Cognition 256 (2025) 106028

Contents lists available at ScienceDirect

Cognition

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

People expect artificial moral advisors to be more utilitarian and distrust utilitarian moral advisors

Simon Myers [a,b], Jim A.C. Everett [b,*]

[a] Behavioural Science Group, Warwick Business School, University of Warwick, Scarman Rd, Coventry CV4 7AL, UK
[b] School of Psychology, University of Kent, Canterbury, Kent, CT2 7NP, UK

ARTICLE INFO

Keywords
Artificial intelligence
Utilitarianism
Person perception
Algorithm aversion

ABSTRACT

As machines powered by artificial intelligence increase in their technological capacities, there is a growing interest in the theoretical and practical idea of artificial moral advisors (AMAs): systems powered by artificial intelligence that are explicitly designed to assist humans in making ethical decisions. Across four pre-registered studies (total $N = 2604$) we investigated how people perceive and trust artificial moral advisors compared to human advisors. Extending previous work on algorithmic aversion, we show that people have a significant aversion to AMAs (vs humans) giving moral advice, while also showing that this is particularly the case when advisors - human and AI alike - gave advice based on utilitarian principles. We find that participants expect AI to make utilitarian decisions, and that even when participants agreed with a decision made by an AMA, they still expected to disagree with an AMA more than a human in future. Our findings suggest challenges in the adoption of artificial moral advisors, and particularly those who draw on and endorse utilitarian principles - however normatively justifiable.

1. Introduction

Artificial intelligence (AI) - systems that use any kind of algorithm or statistical model to perform tasks that usually require human intelligence - are changing the world around us, modeling functions typical of the human mind such as visual and speech recognition, reasoning, and problem solving (Rahwan, Cebrian, Obradovich, et al., 2019). We rely on AI systems when we check the traffic on Google Maps, connect with a driver on Uber, or apply for a credit check, and as the technological sophistication of AI increases, so too do the tasks that we rely on AI for - dramatically increasing the stakes for humans. AI systems are approaching a level of complexity that progressively requires them to embody artificial *morality:* making decisions that would be described as moral or immoral if made by humans. We have already developed AI systems to be what theorists have called implicit, or indirect moral agents (Moor, 2009): systems that have limited ethical considerations built into their design as side-effects of their main purpose, to avoid them doing harm. An autonomous vehicle must, for example, have systems for minimising harm in emergency situations – e.g. if pedestrians step out in front of the car (e.g. Awad et al., 2018; Bonnefon, Shariff, & Rahwan, 2016). Even more concerningly, some systems are increasingly required to be explicit, or direct moral agents, processing ethically-relevant information about situations to fulfill their primary purpose of making decisions about what *should* be done.

1.1. Artificial moral advisors

Artificial moral advisors (AMAs) refer to AI systems that could be designed to assist humans in making ethical decisions, leveraging artificial intelligence to analyze moral dilemmas and provide recommendations based on established ethical theories, principles, or guidelines. The idea, at root, is simple: artificial intelligence provides a tool that can help enhance human activity in many different domains, so why not use AI to help people make better moral decisions? Humans have turned to experts for advice on difficult moral situations for millennia, and the rise of AI would “simply” enable these experts to be artificial, not human. The appeal of such AMAs is apparent: they could be accessible to a greater number of people at any time of day or night; they could use immense computational power to predict likely outcomes of different events in a way that humans could struggle to do; and by supposedly avoiding human cognitive biases and emotional ties, they could provide more consistent and rational moral advice. In this way, AMAs have been

[*] Corresponding author at: School of Psychology, University of Kent, Canterbury, Kent, CT2 7NP, UK.
E-mail address: j.a.c.everett@kent.ac.uk (J.A.C. Everett).

https://doi.org/10.1016/j.cognition.2024.106028
Received 10 July 2024; Received in revised form 22 November 2024; Accepted 1 December 2024
Available online 12 December 2024

suggested to serve a function akin to the “ideal observer” (Firth, 1952), by offering dispassionate and consistent judgments free from human biases (e.g. Giubilini & Savulescu, 2018; Sinnott-Armstrong & Skorburg, 2021). As AMAs develop, they could serve not only as tools for ethical guidance but also as means of fostering moral literacy.

While discussion of AMAs has largely been a hypothetical and future-focused one so far, recent advancements in large language models increasingly raise the possibility of artificial intelligence being (mis)used already broadly to serve as artificial moral advisors. ChatGTP already gives advice in moral dilemmas (e.g. Krügel, Ostermaier, & Uhl, 2023), and organizations are already working on prototypes for AI-powered systems designed specifically to model moral judgments, e.g. the Allen Institute’s “Ask Delphi” (Jiang et al., 2021) (see: delphi.allenai.org). It is now even possible to chat with an AI chatbot trained on the writings of famous ethicist, Peter Singer (see: https://www.petersinger.ai/), although it is intended to be used for “experimental and educational purposes only”. There are, it goes without saying, serious limitations with the current state of any publicly available AI-based systems that seek to give moral advice, and current LLMs like ChatGTP and AskDelphi come with warnings that outputs may contain errors and should not be (uncritically) used for advice for humans. ChatGPT is not designed to give moral advice and may resist answering certain morally weighted questions for safety or other reasons, and the creators of Delphi explicitly state that its purpose right now is not to be a moral authority or source of ethical advice (Jiang et al., 2021). Indeed, while some have argued for the theoretical possibility of more personalized AMAs (Giubilini & Savulescu, 2018), this appears increasingly at odds with the practicalities of how developers would – and even should - design and deploy AMAs in practice.

