntexts.com/assets/pdfs/locke1690book1.pdf > .
Lokhorst, G.-J., & van den Hoven, J. (2011). Responsibility for military robots. In P. Lin, K. Abeney, & George A. Bekey (Eds.). *Robot ethics: the ethical and social implications of robotics* (pp. 145–156). Cambridge, MA: The MIT Press.
Malle, B. F. (2016). Integrating robot ethics and machine morality: The study and design of moral competence in robots. *Ethics and Information Technology, 18*(4), 243–256. https://doi.org/10.1007/s10676-015-9367-8.
Malle, B. F., & Scheutz, M. (2014). Moral competence in social robots. In 2014 IEEE international symposium on ethics in science, technology and engineering, ETHICS 2014. http://doi.org/10.1109/ETHICS.2014.6893446.
Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (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 (pp. 117–124). http://doi.org/10.1145/2696454.2696458.
Malle, B. F., 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 2016 11th ACM/IEEE international conference on human-robot interaction (HRI) (pp. 125–132). IEEE. http://doi.org/10.1109/HRI.2016.7451743.
Malle, B. F., Guglielmo, S., & Monroe, A. E. (2014). A theory of blame. *Psychological Inquiry, 25*(2), 147–186. https://doi.org/10.1080/1047840X.2014.877340.
Markoff, J. (2011). On 'Jeopardy!' Watson win is all but trivial. Retrieved March 25, 2018, from < https://www.nytimes.com/2011/02/17/science/17jeopardy-watson.html > .
McFarland, M. (2014). Elon Musk: 'With artificial intelligence we are summoning the demon.' - The Washington Post. Retrieved February 26, 2018, from < https://www.washingtonpost.com/news/innovations/wp/2014/10/24/elon-musk-with-artificial-intelligence-we-are-summoning-the-demon/?utm_term=.02d648908751 > .
Mele, A., & Sverdlik, S. (1996). Intention, intentional action, and moral responsibility. Retrieved from *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 82*(3), 265–287. <http://www.pdcnet.org/oom/service?url_ver=

Z39.88-2004&rft_val_fmt=&rft.imuse_id=jphil_1969_0066_0023_0829_0839&svc_id=info:www.pdcnet.org/collection>.
Monroe, A. E., Brady, G. L., & Malle, B. F. (2017). This isn't the free will worth looking for: General free will beliefs do not influence moral judgments, agent-specific choice ascriptions do. *Social Psychological and Personality Science, 8*(2), 191–199. https://doi.org/10.1177/1948550616667616.
Monroe, A. E., Dillon, K. D., & Malle, B. F. (2014). Bringing free will down to Earth: People's psychological concept of free will and its role in moral judgment. *Consciousness and Cognition, 27*(1), 100–108. https://doi.org/10.1016/j.concog.2014.04.011.
Mori, M. (1970). The uncanny valley. *Energy, 7*(4), 33–35.
Nahmias, E., Shepard, J., & Reuter, S. (2014). It's OK if "my brain made me do it": People's intuitions about free will and neuroscientific prediction. *Cognition, 133*(2), 502–516. https://doi.org/10.1016/j.cognition.2014.07.009.
Newborn, M. (2011). *Beyond deep blue: Chess in the stratosphere.* New York, NY: Springer.
Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*(6242), 1422–1425. https://doi.org/10.1126/science.aab2374.
O'Connor, T. (2000). Causality, mind, and free will. *Nous, 34*(s14), 105–117. https://doi.org/10.1111/0029-4624.34.s14.6.
Osborne, J., Simon, S., & Collins, S. (2003). Attitudes towards science: A review of the literature and its implications. *International Journal of Science Education, 25*(9), 1049–1079. https://doi.org/10.1080/0950069032000032199.
Pardo del Val, M., & Martínez Fuentes, C. (2003). Resistance to change: A literature review and empirical study. *Management Decision, 41*(2), 148–155. https://doi.org/10.1108/00251740310457597.
Parkin, S. (2016). The artificially intelligent doctor will hear you now - MIT Technology Review. Retrieved March 25, 2018, from <https://www.technologyreview.com/s/600868/the-artificially-intelligent-doctor-will-hear-you-now/>.
Pinker, S. (2011). *The better angels of our nature: Why violence has declined.* New York, NY: Viking.
Pinker, S. (2016). Why computers won't takeover the world, with Steven Pinker|Big Think. Retrieved February 26, 2018, from <http://bigthink.com/videos/steven-pinker-on-artificial-intelligence-apocalypse>.
Preacher, K. J., & Hayes, A. F. (2008). Asymptotic and resampling strategies for assessing and comparing indirect effects in multiple mediator models. *Behavior Research Methods, 40*(3), 879–891. https://doi.org/10.3758/BRM.40.3.879.
Prinz, J. (2007). *The emotional construction of morals.* Oxford: Oxford University Press.
Rabbitt, S. M., Kazdin, A. E., & Scassellati, B. (2015). Integrating socially assistive robotics into mental healthcare interventions: Applications and recommendations for expanded use. *Clinical Psychology Review, 35*, 35–46. https://doi.org/10.1016/j.cpr.2014.07.001.
Rifkin, J. (2009). *The empathic civilization: The race to global consciousness in a world in crisis.* New York, NY: Penguin.
Roberts, S. (2017). Christopher strachey's nineteen-fifties love machine - The New Yorker. Retrieved February 25, 2018, from <https://www.newyorker.com/tech/elements/christopher-stracheys-nineteen-fifties-love-machine>.
