The Moral Psychology of Artificial Intelligence

Jean-François Bonnefon, Iyad Rahwan, Azim Shariff

The Moral Psychology of Artificial Intelligence

Jean-François Bonnefon,[1] Iyad Rahwan,[2] and Azim Shariff[3]

[1]Toulouse School of Economics, Centre National de la Recherche Scientifique (TSM-R), Toulouse, France, 31000; email: jean-francois.bonnefon@tse-fr.eu
[2]Max Planck Institute for Human Development, Center for Humans & Machines, Berlin 14195, Germany
[3]Department of Psychology, University of British Columbia, Vancouver V6T 1Z4, Canada

Xxxx. Xxx. Xxx. Xxx. YYYY. AA:1–25

https://doi.org/10.1146/((please add article doi))

Keywords

psychology, morality, artificial intelligence, agent, patient, algorithmic bias, blame, ethical dilemmas, value alignment, cooperation, humanization, delegation, AI-mediated communication

Abstract

Moral psychology was shaped around three categories of agents and patients: humans, other animals, and supernatural beings. Rapid progress in Artificial Intelligence has introduced a fourth category for our moral psychology to deal with: intelligent machines. Machines can perform as moral agents, making decisions that affect the outcomes of human patients, or solving moral dilemmas without human supervision. Machines can be as perceived moral patients, whose outcomes can be affected by human decisions, with important consequences for human-machine cooperation. Machines can be moral proxies, that human agents and patients send as their delegates to a moral interaction, or use as a disguise in these interactions. Here we review the experimental literature on machines as moral agents, moral patients, and moral proxies, with a focus on recent findings and the open questions that they suggest.

Contents

1. INTRODUCTION

Human-AI encounters have, until recently, been confined to science fiction. Droids and Replicants, Commander Data and Agent Smith, the T-800 and HAL-9000 have all prodded people to consider the moral questions that arise when people interact with advanced machines capable of human-level intelligence. How ought they be treated? How will they treat us? And how do they change how we treat each other?

Today, AI has finally moved beyond fiction and begun its march towards ubiquity. With the pace of AI innovations beginning to be measured in months rather than years, there is a feeling of being in the foothills of the long-promised AI revolution. As of writing, generative AI and large language models have redoubled public interest in AI. Virtual assistants are everyday tools. Recommendation algorithms govern our attention.

As human encounters with robots and other AI have multiplied, so have efforts to understand the moral psychology of AI. Some of this work involves speculative research about the not-yet-possible. Decision-making on how driverless cars ought to be programmed to distribute risk to different individuals is one widely-discussed example. What has been more common, however, is a process of catch-up in which researchers have been racing to understand the psychological dimensions that accompany the rapidly emerging innovations in the AI space. Online bots, AI-assisted medical diagnoses, and the use of predictive algorithms for policing and incarceration, have all provoked important moral questions about trust, bias, and value alignment.

In the current paper, we review the research on the moral roles that intelligent machines have begun to occupy. As moral agents, machines are implicitly or explicitly charged with contributing to or making moral decisions—often about matters of life and death such as who deserves a kidney transplant, whose safety to prioritize in traffic collisions, or who goes to jail. How should we align AI-driven decisions in these domains with human values? As moral patients, machines are the subjects of human moral behavior, be it cooperative or competitive, sympathetic or malicious. Although considering the patiency of non-sentient machines may sound like more fanciful flirtation with science fiction, figuring out how to increase human cooperation with AI is already a present-day challenge. Finally, in the role of moral proxies, machines may serve as moral intermediaries in people's treatment of their fellow humans. In this role, people can use machines to disguise, whitewash, or carry out

their morally questionable behavior.

THE MORAL TYPECASTING OF MACHINES

In this article, we speak of machines as moral agents or patients purely for the purpose of organizing the empirical findings of moral psychology: we do not mean that experts should use these terms the way we do, or that laypersons do use these terms the way we do. Our use of these terms does not imply any ontological commitment; we have nothing to contribute to philosophical debates about whether it is appropriate to attribute agency or patiency to a machine. Furthermore, our use of these terms does not imply that people engage in a binary classification of machines as agents or patients. Agency and patiency are two continuous dimensions of mind perception (Wegner & Gray 2016), and people can perceive machines as occupying various positions in that two-dimensional space.

2. MACHINES AS MORAL AGENTS

2.1. Implicit Moral Machines

A machine can be perceived as a moral agent even when its programming does not explicitly encode moral values—as long as the consequences of its actions can fall in the moral domain. This is what we call an *implicit* moral agent (Moor 2006), or an implicit moral machine. The prototypical case here is that of a machine whose mistakes can create harm. For example, medical AI can harm patients by making a wrong diagnosis, a recommendation algorithm can create harm by steering a child to a violent video, and a face recognition algorithm can harm you by mistaking you for a known terrorist. These implicit moral machines are not necessarily trying to solve moral dilemmas, but their failures have moral implications. Accordingly, from a moral psychology perspective, their performance is the most important consideration. We will consider in turn the expectations that people have about the performance of machines whose mistakes can create harm, and their reactions to these mistakes. More specifically, we will consider the number of mistakes people are willing to tolerate from implicit moral machines, their concerns about the distribution of these mistakes across vulnerable and less vulnerable groups, and the blame they direct towards machines who fail alone or in conjunction with humans.

2.1.1. Performance. How many crashes are you willing to tolerate from self-driving cars, per million kilometers? How many mistakes are you willing to accept from a skin cancer detection algorithm, per million patients? These are very hard questions. An easy way out would be to answer 'zero', that is, to require perfect performance from a machine before it is allowed to replace humans—but this extreme position would forfeit the benefits that machines can deliver even before they are fail-proof. If we require self-driving cars to be perfectly safe before we allow them on the road, we sacrifice the thousands of lives they could have saved by being allowed on the road just a little sooner (Kalra & Groves 2017). If we wait for skin cancer detection algorithms to be perfectly accurate, we sacrifice the thousands of lives that could have been saved by an earlier detection (Esteva et al. 2017). As a result, we may have a moral imperative to allow machines to make some mistakes, and a need to decide how many we will allow.

Generally speaking, we may not want to let machines make decisions if they make more harmful mistakes than humans do—and conversely, we may be willing to let machines make decisions as soon as they make fewer harmful mistakes than humans do. This is the approach taken in several policy reports about autonomous driving, which suggest that the minimal requirement before deploying self-driving cars on our roads is that they are provably safer than the average human driver (Bonnefon et al. 2020b, Luetge 2017, Santoni de Sio 2021). Providing objective evidence that a machine performs better than humans is not trivial to begin with, though (Kalra & Paddock 2016, Kleinberg et al. 2018, Noy et al. 2018), and psychological biases may complicate things even further.

Indeed, it appears that people may have extreme performance requirements for implicit moral machines, because they expect a substantial increase over baseline human performance, while overestimating this baseline human performance. For example, a representative sample of the German population believed that human experts would have a 20-30 percent mistake rate when predicting credit default or recidivism, which is probably an underestimation—and working from this baseline, the same sample required that machines should have a mistake rate lower than 10 percent (Rebitschek et al. 2021). An even stronger bias exists in the domain of autonomous driving (Liu et al. 2019b, Shariff et al. 2021), where people require their self-driving car to be significantly safer than they themselves are, while substantially overestimating the safety of their own driving. In a representative sample of US drivers, the median respondent believed to be in the top 25 percent of drivers, and estimated that two-thirds of car crashes would be avoided if everyone drove like them. From this baseline, they required very high safety from self-driving cars, way above the actual average safety of human drivers.

Similar findings are available for other, less quantifiable aspects of human performance. For example, one of the main concerns that Americans have about implicit moral machines is that they do not understand nuance and complexity as well as humans do (Smith 2019). This concern translates into resistance to medical AI: because patients think their unique characteristics and circumstances will be poorly understood by AI, they prefer to turn to human doctors (Longoni et al. 2019)—leaving unexamined the actual ability of human doctors to take into account these unique characteristics and circumstances. In like vein, people express concerns about the transparency or intelligibility of medical AI recommendations, compared to that of human doctors, without realizing that they overestimate their ability to understand human doctors in the first place (Cadario et al. 2021).

2.1.2. Bias. Because implicit moral machines can harm people through their mistakes, people are rightly concerned about how many mistakes they make. But the distribution of these mistakes also matters. Beyond how many mistakes they make, it matters whether credit-scoring algorithm makes more mistakes about women than men (Bono et al. 2021, Hassani 2021); it matters whether self-driving cars are less likely to detect and protect pedestrians than other road users (Combs et al. 2019); and it matters whether face recognition algorithms are more likely to misclassify dark-skinned faces (Birhane 2022, Buolamwini & Gebru 2018). The nature of the mistakes matters, too. In a landmark investigation (Angwin et al. 2016), the news organization ProPublica published evidence of a racial bias in the results of the COMPAS algorithm, which is used in some US courts to predict (among other outcomes) the risk that a defendant be rearrested in the next two years. The key result of the analysis was that while the algorithm made the same number of mistakes for black defendants and for white defendants, it did not make the same mistakes—mistakes

which were favorable to the defendant were more likely when the defendant was white, and mistakes which were unfavorable to the defendant were more likely when the defendant was black. Comparable results were later found for white versus hispanic defendants (Hamilton 2019).