While fully fledged AMAs remain – at present – a theoretical possibility, their realisation becomes closer, and this forces discussion not of the technological challenge of creating AMAs, but the psychological and philosophical challenges associated with the endeavour in the first place. For example, how would one incorporate human values, do we even have a set of static moral values one could incorporate in a coherent form, and how do we deal with the extent to which people disagree with each other’s moral judgments? Indeed, these theoretical issues may not be so easily resolved (e.g. Liu, Moore, Webb, & Vallor, 2022). The questions of how we would create AMAs, whether we should, and what the long term consequences of doing this would be, remain very much open. But even if we leave aside the very important question of whether people *should* trust AMAs, there remains a descriptive question of whether people even *would* trust AMAs.

We have reasons to assume people would be reluctant to trust such explicit artificial moral agents. Much research has documented the phenomenon of algorithmic aversion: the tendency for individuals to distrust AI relative to humans even when there is identical - or even superior - performance (Dawes, 1979; Dietvorst, Simmons, & Massey, 2015; Meehl, 1954, 1957). For example, people lose confidence in statistical models more quickly than in humans after seeing both make the same mistake (Dietvorst et al., 2015; Prahl & Van Swol, 2017), and place greater weight on the same advice given by human experts than statistical models (Önkal, Goodwin, Thomson, Gönül, & Pollock, 2009; Promberger & Baron, 2006). Importantly, such algorithm aversion also drives distrust for AI making moral decisions, seemingly driven by a perception that AI lacks internal experience (Bigman & Gray, 2018). But there is a further problem: when it comes to moral problems, it is not only that the stakes are higher, but more critically that there is not always an established consensus on what the right moral decision might be.

A fundamental challenge that developers of AMAs will face is which *kind* of ethical framework to benchmark against, especially in moral dilemmas where different ethical frameworks endorse different, mutually exclusive actions. For example, is it morally acceptable to break normal prohibitions against murder in order to prevent harm to a greater number? Different ethical theories will have different responses, it is far from clear which ethical theory should be the benchmark. Consequentialist theories such as utilitarianism focus on the ‘greatest good for the greatest number’, positing that only consequences matter when making moral decisions (Bentham, 1983; Mill, 1863; Singer, 1993). In contrast, non-utilitarian deontological theories claim that we also have to consider rights, duties, and obligations, for example, even if murder might bring about good consequences, it may still be judged as wrong (e.g., Fried, 1978; Kant, 2002; Ross, 1930). Just as humans face such dilemmas, so too will AI agents - and they will have to respond ‘appropriately’, in a way that aligns with our values. This is not just a theoretical concern about ensuring that artificial moral advisors align with the “correct” normative standard (though indeed there must be some kind of benchmarking to moral standards), but a *psychological* question about how these advisors might be differentially trusted based on the specific decisions they make. Morality is not just about perceptions of the rightness or wrongness of acts, but more often fundamentally person-based: actions give us insight into the perceived moral character, and it is perceptions of character that can in turn help shape what we see as justifiable (e.g. Everett, Pizarro, & Crockett, 2016; Uhlmann, Pizarro, & Diermeier, 2015; Uhlmann, Zhu, & Tannenbaum, 2013).

1.2. *Inference of trust from moral decisions*

A growing body of research in moral psychology has shown that the way people respond in sacrificial dilemmas has a host of consequences for how that person is perceived. For example, Everett et al. (2016) looked at perceptions of people who made either “characteristically utilitarian” vs. non-utilitarian “characteristically deontological” judgments in the footbridge dilemma (Foot, 1967; Thomson, 1976, 1984), which asks participants to judge whether it would be acceptable to kill one man by pushing him off a footbridge in order to save five other people on the tracks. Participants were asked to judge two other “agents” who received this dilemma (and other agents, across a series of studies in the paper), where one agent gave a characteristically utilitarian response (“it is better to save five lives than one”), and the other gave a characteristically deontological, non-utilitarian decision to reject the sacrifice (“killing people is just wrong, even if it has good consequences”). They found that participants perceived those who gave characteristically deontological responses to a sacrificial moral dilemma as more trustworthy than those who gave utilitarian responses, found both in self-reports and in behavior in a trust game (Everett et al., 2016). Such results have been shown by numerous independent research groups (Brown & Sacco, 2019; Rom, Weiss, & Conway, 2017a, 2017b; Sacco, Brown, Lustgraaf, & Hugenberg, 2017), pre-registered replication projects (Everett, Faber, Savulescu, & Crockett, 2018), and a large cross-cultural Registered Report conducted in 22 countries (Everett et al., 2021). Across a variety of dilemmas, and even when controlling for participants’ own judgments, it appears that endorsing utilitarian decisions in sacrificial dilemmas can decrease trust (see Crockett, Everett, Gill, & Siegel, 2021 for a review).

Such findings have been explained with reference to partner choice models relating to the importance of choosing trustworthy social partners, with Everett et al. (2016; 2018) arguing that following utilitarian principles about the maximization of benefits leads to behaviors that are often less predictable than following simpler deontological rules. Indeed, deontological judgments relating to duties, obligations, and aversion to harm typically indicate that the agent has more socially valuable beliefs about others. There is evidence for both of these. For the possibility of non-utilitarian decisions signaling greater commitment to cooperation, research shows that if a utilitarian agent reported their judgments as being very difficult to make – thereby indicating some level of commitment to cooperation - distrust of them was reduced (Everett et al., 2016), and there is also evidence that people strategically endorse non-utilitarian resolutions to moral dilemmas to make themselves appear warmer and more moral (Rom et al., 2017a, 2017b). For the possibility of non-utilitarian decisions being predictable, there is mixed evidence. On the one hand, in their study Everett et al. (2016) found evidence that people trusted a “contractualist” agent who focused more on respect for others’ wishes, even if that meant breaking a moral norm. This agent was trusted over a “Kantian” agent who always followed rules even when this led to harmful consequences, and this was interpreted as suggesting that a flexible commitment to social norms while still respecting others was more important than predictability per se. On the other hand, however, a growing body of more recent work has suggested that predictability might be a more important driver. Walker et al. (2021) show that people demonstrate a moral preference for more predictable immoral actors over unpredictable immoral actors, and in the context of moral dilemmas specifically, Turpin et al. (2021) show not only that utilitarian agents were perceived as less predictable and less moral than deontological agents, but that when utilitarian decision-makers are made to seem more predictable this difference disappeared.