Robinson, D. N. (1996). *Wild beasts & idle humours: The insanity defense from antiquity to the present.* Cambridge, Mass.: Harvard University Press.
Rosati, C. S. (2016). Moral motivation. In E. N. Zalta (Ed.), The stanford encyclopedia of philosophy (Winter 201). Retrieved from < https://plato.stanford.edu/archives/win2016/entries/moral-motivation/ > .
Rudman, L. A., & Mescher, K. (2012). Of animals and objects: Men's implicit dehumanization of women and likelihood of sexual aggression. *Personality and Social Psychology Bulletin, 38*(6), 734–746. https://doi.org/10.1177/0146167212436401.
Schein, C., & Gray, K. (2018). The theory of dyadic morality: reinventing moral judgment by redefining harm. *Personality and Social Psychology Review, 22*(1), 32–70. https://doi.org/10.1177/1088868317698288.
Shaw, L. L., Batson, C. D., & Todd, R. M. (1994). Empathy avoidance: Forestalling feeling for another in order to escape the motivational consequences. *Journal of Personality and Social Psychology, 67*(5), 879–887. https://doi.org/10.1037//0022-3514.67.5.879.
Shweder, R. A., Mahapatra, M., & Miller, J. (1987). Culture and moral development. In J. Kagan, & S. Lamb (Eds.). *The emergence of morality in young children* (pp. 1–83). Chicago, IL: University of Chicago Press.
Singer, P. (1975). *Animal liberation.* New York, NY: Random House.
Skitka, L. J. (2010). The psychology of moral conviction. *Social and Personality Psychology Compass, 4*(4), 267–281. https://doi.org/10.1111/j.1751-9004.2010.00254.x.
Soroka, S. N. (2006). Good news and bad news: Responses asymmetric information to economic information. *The Journal of Politics, 68*(2), 372–385. https://doi.org/10.1111/j.1468-2508.2006.00413.x.
Steinert, S. (2014). The five robots-A taxonomy for roboethics. *International Journal of Social Robotics, 6*(2), 249–260. https://doi.org/10.1007/s12369-013-0221-z.
Swaney, P. J., Mahoney, A. W., Hartley, B. I., Remirez, A. A., Lamers, E., Feins, R. H., ... Webster, R. J. (2017). Toward transoral peripheral lung access: Combining continuum robots and steerable needles. *Journal of Medical Robotics Research, 02*(01), 1750001. https://doi.org/10.1142/S2424905X17500015.
Tonkens, R. (2012). The Case against robotic warfare: A response to Arkin. *Journal of Military Ethics, 11*(2), 149–168. https://doi.org/10.1080/15027570.2012.708265.
Validi, S., Bhattacharya, A., & Byrne, P. J. (2015). A solution method for a two-layer sustainable supply chain distribution model. *Computers and Operations Research, 54*, 204–217. https://doi.org/10.1016/j.cor.2014.06.015.
van den Berg, J., Patil, S., & Alterovitz, R. (2017). Motion planning under uncertainty using differential dynamic programming in belief space. In H. I. Christensen, & O. Khatib (Eds.). *Robotics research: The 15th international symposium ISRR* (pp. 473–490). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-29363-9_27.
van Wynsberghe, A. (2013). Designing robots for care: Care centered value-sensitive design. *Science and Engineering Ethics, 19*(2), 407–433. https://doi.org/10.1007/s11948-011-9343-6.
Vanderblit, A. T. (1956). Judges and jurors: Their functions, qualifications and selection. *Boston University Law Review, 36*(1), 1–76.
Wallach, W., & Allen, C. (2009). Moral machines: teaching robots right from wrong. Moral machines: teaching robots right from wrong. New York, NY: Oxford University Press. < http://doi.org/10.1093/acprof:oso/9780195374049.001.0001 > .
Wallach, W., Franklin, S., & Allen, C. (2010). A conceptual and computational model of moral decision making in human and artificial agents. *Topics in Cognitive Science, 2*(3), 454–485. https://doi.org/10.1111/j.1756-8765.2010.01095.x.
Warren, M. A. (1997). *Moral status: Obligations to persons and other living things.* Clarendon Press.
Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology, 52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005.
Waytz, A., & Schroeder, J. (2014). Overlooking others: Dehumanization by commision and omission. *Testing, Psychometrics, Methodology in Applied Psychology, 21*(3), 251–266. https://doi.org/10.4473/XXXX.
Wegner, D. M., & Gray, K. (2017). *The mind club.* New York, NY: Viking.
Wiltshire, T. J. (2015). A prospective framework for the design of ideal artificial moral agents: Insights from the science of heroism in humans. *Minds and Machines, 25*(1), 57–71. https://doi.org/10.1007/s11023-015-9361-2.
Wright, J. R., & Leyton-Brown, K. (2010). Beyond equilibrium : Predicting human behavior in normal-form games. In Proceedings of the twenty-fourth AAAI conference on artificial intelligence (AAAI-10) (pp. 901–907). < http://doi.org/978-1-57735-463-5 > .
Yamane, K., Fujiwara, J., Endo, Y., Machii, K., Kumagai, M., Yokota, T., &S. Matsou (2011). U.S. Patent No. 8,068,973. Washington, Washington, DC: U.S. Patent and Trademark Office.
Zaki, J. (2018). Empathy is a moral force. In K. Gray, & J. Graham (Eds.). *Atlas of moral psychology* (pp. 49–58). New York, NY: The Guilford Press.