This analysis was probably the catalyst for a surge of interest in the design of algorithms whose outcomes satisfy some mathematical definition of fairness across individuals and groups. Much of this literature on algorithmic fairness is grounded in computer science and impossibility theorems, dealing with the problem that there are many possible mathematical definitions of fairness, whose requirements are sometimes impossible to achieve simultaneously (for entry points, see Chouldechova 2017, Kleinberg et al. 2017, Pleiss et al. 2017). Given that not all forms of fairness are simultaneously achievable, it may seem natural to collect experimental data on the forms of fairness that people prefer. This experimental work is mostly disconnected from moral psychology (for a review, see Starke et al. 2022), and its results seem to be highly dependent on the application domain considered in each article. For example, people seem to prefer simple demographic parity when considering university admission algorithms, that is, to require similar admission rates for all demographic groups of applicants (Srivastava et al. 2019); when algorithms decide whether to grant bail to defendants, people prefer that they equalize false positive rates across groups, rather than accuracy across groups (Harrison et al. 2020); and when algorithms decide how to allocate loans, people prefer that they adopt some calibrated version of fairness that prioritize applicants with the highest payback rates (Saxena et al. 2020).

In view of this variation in findings across experimental protocols and domains of application, there seem to be great opportunities for designing methodologically systematic, psychology-driven programs about the kind of fairness people want from machines whose decisions can have disparate impact across groups. In parallel, research is needed to better understand the concerns that people have about algorithmic fairness. At first sight, there are plenty of reasons to expect people to feel deep concern. First, there is ample discussion in the media about the danger that machines will learn, amplify and legitimize the biases embedded in the human decisions they are trained from (O'neil 2017). Second, people may consider that machines are more homogeneous than humans; that is, that a machine being biased is a sign that all machines are comparably biased (Longoni et al. 2022). Third, people may expect machines to not only inherit biases from humans, but also the difficulty of fixing them—perhaps underestimating our ability to reprogram machines, given our relative inability to reprogram humans (Mullainathan 2019).

Experimental results, however, suggest that people do not feel especially outraged when machines discriminate, or at least not as outraged as they would feel if humans discriminated (Bigman et al. 2022, Hidalgo et al. 2021). There is also a growing body of evidence suggesting that the very groups that feel at risk of biased human decisions may be the least averse to letting machines make decisions—seemingly because they are worried enough about the current decisions of humans to be willing to take a chance with machines (Bigman et al. 2021, Fumagalli et al. 2022, Jago & Laurin 2022, Pammer et al. 2021, 2023).

If these results are confirmed, they may create conflicts about how best to listen to the voice of the groups who are currently experiencing discrimination. When making the decision to deploy implicit moral machines, it is ethical to take into account the preferences of the persons who might be adversely and disparately impacted by the machines, and to trust their lived experience of discrimination. But in the context of algorithmic decisions, we may also need to be mindful of the knowledge that non-experts have acquired, and

whether this knowledge is sufficient to express an informed opinion. In this space, there is a great need for clear, interactive simulations and visualizations that can help people 'choose their own algorithm' and get first-hand experience of how implicit moral machines may affect them (Hao & Stray 2019).

2.1.3. Blame and other reactions to harm. So far we have focused on people's requirements and expectations when it comes to letting implicit moral machines make consequential decisions. We now turn to people's reactions when machines do not meet their expectations, compared to their reactions when human agents do not meet expectations. When human agents make harmful mistakes, other humans experience a manifold of negative reactions about the agent. Depending on how bad the mistake was, whether it was preventable, and whether it might have been intentional, people experience emotions such as anger and outrage, place responsibility and blame on the agent, and consider whether to punish the agent or terminate their employment (Cushman 2015, Malle et al. 2022). But do they experience the same emotions about machines, and if so, to a greater or lesser extent? It may seem bizarre, from a rational perspective, to be angry at a machine, to hold it responsible, or to blame it for the outcome of its decision—our anger means nothing to machines, nor our punishments. But from a psychological perspective, people do seem to experience toward machines the same manifold of negative reactions they experience toward humans, perhaps because the machines are perceived as autonomous enough to warrant these reactions (Bigman et al. 2019, Epstein et al. 2020, Franklin et al. 2022).

In fact, when implicit moral machines make mistakes, people may experience stronger reactions than when humans make comparable mistakes. This phenomenon is clear in the domain of automated driving, across many experiments comparing people's reactions to crashes caused by human drivers, and their reactions to crashes caused by self-driving cars (Franklin et al. 2021, Hidalgo et al. 2021, Hong et al. 2020, Liu & Du 2022, Liu et al. 2019a). All other things being equal, people judge crashes as more severe and less acceptable when they are caused by self-driving cars, and place more blame and responsibility on a self-driving car causing a crash, than on a human causing a comparable crash. It is not clear yet whether this pattern generalizes to other domains (Lima et al. 2021, Srinivasan & Sarial-Abi 2021). In particular, we already mentioned that people experience stronger negative reactions when humans discriminate, than when machines do the same (Bigman et al. 2022, Hidalgo et al. 2021)—perhaps because people are angry at the idea that human discrimination may be intentional, while they do not hold the same suspicion about machine discrimination.

While it is theoretically and methodologically interesting to compare the blame incurred by humans and machines that make the same mistake, it may be more realistic to investigate situations in which human and machine jointly produce a mistake. Indeed, there may not be many situations (other than fully autonomous driving) where machines are allowed to make dangerous decisions without any human supervision. Since there will almost always be a human in the same loop as the machine, mistakes will most often be the result of a joint failure of human and machine—so, how do people allocate responsibility and blame between humans and machines, when both contributed to a harmful mistake?

Once more, the bulk of the available evidence comes from the domain of automated driving. Recall that people were less severe toward humans who caused a crash, than toward machines that caused a comparable crash. Remarkably, this pattern reverses when human and machine jointly produce a crash (Awad et al. 2020b, Beckers et al. 2022, Liu

et al. 2021, Wotton et al. 2022). For example, when a semi-autonomous vehicle and its human driver-in-the-loop both fail to steer away from a pedestrian, people typically blame the human the most for the resulting collision. It is not yet clear why people blame machines more than humans when they fail alone, only to blame humans more than machines when they fail together. In any case, it would be useful to collect data in other domains than self-driving cars, in order to assess the transportability of this blame reversal effect (Shank et al. 2019).

2.2. Explicit Moral Machines

Implicit moral machines do not attempt to solve moral dilemmas—explicit moral machines do. Indeed, explicit moral machines either solve moral dilemmas as their main function, or they are susceptible to encounter moral dilemmas in some situations, and must accordingly be equipped to solve these dilemmas when they arise. Some moral dilemmas take the form of a conflict between two ethical principles. For example, a machine that performs content moderation online, at a scale or speed that prevents continuous human oversight, may have to routinely arbitrate between the value of free speech and the duty to suppress offensive or harmful content. In another context, a medical AI may have to arbitrate between immediately providing its best diagnosis even though it cannot explain its reasoning to humans, or recommending further tests to improve explicability, at the risk of delaying a time-sensitive diagnosis. It is very common in AI ethics to provide lists of moral values or ethical principles that AI should simultaneously pursue (e.g., beneficence, privacy, dignity, transparency) but it is far less common to provide guidance on what machines should do when these values are in conflict (Mittelstadt 2019, Morley et al. 2020). One reason is that broad ethical principles such as 'dignity' and 'privacy' are hard to quantify, making it difficult to operationalize their tradeoffs in policy guidelines as well as in experimental work (but see Kozyreva et al. 2023, Nussberger et al. 2022). Perhaps as a result of this difficulty, the psychological literature on explicit moral machines has mostly focused on another kind of dilemma, one that seems more amenable to experimental investigation.

This second kind of moral dilemma typically concerns the allocation of a scarce resource, with detrimental consequences to the humans who are un-prioritized in the allocation decision. Consider for example the problem of kidney paired donation (Freedman et al. 2020). A large share of kidney transplants involve a living donor, who is usually a spouse or a relative of the candidate, but all too frequently, the potential donor is a poor match for the candidate they volunteered to help. In such a situation, one solution is to enter all candidates and prospective donors in a database, which is then fed to an algorithm that seeks 2-way, 3-way, or even more complex chains of donations, so that as many candidates as possible find a compatible donor. This algorithm does not simply seek to maximize the number of donations, though, but uses a complex priority scheme that balances many factors such as the age of the candidates and how long they have been registered in the program, their travel distance to the transplantation center, or their baseline chance to find a donor in the general population. The machine must engage in tradeoffs between all these factors in order to decide who will receive a kidney, and who will remain on the waiting list.

Consider now the example of autonomous vehicles (AVs), for which the scarce resource to allocate is road safety. As implicit moral agents, AVs are expected to lead to an absolute increase in road safety; but AVs are also explicit moral agents in the sense that every action they take can redistribute relative levels of safety between the road users that surround

them (Goodall 2016, Bonnefon et al. 2019). This is illustrated in Figure 1 (left), in which the lateral position of the AV redistribute relative safety between the cyclist to its left, its own passengers, and the truck driver to its right. In a more extreme example (right), a collision is unavoidable, and the AV must decide whether to save its passenger or a pedestrian (Bonnefon et al. 2016). In both situations, the AV must be endowed with the ability to make a moral calculation about whose safety should take priority.