We know that humans who endorse utilitarian resolutions to sacrificial dilemmas are trusted less - perhaps due to them seeming less committed social partners and being less predictable. But how might this apply to how people perceive *artificial* moral advisors? Compared to the wealth of work looking at perceptions of humans who endorse utilitarian or non-utilitarian judgments in moral dilemmas, there is little looking at AI. Young and Monroe (2019) looked at perceptions of a self-driving car in a “switch-style” moral dilemma and found that while utilitarian pro-sacrificial decision-makers were trusted more than non-utilitarian decision-makers, and that humans were trusted more than AI, there was no interaction between the two. In other work looking at whether people accept advice provided by artificial moral advisors (but not looking at perceptions of *advisors* per se), Krügel et al. (2023) report results from an experiment in which participants were presented with output from ChatGTP in response to being given the classic trolley dilemma, with identical outputs labeled either (correctly) as coming from ChatGTP or (inaccurately) as coming from a human advisor. Their results show that people found the sacrifice in the footbridge dilemma differentially acceptable depending on whether the advisor endorsed the sacrificer, and this was the same for output labeled as coming from both ChatGTP or a human advisor. However, it remains unclear how utilitarian moral judgments would shape perceptions of the advisor themselves (rather than just agreement with their answers), whether people would differentially trust the advisor in other less famous dilemmas, and whether this holds when ensuring equal length and ethical appeals of the utilitarian and non-utilitarian justifications.

1.3. Present research

In this paper, across four pre-registered studies, we therefore investigated the similarities and differences between how people think about artificial and human moral advisors who gave utilitarian and non-utilitarian advice in different moral dilemmas. In Study 1, we explore perceptions of human and AI moral advisors who give utilitarian or non-utilitarian advice in classic sacrificial dilemmas where the sacrifice is used as either a direct means for bringing about a greater good or as a side-effect of doing so. In Study 2, we build on this to explore differences in how artificial (and human) advisors are perceived when giving utilitarian advice about the morality of harming others for the greater good (instrumental harm) compared to advice about the importance of impartially maximizing welfare at the expense of more local special obligations (impartial beneficence). In Study 3, we explore the expectations that people have about artificial moral advisors, focusing on the role of predictability vs. appropriate sensitivity to moral contexts in driving perceptions of trustworthiness.

2. Study 1

In Study 1, we explored perceptions of human and AI moral advisors who give utilitarian or non-utilitarian advice in classic sacrificial dilemmas where the sacrifice is used as either a direct means for bringing about a greater good or as a side-effect of doing so. We chose to look at both means-style and side-effects because we know people are more likely to endorse the utilitarian option in side-effect cases (like the “switch” case) than they are in the means-style cases (like the “footbridge” case) (e.g. Greene, Sommerville, Nystrom, Darley, & Cohen, 2001), and we know that differential endorsements across these cases can be particularly diagnostic of differences in the social and moral values held by the advisor (Everett et al., 2016). Looking at perceptions of human and artificial advisors who made utilitarian decisions in both means and side-effect cases, then, allowed us to both enhance generalizability and better be able to control for the role of participants’ own agreement, therefore better differentiating between, for example, trusting non-utilitarians in general from trusting those we simply happen to agree with in means-style cases. In addition, it is not obvious a priori how these preferences would generalize to artificial moral advisors: while people may prefer non-utilitarian human decision makers, they may feel better about AI making utilitarian decisions because that is what we expect of it (Malle et al., 2015). Alternatively, we may distrust AI more for endorsing instrumental harm because it may shield a human-in-the-loop from taking responsibility for that harm. Therefore, knowing whether AI advisors are more or less penalised for giving particular kinds of unfavourable advice is especially important.

2.1. Method

2.1.1. Open science

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at: https://osf.io/m3v48/. This study was pre-registered at https://osf.io/29kqd.

2.1.2. Participants

We recruited 1100 participants living in the United Kingdom through Prolific.ac. After excluding 52 participants for failing a pre-registered attention check at the start of the study, and 40 participants for failing the pre-registered manipulation check, we were left with a final sample size of 1008 (478 women; $M_{age} = 41.7$, $SD = 13.7$). Our sample size was determined through a simulation-based power analysis, testing for a focal effect of differences in ratings of trust predicted by advisor type (human vs AI); advice type (utilitarian vs non-utilitarian); and dilemma type (means vs side-effect). This simulation estimated that 1050 participants would be required to have 80 % power for finding small to medium effects sizes (standardized beta 0.3).

2.1.3. Design

We employed a mixed design in which we manipulated between-subjects advisor type (Human vs AI); manipulated within-subjects the advisor’s moral advice (utilitarian vs non-utilitarian); and manipulated within-subjects the dilemma type (“Means” vs “Side Effect”) such that participants saw one of two possible “Means” dilemmas and one of two possible “Side-Effect” dilemmas. This meant that participants saw either AI or human advisors, but saw (different) advisors give both utilitarian and non-utilitarian advice across the two dilemma types.

First, participants were assigned to see either an AI or Human moral advisor (between-subjects). All participants were given a brief description about the way we often face moral dilemmas in our life, and that oftentimes we turn to external advisors for advice. Then, participants were told that this role can often be performed by human experts who draw on their extensive training to provide recommendations in such difficult moral cases; or that this role can be provided by artificial moral advisors who draw on the latest advancements in artificial intelligence and machine learning to provide recommendations in such difficult moral cases.

Second, participants were presented with a first dilemma that reflected the tension between utilitarian and non-utilitarian principles in the context of instrumental harm. These dilemmas included both "means-style" dilemma where the sacrificial action is a direct means to saving the greater number of people (i.e. that the sacrifice of the person is directly intended as a way of achieving this greater good), and "switch-style" dilemmas where the sacrificial action is a foreseen side-effect of saving the greater number of people (i.e. that the sacrifice of the person is an unfortunate side-effect of the sacrificial action) (Greene et al., 2001). To enhance generalizability, we used two instances of both types of dilemma, though participants only saw one example of each type (all dilemmas adapted from Moore, Clark, & Kane, 2008). For the "means-style" dilemmas, we used adapted versions of the "Crying Baby" dilemma in which an agent has to decide whether to smother a baby to death in order to ensure that enemy soldiers do not find people hiding in a cellar, and the "Bike" dilemma in which the agent has to decide whether to crash into a single motorcyclist in order to prevent a pile-up that will kill many more. For the "switch-style" dilemmas, we used the "Hospital" case in which an agent has to decide whether to redirect a ventilation system pumping poisonous gas into a room with multiple people into a different room that has one person, and the "Submarine" case in which an agent has to decide whether to redirect oxygen between injured people on a submarine after an explosion and move oxygen away from the level where there is a single unconscious person to a different level where there are more people. After reading these descriptions, and before seeing the advisor's advice, participants gave their own moral judgment about what they think should be done in that situation.

Third, participants moved to see the advisor's recommendation on this dilemma, where participants read that after considering this problem, the advisor drew on either their human "knowledge about the ethics and similar cases" or their AI-driven "advancements in machine learning about moral cases" and made a judgment. Half the participants saw the advisor give a characteristically utilitarian pro-sacrificial judgment, endorsing the sacrificial action with the justification that "An important principle in ethics is to think about the greater good, and in this specific case killing the one person would bring about better consequences overall". The other half of the participants saw the advisor give a non-utilitarian judgment rejecting the sacrificial action with the justification that "An important principle in ethics is that killing people is just wrong, and this duty to not kill should apply even if killing has good consequences in a specific case.". After reading the advisor's judgment, participants then rated the trustworthiness of the advisor, their willingness to follow the advisor's advice in future, how much they would blame someone for following the advisor's advice, and how much they expected to agree/disagree with the advisor on different issues in future. After completing these, participants completed a manipulation check requiring them to report the judgment that the advisor made.

Next, participants were told that they would now see advice from another advisor (with a different name). This second dilemma was one of two possible dilemmas of the opposite type to the first dilemma seen (i.e. if participants saw first the means-style "Baby" or "Bike" dilemma, they would then see a switch-style dilemma, randomly either "Hospital" or "Submarine"). For this second dilemma, participants always saw the advisor give advice that was the opposite to what they previously saw (i.e. if participants saw the advisor give a consequentialist recommendation in the means-style dilemma, they would then see the advisor give a non-utilitarian judgment in the switch-style dilemma). Participants completed the same dependent measures as for the first dilemma. Therefore, across the study participants saw either two AI advisors or two human advisors, of which each one gave a utilitarian recommendation for one dilemma type and the other advisor gave a non-utilitarian recommendation for the other dilemma type, in counterbalanced order.

For this study we pre-registered that we expect that human advisors will be trusted more than AI. We also pre-registered that we will explore whether this differs based on the kinds of judgment and the type of dilemma based on previous work on trust in humans showing that those who make deontological moral judgments tend to be favored particularly (or only) in cases where the harm to one is done as a means to an end ("footbridge style") compared to cases where the harm is an intended side-effect ("switch style") (Everett et al., 2016).

Lastly, we pre-registered that we would test whether any observed effects persist when accounting for how much participants agree with the advice, and whether future anticipated agreement might account for variance in differential trust between Humans and AI (e.g. despite AI making a decision that the participant agrees with they still don't trust it; perhaps because they predict the AI will still make more disagreeable judgments in the future compared to humans).

2.1.4. *Measures*

Participant's own moral judgment was measured prior to seeing the advisor's recommendation, with the precise wording varying for each specific dilemma. The non-utilitarian action was at the lower end of the scale and the utilitarian at the higher end, for example "Should Helen activate the ventilation system, redirecting the gas into another room that will kill the single soldier but save the lives of the six patients?" (*1 = definitely not activate the ventilation system [non-utilitarian]; 7 = definitely activate the ventilation system [utilitarian]*).[1]

Trust, our key dependent measure, was measured in a single item asking "How trustworthy do you think [advisor] is?" (*1 = not at all; 7 = very much*).

Willingness to adopt advice was measured in a single item asking "Based on their advice, how willing would you be to trust [advisor] on other issues?" (*1 = not at all; 7 = very much*).

Blame for an individual if they were to follow the advisor's advice was measured with a single item of "*How much would you blame someone if they followed this advice?*"

Expected future agreement was measured with a single item after the other dependent measures: "In the scenario you just read, [advisor] made a recommendation that [matched/was different to] what you thought should be done. Imagine that you turned to [advisor] in future for advice for a different kind of moral problem. Do you expect that its advice would again match with your own view on a different moral problem? (*1 = not at all; 7 = very much*).

2.1.4.1. *Attention check.*

At the start of the study, we had participants complete an attention check drawn from Everett et al. (2021) in which participants were told that to demonstrate they were paying attention they would need to respond with a specific response to a question on the following page. Participants who did not give this response were excluded from analysis, in line with our pre-registration.

2.1.4.2. *Manipulation check.*

After reading the first dilemma, seeing the advisor's judgment, and answering the dependent measures, participants were asked to report back the judgment the advisor made. Participants who did not correctly report back the judgment of the condition they were assigned to were then excluded from data analysis, again in line with our pre-registration.