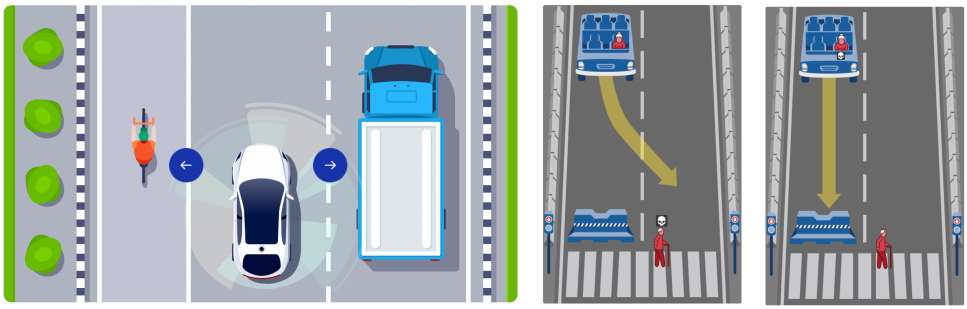

Figure 1

Examples of dilemmas faced by autonomous vehicles (AVs) as explicit moral agents. Left: Depending on its lateral positioning, the AV redistribute safety between the cyclist, the AV's passenger, and the truck driver. Right: In case a collision is unavoidable, the AV may have to decide who to save, for example its passenger or a pedestrian.

While people are often uncomfortable with the idea of letting machines make moral decisions (Bigman & Gray 2018, Dietvorst & Bartels 2022, Shariff et al. 2017), there is a case to be made that it is good to let machines solve moral dilemmas, even and especially when their decisions have unavoidably tragic consequences. If we agree that making such decisions inflict an emotional cost on the decision-maker, both in the short and the long term, we may agree that it is good to delegate this burden to machines, who do not experience psychological suffering (Danaher 2022). In like vein, we know that humans will inevitably be blamed for the way they solved a moral dilemma, since by definition a moral dilemma has no universally accepted solution—hence, we may want to relieve human decision-makers from unavoidable blame, by delegating the decision to a machine (see Textbox **Blaming machines for solving dilemmas** for further discussion).

If explicit moral machines are to make moral tradeoffs, we need to provide them with the goals and priorities they should pursue. This challenge is part of the value alignment problem (Gabriel 2020): To ensure that machines solve moral dilemmas in a way that is compatible with the goals and priorities of humans, we need to know what these human goals and priorities are, and to find a way to teach them to machines. Here we are concerned with the first objective, which falls squarely within the purview of moral psychology. We consider in turn some specific difficulties that moral psychologists face when collecting human moral preferences for the purpose of teaching them to machines: who to ask, how to ask, and what to do with the answers.

2.2.1. Value alignment: Who to ask. Not everyone agrees about what should be done in a moral dilemma, or what values should take priority in a moral tradeoff. So, who do we ask for their moral preferences, when we want to inform the decisions of machines? A

BLAMING MACHINES FOR SOLVING DILEMMAS

Humans will be blamed for the way they solve a moral dilemma, whatever they do. The same holds for machine, only with a twist: the blame incurred by a machine may not be distributed across possible decisions the same way it is distributed across human decisions. For example, in classic dilemmas where an agent must decide whether to save several lives by sacrificing one, humans are blamed more when they choose to sacrifice one, but this pattern is eliminated or even reversed when a machine makes the decision (Malle et al. 2015, Komatsu et al. 2021). This implies that delegating difficult moral decisions to machines may not merely remove a psychological burden from humans, but also change social expectations about which decision should be made (Gill 2020).

good place to start is to ask ethicists, who are trained to think about these issues, and have a deep understanding of their implications. Ethicists, however, are not immune to biases (Schwitzgebel & Cushman 2015), and they do not always come to an agreement—for example, a German national ethics committee could not reach a consensus on what an autonomous vehicle should do when deciding whether to save its passengers or to save other road users (Luetge 2017). There is another expert group we can ask for their preferences, namely, the people who build the machines, and who have a detailed understanding of what AI can actually do. For example, we could ask the autonomous vehicle industry what they believe AVs should do when deciding to save their passengers or other road users. One problem is that the industry is very reluctant to engage in this debate (Martinho et al. 2021), since any position they take may alienate either their consumer base or the general population. Experts from the AV industry may also feel a duty to protect their customers, which could explain why they have a stronger preference to save passengers, compared to the general population (Zhu et al. 2022).

Asking AI developers and ethicists for their informed preferences is important in view of their expertise, but we also need to document the preferences of the laypersons who will adopt the technology. Consider again the dilemma of an AV which needs to decide whether to prioritize the life of its passengers or that of other road users. However rare this dilemma might be, it weighs heavily with the minds of consumers, to the point of being cited as one of the top issue that will determine their decision to adopt AVs (Gill 2021). In this context, learning about consumers' preferences is not merely a marketing exercise. The main promise of AVs is that they can reduce the number of road casualties, by being safer than human drivers; but these lives will not be saved if consumers opt out of the technology because they are unsatisfied or even outraged with the way AVs solve moral dilemmas (Bonnefon et al. 2020a, De Freitas & Cikara 2021). As a result, learning the moral preferences of consumers may be a prerequisite for explicit moral machines to deliver their benefits. Explicit moral machines do not only impact the outcomes of their adopters, though. By design, they can create externalities for other stakeholders. For example, AVs do not merely affect the safety of their passengers, but also distribute risk to all road users around them. As a result, other road users (including pedestrians) should be given a voice when collecting preferences about the moral priorities that determine the behavior of AVs.

In sum, value alignment requires to collect human moral preferences to inform the behavior of explicit moral machines—a process that requires to decide whose values will be

collected, and how they should be weighted when different groups have different preferences. These normative issues are complex, but they arguably fall beyond the purview of moral psychology. Moral psychologists have an important role to play, however, in bringing their expertise to the matter of how best to measure moral preferences about explicit moral machines.

2.2.2. Value alignment: How to ask. Measuring moral preferences is never easy, and measuring preferences about explicit moral machines comes with its own set of challenges. First, explicit moral machines may need to balance a great number of conflicting values or priorities. For example, kidney paired donation algorithms may balance up to a dozen priorities, including the quality of the match between donor and candidate, the statistical rarity of potential donors for a given candidate, the age of the candidate at registration in the program, as well as their waiting time in the program, the blood types of donor and candidate, or the candidate having donated a kidney themselves. When distributing risk around them, AVs may need to consider the number of potential victims, their mode of transportation, their age, whether they are currently on the road or the sidewalk, and yet other variables. The high-dimensional nature of these choices may lead to an exploding number of experimental treatments, resulting in a need for an unpractical number of research participants. The Moral Machine Experiment (Awad et al. 2018) considered nine possible priorities for AVs to decide which group of road users to save or to sacrifice, which led to millions of possible scenarios. Exploring this enormous space was only possible because the experiment went viral, collecting data from millions of participants. Not every experiment can go viral, though, which means that moral psychologists have difficult choices to make when deciding how complex they want their scenario space to be.

Many other design choices will impact the feasibility of such experiments, and the interpretation of their results. For example, the Moral Machine Experiment purposefully used stylized scenarios when a collision is unavoidable (Awad et al. 2020a), but more realistic scenarios would have manipulated the probability of the collisions (Krügel & Uhl 2022). Participants were asked what the AV should do, but other questions can lead to different results, for example asking participants what AV behavior they would prefer as passengers (Bonnefon et al. 2016, Liu & Liu 2021, Takaguchi et al. 2022), or from another road user perspective (Mayer et al. 2021, Martin et al. 2021), or from under a veil of ignorance (Huang et al. 2019). Other experiments may opt out of asking participants to state their preferences, and try instead to reveal their preferences, by placing them in a virtual environment where they need to make themselves the same moral decisions that AVs will face (Faulhaber et al. 2019, Samuel et al. 2020). Given the relative novelty of explicit moral machines as a topic of investigation for moral psychology, the field may be best served by embracing this diversity of methods and designs, in order to build a comprehensive description of the moral values that people may want to see embedded in machines. This inclusive approach is especially important in view of what we will do with these data, as we discuss in the next and final section.

2.2.3. Value alignment: What to do with answers. It seems consensual to say that moral psychology has an important descriptive role in documenting the values and priorities that laypersons would want explicit moral machines to pursue (Awad et al. 2022). What is much more controversial is to decide what prescriptive weight these data should have in the policies that will regulate the behavior of the machines. Clearly, no one wants these

policies to be driven solely by the preferences of laypersons—but should these preferences be discarded entirely?

A promising approach to that question is to jointly consider the degree of consensus or division among experts, and the degree of consensus or division among laypersons (Savulescu et al. 2021). Consider first the situation where experts show strong consensus about what a machine should do. If laypersons show the same consensus, the case is closed. If laypersons are divided about what the machine should do, then the proper course of action is probably to follow the expert consensus while building up a strong and clear case for this consensus, in terms that the public can understand. If laypersons show a strong consensus against the consensus of the experts, the situation is more difficult, but it is also possible that the public consensus is based on bias more than reason, which is something that moral psychologists are equipped to show.