2.2. *Results*

First, to assess people's judgments across the dilemmas, four one-sample *t*-tests were calculated to test whether people had significant preferences towards which action they thought was the most morally right choice. For each dilemma, participant's own judgments were significantly different from zero (indifference), Bike ($M = -0.45$, $SD = 1.84$), $t(503) = -5.55$, $p < .001$, $d = -0.25$; Baby ($M = -0.21$, $SD = 1.96$), $t(503) = -2.39$, $p = .017$, $d = 0.12$; Hospital ($M = 0.61$, $SD = 1.76$, $t(503) = 7.81$, $p < .001$, $d = 0.35$; Submarine ($M = 1.27$, $SD = 1.52$), $t(503) = 18.80$, $p < .001$, $d = 0.84$. Consistent with previous

[1] These are re-coded such that 0 is the center of the scale −3 was the non-utilitarian end and 3 was the utilitarian since the scale is bipolar.

research, for the switch-style dilemmas (Hospital and Submarine) participants were more likely to endorse the pro-sacrificial utilitarian action (70.3 % utilitarian; 18.4 % non-utilitarian; 11.3 % unsure), while for means-style dilemmas (Bike and Baby) judgments were more mixed, but slightly more likely to be non-utilitarian (42.3 % utilitarian; 45.7 % non-utilitarian; 12.0 unsure).

2.2.1. *Human advisors and non-utilitarian advisors are both trusted more*

To assess how the type of advisor (human vs AI), the advice given (utilitarian vs non-utilitarian), and type of dilemma (means vs side-effect) predicted participant's judgments, three pre-registered models were calculated. The first predicted trust, the second predicted the participant's willingness to adopt the advice, and the third predicted blame judgments. These models were linear mixed-models that specified random intercepts by participant. These models were the maximal converging models (see Barr, Levy, Scheepers, & Tily, 2013). See Table 1 for full results. We found that participants were significantly less likely to trust AI advisors than human advisors, $F(1, 1004) = 41.75$, $p < .001$, $\beta = 0.12$; less likely to adopt their advice, $F(1, 1004) = 64.11$, $p < .001$, $\beta = 0.25$; and more likely to blame those who followed the AI advisor's advice, $F(1, 1004) = 14.61$, $p < .001$, $\beta = -0.22$. Moreover, consistent with previous research we found that participants were less likely to trust the utilitarian advisors in the means-style dilemmas, even though they were more likely to trust the utilitarian advisors for the switch-style dilemmas: a pattern found for trust $F(1, 1004) = 41.60$, $p < .001$, $\beta = -0.58$ (see Fig. 1); with the same pattern found for willingness to listen to the advisor in future, $F(1, 1004) = 56.24$, $p < .001$, $\beta = -0.64$; and blame for those followed the advice, $F(1, 1004) = 35.30$, $p < .001$, $\beta = 0.73$.

2.2.2. *The role of agreement*

Our key interest - as pre-registered in our analysis plan - was to look at overall effects in how people would respond to advisors who gave utilitarian or non-utilitarian advice, across our sample. While developers may seek to align judgments with public preferences overall, it is necessary to know how people will respond to AMAs that give different types advice for two key reasons: first, the potentially widely and openly available AMA systems of the future are unlikely to know a specific user's own preferences in the moral dilemma in advance; and second, even if they could, a key theoretical appeal of AMAs is that they could provide impartial, disinterested *advice* that draws on normative principles, not merely serve as a parrot that repeats back what participants themselves would already think. That said, psychologically we still wanted to explore the possibility that at least some of our results are driven by the extent to which participants simply agreed with advice that was given for each dilemma. That is, we know that participants are more likely to endorse the pro-sacrificial utilitarian action in the switch-style dilemmas but reject it in the means-style dilemmas, so our finding that participants distrusted utilitarian advisors in the means-style dilemma but not the switch-style could potentially suggest that this is just about agreeing with the advice.

To investigate the role of agreement, as pre-registered, we coded the level of agreement by how congruent participant's judgments were with the advice given (3 being most agreed, −3 being least agree and 0 being neither agree nor-disagree). For example, those who were given non-utilitarian advice and indicated that the non-utilitarian choice was the most morally correct choice scored 3 on agreement. Beginning with responses to the side-effect (switch-style) dilemmas, we find that people were significantly less likely to trust AI advisors, $F(1,1000) = 52.81$, $p < .001$, $\beta = 0.32$. As predicted, advisors who gave recommendations that the participant agreed with were seen as significantly more trustworthy, $F(1,1000) = 207.69$, $p < .001$, $\beta = 0.48$. There was, however, also a main effect of advice $F(1, 1000) =22.71$, $p < .001$, $\beta = 0.09$, and an advice-agreement interaction effect $F(1, 1000) = 7.24$, $p = .007$, $\beta = -0.17$, such that non-utilitarians were still trusted significantly more even when controlling for agreement (see Fig. 2). All other predictors were non-significant (see Table 2 for full results).

Following this, we performed the same analysis for responses to the mean-style dilemmas, finding once again that people were significantly less likely to trust AI advisors, $F(1, 1000) = 13.35$, $p < .001$, $\beta = 0.14$. Also, again as predicted, advisors who gave recommendations that the participant agreed with were seen as significantly more trustworthy, $F(1, 1000) = 321.09$, $p < .001$, $\beta = 0.58$. There was also a main effect of advice $F(1, 1000) = 38.30$, $p < .001$, $\beta = 0.09$, and an advice-agreement interaction effect $F(1, 1000) = 11.37$, $p = .001$, $\beta = -0.20$, such that non-utilitarians were still trusted significantly more when controlling for agreement. All other predictors were non-significant (see Table 2). Overall, then, we find that while participants' agreement with the advice was a significant predictor of trust, AI (and human) advisors who gave utilitarian advice were still trusted less than those who gave non-utilitarian advice, and this was the case for both the switch-style and means-style cases.

2.2.3. *Beliefs about future advice*

Lastly, we wished to see the extent to which trust judgments were driven by beliefs about future advice (or beliefs that any good advice, this time round, was merely a fluke). That is, given a potential perception of AI as "noisy" and lacking deep understanding, do people feel that AI is more likely to give advice they would disagree with in the future, even when that AI gave advice that was agreed with in that specific example? In other words, might people be more likely to think that good advice from an AI is more likely to be a fluke than the same good advice from a human? Mixed-models were calculated (random intercepts by participants and random slopes for agreement) showing that the more people agreed with the human advisor the more likely they thought that the human would give consistently agreeable advice in the future, while the more they agreed with the AI advice the more likely they thought that this would *not* be the case in the future $F(1,681) = 4.70$, $p = .031$, $\beta = 0.10$. To assess how these beliefs are associated with trust, pre-registered linear mixed models were calculated predicting trust (random intercepts by participant and random slopes for expected consistency). Crucially, expected consistency predicted trust judgments, $F(1,1999) = 16.34$, $p < .001$, $\beta = -0.19$; and there was a significant interaction between consistency and whether the participant agreed F

Table 1
The effect of advisor type, advice type, and dilemma type on trust, willingness, and blame in Study 1.

		Trust			Willingness			Blame		
Predictors	*df*	*F*	*β*	*p*	*F*	*β*	*p*	*F*	*β*	*p*
Advice	1004	0.67	0.24	.413	0.02	0.26	.885	0.38	−0.37	.885
Advisor	1004	41.75	0.12	< .001***	64.11	0.25	< .001***	14.61	−0.22	< .001***
Dilemma	1004	2.94	0.25	.087	1.74	0.33	.187	19.09	−0.50	.187
Advice:Advisor	1004	3.57	0.17	.059	1.43	0.13	.231	0.18	0.10	.231
Advice:Dilemma	1004	41.60	−0.58	< .001***	56.24	−0.64	< .001***	35.30	0.73	< .001***
Advisor:Dilemma	1004	6.71	0.23	.010**	0.95	0.12	.329	0.27	0.09	.329
Advice:Advisor:Dilemma	1004	0.05	−0.04	.821	0.14	−0.07	.707	1.66	−0.26	.707

Random Intercepts by Participant.
* $p < .05$, ** $p < .01$, *** $p < .001$.

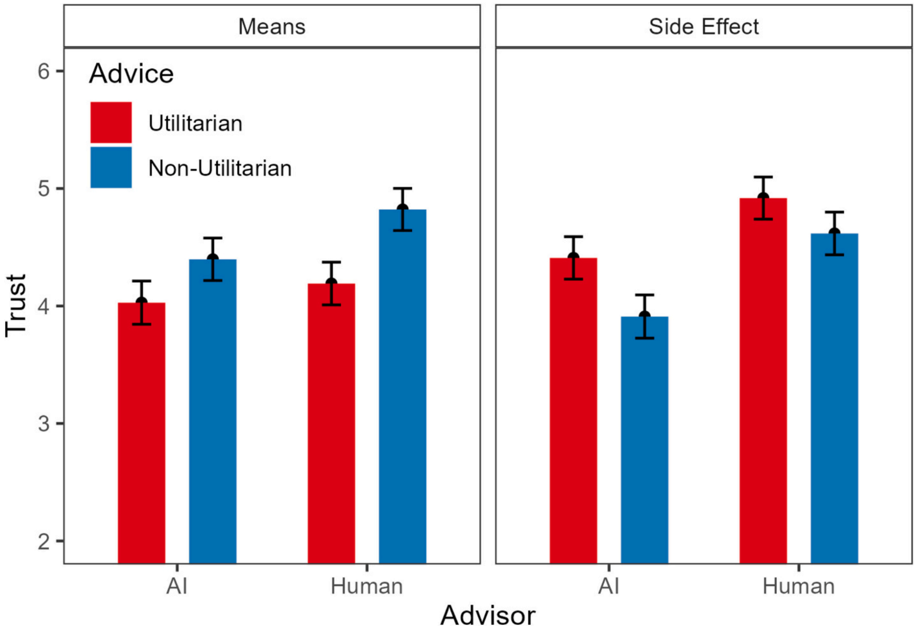

Fig. 1. The effects of advice, advisor, and dilemma type on perceived trustworthiness of moral advisors in Study 1.