But consider now the situation where experts themselves are divided, and where this division reflects a reasonable moral disagreement. In that case, it may be appropriate to follow the public consensus, if there is one. But this requires to be very careful about establishing this consensus, and making sure it does not reflect, for example, the biases and prejudices of the majority. This is why we believe it is especially important for moral psychologists to explore an exhaustive range of methods and controls, to make sure that the public consensus is robust across experimental designs and demographics, as well as free of prejudice and bias, before it is allowed to arbitrate over the disagreements of experts.

3. MACHINES AS MORAL PATIENTS

So far, we have considered situations where machines are (implicit or explicit) moral agents, that is, situations in which machines perform actions whose consequences affect people. We will now flip the table, and consider situations in which machines are moral patients, that is, situations in which people perform actions that affect machines. This may sound strange, since machines have no affects, nor needs or desires for anything. Even though people are well aware of this, they can still feel empathy for machines (see Textbox **Empathy for the Machine**), or consider that machines 'want' things, in a certain sense, things that can be given or denied. In other terms, and as we will consider in more detail in the rest of this section, people sometimes assume that machines have preferences—which can turn machines into moral patients, who experience preferred or dis-preferred outcomes as a result of the actions taken by other agents (Pauketat & Anthis 2022).

This is especially important when people have an opportunity to cooperate with a machine. Cooperative interactions with intelligent, autonomous machines are not yet a common experience, but this is likely to change in the future. Cooperation with machines is already a reality in industry settings (Villani et al. 2018), and soon enough, road users will have to cooperate with autonomous vehicles to make traffic safe for everyone (Schwarting et al. 2019). Many participants in online communities or social networks already have with bots the same kind of cooperative (or uncooperative) interactions that they have with humans (Seering et al. 2018, Shao et al. 2018, Tsvetkova et al. 2017, Stella et al. 2018): People and bots can retweet or block one another; Reddit users sometimes congratulate bots for good behavior, but sometimes report them to moderators; and Wikipedia editors can cooperate with bots on an article, or engage in an editing war against one another. As interactions with intelligent machines become more commonplace, how will humans and machines initiate and sustain cooperation?

EMPATHY FOR THE MACHINE

While people understand that robots do not experience physical pain or psychological distress, they can nevertheless feel emotionally uncomfortable when humans direct toward robots the kind of behavior that would qualify as abuse if directed toward other humans. For example, research participants show physiological signs of discomfort when watching a baby dinosaur robot being punched and choked (Rosenthal-von der Pütten et al. 2013), or a robot hand being cut by a knife (Suzuki et al. 2015); they hesitate when asked to strike a robot (Darling et al. 2015), or to topple a block tower that a robot built and pretends to care about (Briggs & Scheutz 2014); and they are likely to ask a research confederate to stop when they see the confederate insulting and roughing up a robot (Connolly et al. 2020).

Mutually beneficial cooperation between humans often rely on a positive concern for the outcomes of others—a preference for the satisfaction of the preferences of others. Cooperation is easier if my other-regarding preferences are prosocial, that is, if I derive some measure of satisfaction from doing good to others. Conversely, cooperation is usually more difficult if my other-regarding preferences are antisocial or even just callous—that is, if I derive satisfaction from doing ill to others, or if I am entirely indifferent about what happens to others, and only care about my own outcomes. But what happens when humans have an opportunity to cooperate with machines? What are their machine-regarding preferences? This is the topic of the next section.

3.1. Machine-Regarding Preferences

Cooperation between humans does not necessarily involve money. People can volunteer their time and skills to help others, provide advice, share tools, advocate for a cause, or donate blood. While it is possible to study all these currencies in behavioral experiments that investigate cooperation, experimental economics has popularized the assumption that it is possible to capture the manifold of human cooperation by using lab-based games with financial incentives, such as dictator games, prisoners' dilemmas, ultimatum games, or public good games. Incentivized games provide a controlled, stylized environment to measure other-regarding preferences and prosocial behavior, that allows for easy comparison of studies and experimental treatments. As a result, many studies of human-machine cooperation have used the same games, only replacing some human players by intelligent machines, in order to document changes in human behavior when playing incentivized games with machines, as compared to humans (March 2021). These studies carry over the assumption that just as money can be used as a proxy for the many currencies of human-human cooperation, it can be used as a proxy for the many currencies of human-machine cooperation. In the rest of this section, we proceed with this assumption—but see Textbox **What do machines do with money?** for a closer examination.

Findings on human-machine cooperation in incentivized games show remarkable convergence. In a nutshell, people do show some measure of prosocial machine-regarding preferences, and cooperation does not disappear when humans play with machines—but it does not reach the level of human-human cooperation. In other words, all findings suggest the existence of a machine penalty in cooperative games. For example, in a one-shot trust

WHAT DO MACHINES DO WITH MONEY?

If you had to split some money between yourself and, say, a tree, you would probably wonder about what happens to the money you give to the tree, since trees have no use for money. The same question holds in experiments where people share money with machines, or help machines make money. Presumably, what truly happens in most cases is that the money earned by the machine goes back to the research fund of the experimenters—but this is not usually made clear to research participants. Indeed, in a survey of 160 experiments, von Schenk et al. (2022) observed that 82% of instructions did not give any explanation about what happened to machine earnings. (The rest was split between pretending that machines would keep the money, reminding that machines had no use for money, and explaining that the machine earnings would actually be transferred to a human.) So, if people wonder what machines could use money for, and if experimenters have no answer to offer, is money the right currency to study human-machine cooperation? There are two arguments for believing so. First, it is not like any other currency would be better, since machines do not care about anything, in the sense of feeling a desire or a need for something. Second, people seem to agree that machines still 'want' money, in the sense of being programmed to do so, to the same extent that they 'want' retweets or other cooperative currencies used in online communities (Makovi et al. 2023). As a result, money in incentivized games seems an acceptable proxy for the currencies used in real-life human-machine cooperation.

game, human second-movers expected the same level of cooperation from human and machine first-movers, but only 34% reciprocated the trust of a machine, compared to 75% who reciprocated the trust of a human; and likewise, in a one-shot prisoners' dilemma, people expected the same level of cooperation from humans and machines, but cooperated with only 36% of machines, compared to 49% of humans (Karpus et al. 2021). In a one-shot dictator game, people allocated 39% of their endowment to a human, but only 16% to a machine; and in a one-shot public good game, people contributed about 55% of their endowment to the common pool when playing with humans, but only 40% when playing with machines (Nielsen et al. 2022b).

One-shot games thus suggest that people do not initiate cooperation with machines to the same level they initiate cooperation with humans: cooperation does not drop to zero, but it suffers from a machine penalty. Repeated games allow to study the dynamics of the machine penalty, and its evolution through repeated interaction (Crandall et al. 2018). Findings suggest that the machine penalty carries unchanged over repeated interactions, but their interpretation can be complicated by the fact that in repeated games, human decisions can be impacted by the strategy chosen by the machine, which may be different than the strategies commonly adopted by humans (Sandoval et al. 2016). One way to address this difficulty is to use deception, that is, to pair players with humans they believe to be machines, or to pair them with machines they believe to be humans. Such deception allows to measure the mere effect of believing that one's partner is human or machine, independently of the strategy adopted by the partner. Figure 2 displays the results of one such experiment (Ishowo-Oloko et al. 2019), in which human players either knew each other to be humans, or believed each other to be machines. As is common with repeated prisoners' dilemmas, cooperation steadily decreases over time when both players know each other to be humans. When both players believe each other to be machines, the negative

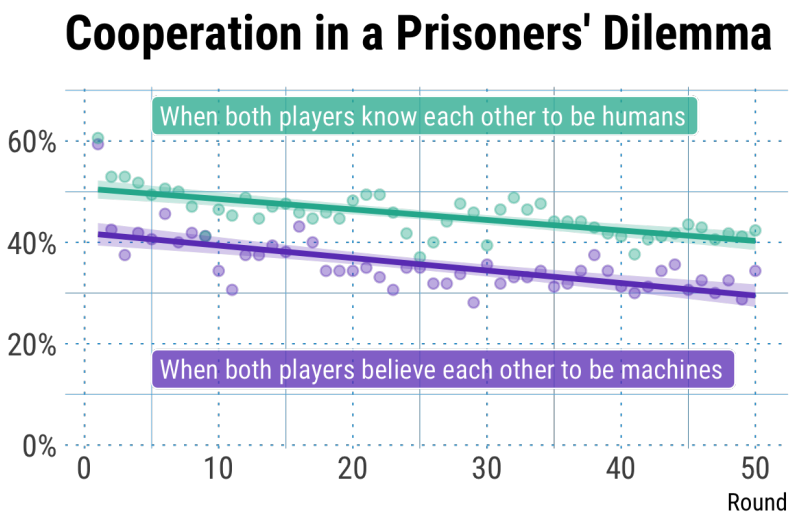

Figure 2

Over 50 rounds, cooperation between two human players in a prisoners' dilemma steadily decreases. The dynamics is the same when the players believe each other to be machines, hence the parallel regression lines, but the machine penalty carries over time, hence the vertical distance between the two lines. Data replotted from the source file of Ishowo-Oloko et al. (2019).

dynamics is very similar, and the machine penalty carries over unchanged over time.