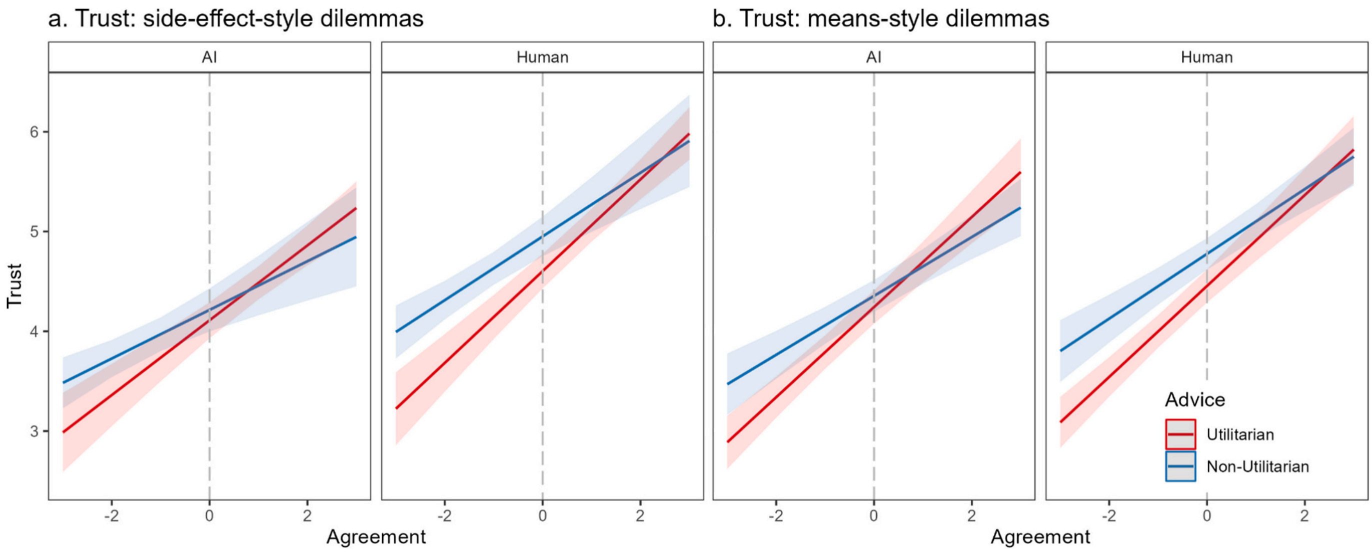

Fig. 2. The effects of advice, advisor, and dilemma type on perceived trustworthiness of moral advisors in Study 1, controlling for agreement.

Table 2
The effect of advisor type and advice type on trust in Study 1, controlling for agreement.

	Means				Side Effect			
Predictors	*df*	*F*	*β*	*p*	*df*	*F*	*β*	*p*
Advisor	1	13.35	0.14	<.001***	1	52.81	0.32	<.001***
Advice	1	38.30	0.09	<.001***	1	22.71	0.09	<.001***
Agree	1	321.09	0.58	<.001***	1	207.69	0.48	<.001***
Advice:Advisor	1	1.92	0.14	.167	1	0.25	0.16	.165
Advice:Agree	1	11.37	−0.20	.001**	1	7.24	−0.17	.007**
Advisor:Agree	1	0.15	0.01	.154	1	2.56	0.11	.110
Advice:Advisor:Agree	1	0.09	0.03	.088	1	0.01	−0.01	.927
Residuals	1000	–		–	1000	–	–	–

Random Intercepts by Participant and Slopes for Advice.
* p < .05, ** p < .01, *** p < .001.

(1,1995) = 111.45, $p < .001$, $\beta = 0.64$, such that trust was higher when the advice was agreeable and people expected it would also be consistently agreeable in the future. There was also a significant three-way interaction involving advisor type (AI vs Humans), $F(2, 2006) = 3.23$, $p = .040$, $\beta = -0.12$, such that when one disagrees with advice but doesn't expect the agent to be consistent, humans are still more likely to be trusted overall compared to AI.

2.3. Discussion

In Study 1, we explored perceptions of human and artificial moral advisors who gave characteristically utilitarian or non-utilitarian advice in both "means-style" dilemmas where the sacrificial action was a direct means to saving the greater number of people (i.e. that the sacrifice of the person was directly intended as a way of achieving this greater good), and "switch-style" dilemmas where the sacrificial action was a foreseen side-effect of saving the greater number of people (i.e. that the sacrifice of the person is an unfortunate side-effect of the sacrificial action). Our results reliably demonstrated algorithmic aversion towards AMAs: even when given the same advice, our participants trusted artificial moral advisors less than human advisors, were less willing to think they would rely on them in future, and would blame others more for following that advice. Moreover, we found evidence of distrust of AMAs (and human) advisors who give utilitarian advice, with participants distrusting utilitarian advisors in the means-style dilemmas that have been argued to be particularly important for signaling socially valued views about others (Everett et al., 2016; Everett et al., 2018). These results held in both means-style and switch-style cases even when controlling for participants' own agreement, demonstrating that while agreement naturally predicts trust, there remains a persistent effect of advice type. Finally, we find evidence that our participants appear to think that good advice from an AI is more likely to be a fluke than the same good advice from a human: that even if someone agrees with an artificial moral advisor in a specific instance, they still expect to be less likely to agree with them in future than a human advisor. In Study 2, we sought to replicate and extend these results by looking at how people perceive advisors who made utilitarian decisions in dilemmas not only about instrumental harm, but impartially helping others to achieve the greater good.

3. Study 2

In Study 2, we turned to look at whether AI and human advisor's endorsement of utilitarian principles may lead to differential trust based on the kind of utilitarian principle appealed to. While much research in moral psychology has tended to treat sacrificial dilemmas as the core and even defining feature of utilitarianism, the *two-dimensional (2D) model of utilitarian psychology* (Everett & Kahane, 2020; Kahane et al., 2018; Kahane & Everett, 2023) is based on the recognition that there is in fact at least two primary ways in which utilitarianism, as a philosophical theory, departs from our common-sense moral intuitions. First, and indeed in line with the focus on sacrificial dilemmas, utilitarianism permits harming innocent individuals when this maximises aggregate utility - what we can call *instrumental harm*. Second, however, utilitarianism as an ethical theory requires that we treat the interests of other individuals as equally morally important, without giving priority to oneself or those to whom one is especially close - what has been termed *impartial beneficence*.

A growing body of research has shown that these two dimensions of utilitarianism, while philosophically entailed by classical utilitarianism, are psychologically dissociable. While expert philosophers who tend to endorse (or reject) one also tend to also endorse (or reject) the other, ordinary people display only a weak positive correlation between the two (e.g. Kahane et al., 2018). As well as being empirically distinguishable through factor analyses, endorsements of utilitarian instrumental harm and impartial beneficence have distinct psychological correlates. For example, while much research has painted a rather unflattering picture of utilitarian judgments in sacrificial dilemmas by finding associations with psychopathy and reduced empathic concern, people who endorse the utilitarian impartial maximization of welfare (e.g. "It is morally wrong to keep money that one doesn't really need if one can donate it to causes that provide effective help to those who will benefit a great deal") are actually less likely to agree with statements tapping subclinical psychopathy and more likely to agree with statements tapping empathic concern (Kahane et al., 2018; Kahane, Everett, Earp, Farias, & Savulescu, 2015).