3.2. Overcoming the Machine Penalty

The machine penalty is not only a phenomenon we need to understand, it is also arguably a problem we must solve. For the last 20 years, celebrated milestones in AI research were often tied to competition against humans—be it when IBM Watson defeated the two highest-ranked *Jeopardy!* players (Ferrucci et al. 2010), or when DeepMind's AlphaGo defeated the top Go player Lee Sedol (Silver et al. 2016). While surpassing human performance is an important goal of AI (in particular when it behaves as an implicit moral agent), there is an increasing recognition that in order to fulfill the true potential of AI, we need to put as much effort in human-AI cooperation as in human-AI competition. There is a technical side to this challenge, since it may require to design AI systems that understand and respond appropriately to human intentions and goals (Dafoe et al. 2021). But there is also a psychological side to the challenge, which requires to understand why humans are reluctant to cooperate with machines, and to design interventions that can overcome this machine penalty.

It is perhaps natural to start with interventions that give machines more human-like traits. After all, if people do not cooperate with machines as much as with humans, perhaps we can narrow the gap between cooperation rates by making machines look or feel more like humans. This humanization strategy may help people activate with machines the same cooperation templates they activate with humans, or the frames of reference they use to interpret the behavior of cooperation partners, thus increasing their trust and comfort in

this new situation (Nielsen et al. 2022a). We may then expect the humanization strategy to increase in efficacy when machines are humanized to a large degree, as compared to when machines are humanized to a minimal degree. Experimental findings, however, tell a more complicated story.

Minimally humanized robots typically fail to elicit more cooperation than non-humanized robots. Examples of minimal humanization include giving the robot an ovoid shape augmented with eyes, as compared to an insectoid appearance (De Kleijn et al. 2019); or endowing a non-humanoid robot with some emotional displays, such as stylized angry, sad or happy eyes, as well as recorded sighs and laughter (Hsieh & Cross 2022). These experiments do not report significant effects on cooperation, suggesting that minimal humanization is insufficient to overcome the machine penalty. Climbing up the humanization gradient does not improve cooperation much, and can even make things worse, due to the uncanny valley effect—that is, the feeling of strangeness and discomfort elicited by a machine that is largely but not quite human-like. For example, a study using 80 robotic faces going from entirely machine-like to entirely human-like found that cooperation was at its *lowest* for machines that placed at two-thirds of the humanization gradient (Mathur & Reichling 2016), and other studies showed that even more human-like robots failed to eliminate the machine penalty (Złotowski et al. 2016, Cominelli et al. 2021). Intriguingly, the few studies that succeeded in reducing the machine penalty through (moderate) humanization did so by gendering the machine as female, through stylized cues such as suggestions of long hair or breasts (Bernotat et al. 2021, Eyssel & Hegel 2012). While this strategy may indeed prove somewhat efficient at reducing the machine penalty, it seems ethically problematic to exploit and perpetuate gender stereotypes about women being less competitive or more nurturing, just as it would seem problematic to systematically give AI assistants a female voice (Fossa & Sucameli 2022).

All humanization strategies we reviewed so far were non-deceptive, in the sense that while the machine was made more human-like, it was never described as being human. If we remove that constraint, we reach the highest possible level of humanization, machines that pretend to be humans. This is done easily enough in most experimental protocols that use incentivized games, since these protocols are usually designed to remove all visual or verbal interactions between players. If players are identified with headshots, machines can create synthetic faces for themselves, faces which can be both realistic and especially trust-inducing (Nightingale & Farid 2022). Unsurprisingly, this deceitful form of humanization eliminates the machine penalty (Ishowo-Oloko et al. 2019): If people do not know they are cooperating with machines, they do not manifest the machine penalty. Once more, though, this solution creates ethical issues, since AI codes of ethics typically emphasise that machines should never be allowed to pass as humans (O'Leary 2019).

In sum, humanization strategies usually fail to reduce the machine penalty, and the ones that succeed (partially or totally) fall short of current ethical standards. As a result, there is a need for further research that would seek to improve human-machine cooperation without resorting to the humanization of machines. One promising direction may be to embrace the fact that intelligent machines are newcomers in our social and cooperative interactions, and to accept that dealing with these newcomers may require new social norms (Makovi et al. 2023). In other words, rather than making machines more human-like in the hope that people will apply to them the old social norms they apply to humans, we could experiment on the new social norms that will develop around the new entrants in our social world, intelligent machines.

4. MACHINES AS MORAL PROXIES

By design, AI enables machines to make autonomous decisions on behalf of human stakeholders. This raises the possibility of delegating unethical behavior, in a way that distances the human from the act. AI also offers a further possibility, namely of *mediating* human communication in a morally-relevant manner. We explore each of these possibilities in turn.

4.1. Delegation to Machines

People delegate a growing number of tasks to AI agents (de Melo et al. 2018). Current and near-term possibilities are as diverse as setting prices in online markets (Calvano et al. 2020a), interrogating suspects (McAllister 2016), and marketing to customers (Cheng & Jiang 2022). This creates many opportunities to delegate unethical behavior to machines.

First, AI can be used by people who have malicious intentions to scale up criminal or unethical behavior. Recent advancements in deep learning, specifically Generative Adversarial Networks (GANs), have made it easier to create fake content that looks genuine (Caldwell et al. 2020). Those who have malicious intentions can benefit from using AI hench-agents because AI can act independently and has the potential to cause harm with unparalleled efficiency and at scale. Moreover, these AI hench-agents may be harder to trace back to the original source. AI-powered deepfakes can create fake identities, which allows phishing attacks to become more personalized and effective. These attacks, also known as spear phishing (Seymour & Tully 2016), put a new spin on identity theft, and can have devastating results (Jagatic et al. 2007). Reflecting on this emerging worry, a panel of experts has nominated deepfakes as the most dangerous tool for AI-enabled crime (Caldwell et al. 2020).

Delegation of criminal or ethically questionable behavior to AI agents might be attractive for reasons other than scalability. When people delegate tasks to AI agents instead of humans, it creates a combination of psychological factors that can lead to unethical behavior, such as anonymity (Ostermaier & Uhl 2017), psychological distance from victims (Köbis et al. 2019), and undetectability (Hancock & Guillory 2015, Rauhut 2013). The often-incomprehensible workings of algorithms create ambiguity (Miller 2019). Letting such "black box" algorithms execute tasks on one's behalf increases plausible deniability, and obfuscates the attribution of responsibility for the harm caused. If any harm does become apparent, blame and responsibility can be deflected to the delegate, which may alleviate the (legal or psychological) guilt experienced by the remitter. Indeed, people tend to prefer delegation, even if it entails explicit instruction to break ethical rules, such as when using henchpersons (Drugov et al. 2014).

Ambiguity is another mechanism through which unethical behavior can be delegated to machines. More often than not, people do not explicitly instruct their delegates to break ethical rules but instead merely define their desired outcome and turn a blind eye to how it's achieved. By doing so, the remitter avoids direct contact with the victims and can willfully ignore, through deliberate ignorance (Hertwig & Engel 2016), any possible ethical rule violations that may occur as a result of the delegation (Drugov et al. 2014, Van Zant & Kray 2014).

Delegation to AI may also cause moral violations without any bad intent (Thomas et al. 2019). For example, someone may use algorithmic prices to sell goods on online markets, without being aware that algorithms might coordinate and set collusive prices (Calvano et al. 2020b, Wellman & Rajan 2017). Marketers who rely on AI-powered sales strategies

might be unaware of the fact that the AI agent employs deceptive tactics to reach sales goals.

Not all delegation is bad, of course. One may indeed delegate morally desirable actions to AI agents. Specifically, delegating morally desirable actions, such as charitable donations, to an AI agent may act as a commitment device (Bryan et al. 2010) that increases the magnitude and frequency of such actions. There are also opportunities to delegate an advisory role to AI agents, enabling them to dynamically suggest moral behavior to the human (Giubilini & Savulescu 2018).

4.2. Machine Masquerade

We close this article with our shortest and most speculative section. So far we considered the possibility for people to send a machine proxy in a moral interaction, in the sense that they delegate their decisions to the machine. In this final section, we consider the possibility for people to participate themselves in an interaction, only under a disguise provided by the machine. Under this machine masquerade (known as 'AI-mediated communication', Hancock et al. 2020), people use technology to modify the way they write, talk, or look, in order to change the behavior of their partner. Moral psychology has given little attention so far to AI-mediated communication, but this is likely to change given the incoming availability of machine masquerade tools, the way they will transform moral interactions, and the ethical challenges they raise.

Many people are already familiar with machine-generated replies to text messages or emails, as well as image filters that improve the appearance of the subject; but AI is poised to allow much more powerful and flexible forms of transformation. Written text, profile pictures, as well as voice and facial dynamics in live online interactions can already be altered to achieve various presentation goals. While not everyone will have immediate access to all these technologies (Goldenthal et al. 2021), their adoption can be very fast. Consider the case of OpenAI's ChatGPT, a (as of the writing of this article) state-of-the-art language model with a user-friendly interface that allows people to easily experiment with various prompts and requests. Within weeks of its public launch, ChatGPT attracted more than a hundred million of users, affording them seemingly endless possibilities. Students could use ChatGPT to sound more competent; business owners could use it to sound more trustworthy; and social media users could ask it to generate posts in line with the image they wished to project, or the moral virtues they wished to signal.

We know very little yet about how people will seize and judge these opportunities, at which scale, and to which effects. Existing work suggests that people who use machines to write for them are perceived as less trustworthy, in studies using hypothetical emails (Liu et al. 2022), hypothetical AirBnB profiles (Jakesch et al. 2019), and actual text conversations (Hohenstein et al. 2021); but there is much more to be done to understand the material and reputational benefits that people can achieve if their use of machines is not discovered, and how much of these benefits are conserved depending on the way the use of machines is disclosed. Compare for example a social media user who is posting content that they secretly asked a machine to generate in order to signal a commitment to gender equality; and a social media user who is disclosing on their profile that they are systematically asking a machine to alter their posts in order to remove gender biases. Such scenarios are no longer far-fetched, and we need moral psychology to understand the effects they will have, as well as the reactions they will trigger.

Machine masquerade is not restricted to written text—it can alter the way we look and the way we sound. People can already experiment with generative AI to create their profile pictures, and the technology to alter voices is already perfected. This means that people can ask machines to alter their face in order to appear more dominant or more trustworthy, or to alter their voice to sound more articulate or more cheerful (Guerouaou et al. 2022). These alterations can change the outcomes of moral interactions; but they can also raise new ethical issues for moral psychology to investigate, such as the conflict between reducing discrimination while threatening inclusion. For example, machines can remove the foreign accent of call center employees, which decreases the likelihood they will receive racist abuse from angry customers—but this can be construed as a step in the wrong direction, as it would amount to considering that the fix to racism is not to reduce prejudice, but to accommodate it by whitening the voice of its victims (Simpson 2022).

In sum, machine masquerade offers a vast new field of investigation for moral psychology, aimed at understanding how people will use technology to alter their presentation, either for the purpose of changing the outcomes of moral interactions or for the purpose of managing their moral reputation; how this processes may be moderated by different forms of disclosure; and how society will deal with the new ethical dilemmas raised by this technology.

5. CONCLUSION

We have not addressed every issue at the intersection of AI and moral psychology. Questions about how people perceive AI plagiarism, about how the presence of AI agents can reduce or enhance trust between groups of humans, about how sexbots will alter intimate human relations, are the subjects of active research programs. Many more yet unasked questions will only be provoked as new AI abilities develops. Given the pace of this change, any review paper will only be a snapshot. Nevertheless, the very recent and rapid emergence of AI-driven technology is colliding with moral intuitions forged by culture and evolution over the span of millennia. Grounding an imaginative speculation about the possibilities of AI with a thorough understanding of the structure of human moral psychology will help prepare for a world shared with, and complicated by, machines.

ACKNOWLEDGMENTS

JFB acknowledges support from grant ANR-19-PI3A-0004, grant ANR-17-EURE-0010, and the research foundation TSE-Partnership. AFS acknowledges support from a Canada 150 Research Chair grant from the Social Sciences and Humanities Research Council of Canada.

LITERATURE CITED

Angwin J, Larson J, Mattu S, Kirchner L. 2016. Machine bias. *ProPublica*

Awad E, Dsouza S, Bonnefon JF, Shariff A, Rahwan I. 2020a. Crowdsourcing moral machines. *Communications of the ACM* 63(3):48–55

Awad E, Dsouza S, Kim R, Schulz J, Henrich J, et al. 2018. The moral machine experiment. *Nature* 563(7729):59–64

Awad E, Levine S, Anderson M, Anderson SL, Conitzer V, et al. 2022. Computational ethics. *Trends in Cognitive Sciences*

Awad E, Levine S, Kleiman-Weiner M, Dsouza S, Tenenbaum JB, et al. 2020b. Drivers are blamed more than their automated cars when both make mistakes. *Nature human behaviour* 4(2):134–143

Beckers N, Siebert LC, Bruijnes M, Jonker C, Abbink D. 2022. Drivers of partially automated vehicles are blamed for crashes that they cannot reasonably avoid. *Scientific reports* 12(1):1–11

Bernotat J, Eyssel F, Sachse J. 2021. The (fe) male robot: how robot body shape impacts first impressions and trust towards robots. *International Journal of Social Robotics* 13(3):477–489

Bigman YE, Gray K. 2018. People are averse to machines making moral decisions. *Cognition* 181:21–34

Bigman YE, Waytz A, Alterovitz R, Gray K. 2019. Holding robots responsible: The elements of machine morality. *Trends in cognitive sciences* 23(5):365–368

Bigman YE, Wilson D, Arnestad MN, Waytz A, Gray K. 2022. Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*

Bigman YE, Yam KC, Marciano D, Reynolds SJ, Gray K. 2021. Threat of racial and economic inequality increases preference for algorithm decision-making. *Computers in Human Behavior* 122:106859

Birhane A. 2022. The unseen black faces of ai algorithms. *Nature* 610:451–452

Bonnefon J, Shariff A, Rahwan I. 2020a. The moral psychology of ai and the ethical opt-out problem, In *The ethics of artificial intelligence*, ed. SM Liao, pp. 109–126, Oxford University Press, Oxford

Bonnefon JF, Černy D, Danaher J, Devillier N, Johansson V, et al. 2020b. Ethics of connected and automated vehicles: Recommendations on road safety, privacy, fairness, explainability and responsibility. Brussels, Belgium: Publication Office of the European Union

Bonnefon JF, Shariff A, Rahwan I. 2016. The social dilemma of autonomous vehicles. *Science* 352(6293):1573–1576

Bonnefon JF, Shariff A, Rahwan I. 2019. The trolley, the bull bar, and why engineers should care about the ethics of autonomous cars [point of view]. *Proceedings of the IEEE* 107(3):502–504

Bono T, Croxson K, Giles A. 2021. Algorithmic fairness in credit scoring. *Oxford Review of Economic Policy* 37(3):585–617

Briggs G, Scheutz M. 2014. How robots can affect human behavior: Investigating the effects of robotic displays of protest and distress. *International Journal of Social Robotics* 6(3):343–355

Bryan G, Karlan D, Nelson S. 2010. Commitment devices. *Annu. Rev. Econ.* 2(1):671–698

Buolamwini J, Gebru T. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification, In *Conference on fairness, accountability and transparency*, pp. 77–91, PMLR

Cadario R, Longoni C, Morewedge CK. 2021. Understanding, explaining, and utilizing medical artificial intelligence. *Nature human behaviour* 5(12):1636–1642

Caldwell M, Andrews JT, Tanay T, Griffin LD. 2020. Ai-enabled future crime. *Crime Science* 9(1):1–13

Calvano E, Calzolari G, Denicolò V, Harrington Jr JE, Pastorello S. 2020a. Protecting consumers from collusive prices due to ai. *Science* 370(6520):1040–1042

Calvano E, Calzolari G, Denicolo V, Pastorello S. 2020b. Artificial intelligence, algorithmic pricing, and collusion. *American Economic Review* 110(10):3267–3297

Cheng Y, Jiang H. 2022. Customer–brand relationship in the era of artificial intelligence: understanding the role of chatbot marketing efforts. *Journal of Product & Brand Management* 31(2):252–264

Chouldechova A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5(2):153–163

Combs TS, Sandt LS, Clamann MP, McDonald NC. 2019. Automated vehicles and pedestrian safety: exploring the promise and limits of pedestrian detection. *American journal of preventive medicine* 56(1):1–7

Cominelli L, Feri F, Garofalo R, Giannetti C, Meléndez-Jiménez MA, et al. 2021. Promises and trust in human–robot interaction. *Scientific Reports* 11(1):1–14

Connolly J, Mocz V, Salomons N, Valdez J, Tsoi N, et al. 2020. Prompting prosocial human interventions in response to robot mistreatment, In *Proceedings of the 2020 ACM/IEEE international conference on human-robot interaction*, pp. 211–220

Crandall JW, Oudah M, Tennom, Ishowo-Oloko F, Abdallah S, et al. 2018. Cooperating with machines. *Nature communications* 9(1):1–12

Cushman F. 2015. Deconstructing intent to reconstruct morality. *Current Opinion in Psychology* 6:97–103

Dafoe A, Bachrach Y, Hadfield G, Horvitz E, Larson K, Graepel T. 2021. Cooperative AI: machines must learn to find common ground. *Nature* 593:33–36

Danaher J. 2022. Tragic choices and the virtue of techno-responsibility gaps. *Philosophy & Technology* 35(2):1–26

Darling K, Nandy P, Breazeal C. 2015. Empathic concern and the effect of stories in human-robot interaction, In *2015 24th IEEE international symposium on robot and human interactive communication (RO-MAN)*, pp. 770–775, IEEE

De Freitas J, Cikara M. 2021. Deliberately prejudiced self-driving vehicles elicit the most outrage. *Cognition* 208:104555

De Kleijn R, van Es L, Kachergis G, Hommel B. 2019. Anthropomorphization of artificial agents leads to fair and strategic, but not altruistic behavior. *International Journal of Human-Computer Studies* 122:168–173

de Melo CM, Marsella S, Gratch J. 2018. Social decisions and fairness change when people's interests are represented by autonomous agents. *Autonomous Agents and Multi-Agent Systems* 32:163–187

Dietvorst BJ, Bartels DM. 2022. Consumers object to algorithms making morally relevant tradeoffs because of algorithms' consequentialist decision strategies. *Journal of Consumer Psychology* 32(3):406–424

Drugov M, Hamman J, Serra D. 2014. Intermediaries in corruption: an experiment. *Experimental Economics* 17:78–99

Epstein Z, Levine S, Rand DG, Rahwan I. 2020. Who gets credit for ai-generated art? *Iscience* 23(9):101515

Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, et al. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 542(7639):115–118

Eyssel F, Hegel F. 2012. (s)he's got the look: Gender stereotyping of robots. *Journal of Applied Social Psychology* 42(9):2213–2230

Faulhaber AK, Dittmer A, Blind F, Wächter MA, Timm S, et al. 2019. Human decisions in moral dilemmas are largely described by utilitarianism: Virtual car driving study provides guidelines for autonomous driving vehicles. *Science and engineering ethics* 25:399–418

Ferrucci D, Brown E, Chu-Carroll J, Fan J, Gondek D, et al. 2010. Building watson: An overview of the deepqa project. *AI magazine* 31(3):59–79

Fossa F, Sucameli I. 2022. Gender bias and conversational agents: an ethical perspective on social robotics. *Science and Engineering Ethics* 28(3):1–23

Franklin M, Ashton H, Awad E, Lagnado D. 2022. Causal framework of artificial autonomous agent responsibility, In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 276–284

Franklin M, Awad E, Lagnado D. 2021. Blaming automated vehicles in difficult situations. *Iscience* 24(4):102252

Freedman R, Borg JS, Sinnott-Armstrong W, Dickerson JP, Conitzer V. 2020. Adapting a kidney exchange algorithm to align with human values. *Artificial Intelligence* 283:103261

Fumagalli E, Rezaei S, Salomons A. 2022. Ok computer: Worker perceptions of algorithmic recruitment. *Research Policy* 51(2):104420

Gabriel I. 2020. Artificial intelligence, values, and alignment. *Minds and machines* 30(3):411–437

Gill T. 2020. Blame it on the self-driving car: how autonomous vehicles can alter consumer morality. *Journal of Consumer Research* 47(2):272–291

Gill T. 2021. Ethical dilemmas are really important to potential adopters of autonomous vehicles. *Ethics and Information Technology* 23(4):657–673

Giubilini A, Savulescu J. 2018. The artificial moral advisor. the "ideal observer" meets artificial intelligence. *Philosophy & technology* 31:169–188

Goldenthal E, Park J, Liu SX, Mieczkowski H, Hancock JT. 2021. Not all ai are equal: exploring the accessibility of ai-mediated communication technology. *Computers in Human Behavior* 125:106975

Goodall NJ. 2016. Away from trolley problems and toward risk management. *Applied Artificial Intelligence* 30(8):810–821

Guerouaou N, Vaiva G, Aucouturier JJ. 2022. The shallow of your smile: the ethics of expressive vocal deep-fakes. *Philosophical Transactions of the Royal Society B* 377(1841):20210083

Hamilton M. 2019. The biased algorithm: Evidence of disparate impact on hispanics. *American Criminal Law Review* 56:1553

Hancock JT, Guillory J. 2015. Deception with technology. *The handbook of the psychology of communication technology* :270–289

Hancock JT, Naaman M, Levy K. 2020. Ai-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication* 25(1):89–100

Hao K, Stray J. 2019. Can you make ai fairer than a judge? play our courtroom algorithm game. *MIT Technology Review*

Harrison G, Hanson J, Jacinto C, Ramirez J, Ur B. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models, In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 392–402

Hassani BK. 2021. Societal bias reinforcement through machine learning: a credit scoring perspective. *AI and Ethics* 1(3):239–247

Hertwig R, Engel C. 2016. Homo ignorans: Deliberately choosing not to know. *Perspectives on Psychological Science* 11(3):359–372

Hidalgo CA, Orghian D, Canals JA, De Almeida F, Martín N. 2021. How humans judge machines. MIT Press

Hohenstein J, DiFranzo D, Kizilcec RF, Aghajari Z, Mieczkowski H, et al. 2021. Artificial intelligence in communication impacts language and social relationships. *arXiv preprint arXiv:2102.05756*

Hong JW, Wang Y, Lanz P. 2020. Why is artificial intelligence blamed more? analysis of faulting artificial intelligence for self-driving car accidents in experimental settings. *International Journal of Human–Computer Interaction* 36(18):1768–1774

Hsieh TY, Cross ES. 2022. People's dispositional cooperative tendencies towards robots are unaffected by robots' negative emotional displays in prisoner's dilemma games. *Cognition and Emotion* :1–25

Huang K, Greene JD, Bazerman M. 2019. Veil-of-ignorance reasoning favors the greater good. *Proceedings of the national academy of sciences* 116(48):23989–23995

Ishowo-Oloko F, Bonnefon JF, Soroye Z, Crandall J, Rahwan I, Rahwan T. 2019. Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence* 1(11):517–521

Jagatic TN, Johnson NA, Jakobsson M, Menczer F. 2007. Social phishing. *Communications of the ACM* 50(10):94–100

Jago AS, Laurin K. 2022. Assumptions about algorithms' capacity for discrimination. *Personality and Social Psychology Bulletin* 48(4):582–595

Jakesch M, French M, Ma X, Hancock JT, Naaman M. 2019. Ai-mediated communication: How the perception that profile text was written by ai affects trustworthiness, In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pp. 1–13

Kalra N, Groves DG. 2017. The enemy of good: Estimating the cost of waiting for nearly perfect automated vehicles. Rand Corporation

Kalra N, Paddock SM. 2016. Driving to safety: How many miles of driving would it take to demon-

strate autonomous vehicle reliability? *Transportation Research Part A: Policy and Practice* 94:182–193

Karpus J, Krüger A, Verba JT, Bahrami B, Deroy O. 2021. Algorithm exploitation: Humans are keen to exploit benevolent ai. *iScience* 24(6):102679

Kleinberg J, Lakkaraju H, Leskovec J, Ludwig J, Mullainathan S. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133(1):237–293

Kleinberg J, Mullainathan S, Raghavan M. 2017. Inherent trade-offs in the fair determination of risk scores, In *8th Innovations in Theoretical Computer Science Conference*

Köbis NC, Verschuere B, Bereby-Meyer Y, Rand D, Shalvi S. 2019. Intuitive honesty versus dishonesty: Meta-analytic evidence. *Perspectives on Psychological Science* 14(5):778–796

Komatsu T, Malle BF, Scheutz M. 2021. Blaming the reluctant robot: parallel blame judgments for robots in moral dilemmas across us and japan, In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, pp. 63–72

Kozyreva A, Herzog SM, Lewandowsky S, Hertwig R, Lorenz-Spreen P, et al. 2023. Resolving content moderation dilemmas between free speech and harmful misinformation. *PNAS* 120:e2210666120

Krügel S, Uhl M. 2022. Autonomous vehicles and moral judgments under risk. *Transportation research part A: policy and practice* 155:1–10

Lima G, Grgić-Hlača N, Cha M. 2021. Human perceptions on moral responsibility of ai: A case study in ai-assisted bail decision-making, In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–17

Liu P, Du M, Li T. 2021. Psychological consequences of legal responsibility misattribution associated with automated vehicles. *Ethics and information technology* 23(4):763–776

Liu P, Du Y. 2022. Blame attribution asymmetry in human–automation cooperation. *Risk Analysis* 42(8):1769–1783

Liu P, Du Y, Xu Z. 2019a. Machines versus humans: People's biased responses to traffic accidents involving self-driving vehicles. *Accident Analysis & Prevention* 125:232–240

Liu P, Liu J. 2021. Selfish or utilitarian automated vehicles? deontological evaluation and public acceptance. *International Journal of Human–Computer Interaction* 37(13):1231–1242

Liu P, Yang R, Xu Z. 2019b. How safe is safe enough for self-driving vehicles? *Risk analysis* 39(2):315–325

Liu Y, Mittal A, Yang D, Bruckman A. 2022. Will ai console me when i lose my pet? understanding perceptions of ai-mediated email writing, In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pp. 1–13

Longoni C, Bonezzi A, Morewedge CK. 2019. Resistance to medical artificial intelligence. *Journal of Consumer Research* 46(4):629–650

Longoni C, Cian L, Kyung EJ. 2022. Algorithmic transference: People overgeneralize failures of ai in the government. *Journal of Marketing Research* :00222437221110139

Luetge C. 2017. The german ethics code for automated and connected driving. *Philosophy & Technology* 30(4):547–558

Makovi K, Sargsyan A, Li W, Bonnefon JF, Rahwan T. 2023. Trust within human-machine collectives depends on the perceived consensus about cooperative norms. *Nature Communications*

Malle BF, Guglielmo S, Voiklis J, Monroe AE. 2022. Cognitive blame is socially shaped. *Current Directions in Psychological Science* 31(2):169–176

Malle BF, 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 *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 117–124, IEEE

March C. 2021. Strategic interactions between humans and artificial intelligence: Lessons from experiments with computer players. *Journal of Economic Psychology* 87:102426

Martin R, Kusev P, Van Schaik P. 2021. Autonomous vehicles: How perspective-taking accessibility alters moral judgments and consumer purchasing behavior. *Cognition* 212:104666

Martinho A, Herber N, Kroesen M, Chorus C. 2021. Ethical issues in focus by the autonomous

vehicles industry. *Transport reviews* 41(5):556–577
Mathur MB, Reichling DB. 2016. Navigating a social world with robot partners: A quantitative cartography of the uncanny valley. *Cognition* 146:22–32
Mayer MM, Bell R, Buchner A. 2021. Self-protective and self-sacrificing preferences of pedestrians and passengers in moral dilemmas involving autonomous vehicles. *PLoS one* 16(12):e0261673
McAllister A. 2016. Stranger than science fiction: The rise of ai interrogation in the dawn of autonomous robots and the need for an additional protocol to the un convention against torture. *Minn. L. Rev.* 101:2527
Miller T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267:1–38
Mittelstadt B. 2019. Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence* 1(11):501–507
Moor JH. 2006. The nature, importance, and difficulty of machine ethics. *IEEE intelligent systems* 21(4):18–21
Morley J, Floridi L, Kinsey L, Elhalal A. 2020. From what to how: an initial review of publicly available ai ethics tools, methods and research to translate principles into practices. *Science and engineering ethics* 26(4):2141–2168
Mullainathan S. 2019. Biased algorithms are easier to fix than biased people. *The New York Times*
Nielsen YA, Pfattheicher S, Keijsers M. 2022a. Prosocial behavior toward machines. *Current Opinion in Psychology* 43:260–265
Nielsen YA, Thielmann I, Zettler I, Pfattheicher S. 2022b. Sharing money with humans versus computers: On the role of honesty-humility and (non-) social preferences. *Social Psychological and Personality Science* 13(6):1058–1068
Nightingale SJ, Farid H. 2022. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences* 119(8):e2120481119
Noy IY, Shinar D, Horrey WJ. 2018. Automated driving: Safety blind spots. *Safety science* 102:68–78
Nussberger AM, Luo L, Celis LE, Crockett MJ. 2022. Public attitudes value interpretability but prioritize accuracy in artificial intelligence. *Nature Communications* 13(1):1–13
O'Leary DE. 2019. Google's duplex: Pretending to be human. *Intelligent Systems in Accounting, Finance and Management* 26(1):46–53
O'neil C. 2017. Weapons of math destruction: How big data increases inequality and threatens democracy. Crown
Ostermaier A, Uhl M. 2017. Spot on for liars! how public scrutiny influences ethical behavior. *PloS one* 12(7):e0181682
Pammer K, Gauld C, McKerral A, Reeves C. 2021. "they have to be better than human drivers!" motorcyclists' and cyclists' perceptions of autonomous vehicles. *Transportation research part F: traffic psychology and behaviour* 78:246–258
Pammer K, Predojevic H, McKerral A. 2023. Humans vs, machines; motorcyclists and car drivers differ in their opinion and trust of self-drive vehicles. *Transportation Research Part F: Traffic Psychology and Behaviour* 92:143–154
Pauketat JV, Anthis JR. 2022. Predicting the moral consideration of artificial intelligences. *Computers in Human Behavior* 136:107372
Pleiss G, Raghavan M, Wu F, Kleinberg J, Weinberger KQ. 2017. On fairness and calibration, In *Advances in Neural Information Processing Systems*, pp. 5684–5693
Rauhut H. 2013. Beliefs about lying and spreading of dishonesty: Undetected lies and their constructive and destructive social dynamics in dice experiments. *PloS one* 8(11):e77878
Rebitschek FG, Gigerenzer G, Wagner GG. 2021. People underestimate the errors made by algorithms for credit scoring and recidivism prediction but accept even fewer errors. *Scientific reports* 11(1):1–11
Rosenthal-von der Pütten AM, Krämer NC, Hoffmann L, Sobieraj S, Eimler SC. 2013. An exper-

imental study on emotional reactions towards a robot. *International Journal of Social Robotics* 5(1):17–34

Samuel S, Yahoodik S, Yamani Y, Valluru K, Fisher DL. 2020. Ethical decision making behind the wheel–a driving simulator study. *Transportation research interdisciplinary perspectives* 5:100147

Sandoval EB, Brandstetter J, Obaid M, Bartneck C. 2016. Reciprocity in human-robot interaction: a quantitative approach through the prisoner's dilemma and the ultimatum game. *International Journal of Social Robotics* 8(2):303–317

Santoni de Sio F. 2021. The european commission report on ethics of connected and automated vehicles and the future of ethics of transportation. *Ethics and Information Technology* 23(4):713–726

Savulescu J, Gyngell C, Kahane G. 2021. Collective reflective equilibrium in practice (crep) and controversial novel technologies. *Bioethics* 35(7):652–663

Saxena NA, Huang K, DeFilippis E, Radanovic G, Parkes DC, Liu Y. 2020. How do fairness definitions fare? testing public attitudes towards three algorithmic definitions of fairness in loan allocations. *Artificial Intelligence* 283:103238

Schwarting W, Pierson A, Alonso-Mora J, Karaman S, Rus D. 2019. Social behavior for autonomous vehicles. *Proceedings of the National Academy of Sciences* 116(50):24972–24978

Schwitzgebel E, Cushman F. 2015. Philosophers' biased judgments persist despite training, expertise and reflection. *Cognition* 141:127–137

Seering J, Flores JP, Savage S, Hammer J. 2018. The social roles of bots: evaluating impact of bots on discussions in online communities. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW):1–29

Seymour J, Tully P. 2016. Weaponizing data science for social engineering: Automated e2e spear phishing on twitter. *Black Hat USA* 37:1–39

Shank DB, DeSanti A, Maninger T. 2019. When are artificial intelligence versus human agents faulted for wrongdoing? moral attributions after individual and joint decisions. *Information, Communication & Society* 22(5):648–663

Shao C, Ciampaglia GL, Varol O, Yang KC, Flammini A, Menczer F. 2018. The spread of low-credibility content by social bots. *Nature communications* 9(1):1–9

Shariff A, Bonnefon JF, Rahwan I. 2017. Psychological roadblocks to the adoption of self-driving vehicles. *Nature Human Behaviour* 1(10):694–696

Shariff A, Bonnefon JF, Rahwan I. 2021. How safe is safe enough? psychological mechanisms underlying extreme safety demands for self-driving cars. *Transportation research part C: emerging technologies* 126:103069

Silver D, Huang A, Maddison CJ, Guez A, Sifre L, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature* 529(7587):484–489

Simpson M. 2022. Tech start-up denies whitening call center voices. *Canada Today*

Smith A. 2019. Public attitudes toward computer algorithms. pew research center. 2018. *URL: https://www.pewresearch.org/internet/2018/11/16/public-attitudes-toward-computer-algorithms*

Srinivasan R, Sarial-Abi G. 2021. When algorithms fail: Consumers' responses to brand harm crises caused by algorithm errors. *Journal of Marketing* 85(5):74–91

Srivastava M, Heidari H, Krause A. 2019. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning, In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 2459–2468

Starke C, Baleis J, Keller B, Marcinkowski F. 2022. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society* 9(2):20539517221115189

Stella M, Ferrara E, De Domenico M. 2018. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences* 115(49):12435–12440

Suzuki Y, Galli L, Ikeda A, Itakura S, Kitazaki M. 2015. Measuring empathy for human and robot hand pain using electroencephalography. *Scientific reports* 5(1):1–9

Takaguchi K, Kappes A, Yearsley JM, Sawai T, Wilkinson DJ, Savulescu J. 2022. Personal ethical settings for driverless cars and the utility paradox: An ethical analysis of public attitudes in uk and japan. *Plos one* 17(11):e0275812

Thomas PS, Castro da Silva B, Barto AG, Giguere S, Brun Y, Brunskill E. 2019. Preventing undesirable behavior of intelligent machines. *Science* 366(6468):999–1004

Tsvetkova M, García-Gavilanes R, Floridi L, Yasseri T. 2017. Even good bots fight: The case of wikipedia. *PloS one* 12(2):e0171774

Van Zant AB, Kray LJ. 2014. "i can't lie to your face": Minimal face-to-face interaction promotes honesty. *Journal of Experimental Social Psychology* 55:234–238

Villani V, Pini F, Leali F, Secchi C. 2018. Survey on human–robot collaboration in industrial settings: Safety, intuitive interfaces and applications. *Mechatronics* 55:248–266

von Schenk A, Klockmann V, Köbis N. 2022. Social preferences towards machines and humans. *Available at SSRN*

Wegner D, Gray K. 2016. The mind club: who thinks, what feels, and why it matters viking. *New York NY*

Wellman MP, Rajan U. 2017. Ethical issues for autonomous trading agents. *Minds and Machines* 27:609–624

Wotton ME, Bennett JM, Modesto O, Challinor KL, Prabhakharan P. 2022. Attention all 'drivers': You could be to blame, no matter your behaviour or the level of vehicle automation. *Transportation research part F: traffic psychology and behaviour* 87:219–235

Zhu A, Yang S, Chen Y, Xing C. 2022. A moral decision-making study of autonomous vehicles: Expertise predicts a preference for algorithms in dilemmas. *Personality and Individual Differences* 186:111356

Złotowski J, Sumioka H, Nishio S, Glas DF, Bartneck C, Ishiguro H. 2016. Appearance of a robot affects the impact of its behaviour on perceived trustworthiness and empathy. *Paladyn, Journal of Behavioral Robotics* 7(1)