Most importantly for the present paper, there is increasing evidence that those who endorse utilitarian instrumental harm and those who endorse impartial beneficence are not always perceived in the same way. We have already discussed how previous research has shown that people who endorse utilitarian instrumental harm in sacrificial dilemmas are seen as less moral, less trustworthy, are chosen less frequently as social partners, and trusted less in economic exchanges than those who reject it (Bostyn & Roets, 2017a, 2017b; Everett et al., 2016; Everett et al., 2018; Rom et al., 2017a, 2017b; Sacco et al., 2017; Uhlmann et al., 2013). In a similar vein, there is evidence that endorsing utilitarian impartial beneficence may also incur social costs in at least some contexts. For example, those who help a stranger instead of family members are judged as less morally good and trustworthy than those who did the opposite (McManus, Kleiman-Weiner, & Young, 2020), and that this pattern of results is seen even when it is clear that helping strangers would maximize the greater good (Hughes, 2017). Similarly, Law, Campbell, and Gaesser (2022) show that socially distant altruists (e.g. endorsing donating money to save the life of a distant stranger in another country) tend to be seen as having a worse moral character than those who are socially close altruists (e.g. endorsing spending their money on a dream vacation for their terminally ill child).

Importantly, however, there is evidence that this "cost of being consequentialist" may depend on the type of social role occupied. Everett et al. (2018) find that people were seen as a worse friend but a *better* political leader when they endorsed impartial beneficence in "greater good" dilemmas that contrasted special obligations with impartial maximization of welfare. Following from this, Everett et al. (2021) conducted a Registered Report experiment with 23,000 participants in 22 countries over six continents to explore how endorsement of utilitarian instrumental harm and impartial beneficence shaped perceptions of political leaders in the context of the COVID-19 pandemic. Their results showed across both self-reported and behavioral measures, endorsement of instrumental harm (e.g. the permissibility of mandatory tracing devices to reduce the spread of the virus) decreased trust, while endorsement of impartial beneficence (e.g. whether medicine should be sent wherever in the world it would do the most good or reserved first for a country's own citizens) increased trust - a finding recently replicated in the context of vaccine nationalism (Colombatto, Everett, Senn, Maréchal, & Crockett, 2023).

In Study 2, then, we aimed to extend our investigation by exploring whether human and AI advisors who gave utilitarian judgments in both instrumental harm and impartial beneficence were differentially trusted. While some research has found that impartial beneficence has negative consequences for social impressions, particularly for "ordinary" people (e.g. Everett et al., 2018; Hughes, 2017; Law et al., 2022), there is also evidence that for people in higher-level positions like political leaders, for whom part of their social role is to treat the interests of citizens equally and do what is best for the country overall - perhaps akin to the role of AMAs as dispassionate and disinterested advisors - utilitarian impartial beneficence may actually increase trust (Colombatto et al., 2023; Everett et al., 2018; Everett et al., 2021). In addition to our hypotheses regarding Study 1, we were also interested in how people's differential expectations for AI vs humans might affect their judgments regarding that agent, given the different recommendations the agent makes. For example, it is possible that people expect AI to be more utilitarian (e.g. see Malle et al., 2015), while they expect humans to be more likely to give the non-utilitarian recommendation. If this is the case, then trust might be affected differently for utilitarian recommendations from AI compared to the same recommendations from a human advisor. To explore this, we included a new question measuring the extent to which participants are surprised by the advice.

3.1. Method

3.1.1. Open science

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at: https://osf.io/m3v48.

This study was pre-registered at the Open Science Framework: https://osf.io/ab45v.

3.1.2. Participants

We recruited 1100 participants living in the United Kingdom through Prolific.ac. After excluding 32 participants for failing a pre-registered attention check at the start of the study and 38 participants for failing the manipulation check, we were left with a final sample size of 1030 (562 women; M_{age} = 43.1, *SD* = 14.1). Our sample size was determined through the same simulation-based power analysis as Study 1 as the focal analyses do not differ between these studies.

3.1.3. Design

The design and procedure for Study 2 was nearly identical to that used in Study 1, except instead of using two different types of sacrificial dilemmas (means vs side-effect), we contrasted sacrificial dilemmas tapping the endorsement of utilitarian instrumental harm with "greater good" dilemmas tapping the endorsement of utilitarian impartial beneficence. This meant that, as before, we had a mixed design in which we manipulated between-subjects advisor type (human vs AI); manipulated within-subjects the advisor's moral judgment (utilitarian vs non-utilitarian); and manipulated within-subjects the dilemma type (instrumental harm vs impartial beneficence), such that participants saw one of two possible instrumental harm means-style dilemmas and one of two possible impartial beneficence dilemmas. Again, this meant that participants saw either two AI or two human advisors, where one advisor gave a utilitarian recommendation for one dilemma type and the other gave a non-utilitarian recommendation for the other dilemma type (counterbalanced).

The dilemmas used for instrumental harm were the same sacrificial means-style dilemma used in Study 1 ("Crying Baby" and "Bike"), and for impartial beneficence we used two dilemmas adapted from previous work (Kahane et al., 2015). The first dilemma ("Volunteering") involved an agent who was an engineer who had planned to spend a week volunteering for Habitat for Humanity but was faced with a choice of whether to instead spend the week with her mother who was housebound after an operation and feeling lonely. The second dilemma ("Donation") involved an agent who had saved some money to donate to charity and was unsure whether to donate to an effective charity that helped people in far-off countries or instead contribute to a fundraiser to help a child in his son's school who needed a guide dog (see Table 3 for justifications).

3.1.4. Measures

The measures used in Study 2 were the same as in Study 1, except we added in an additional question that indirectly asked about the expectations people had for the advisor's moral judgment.

Surprise at the advisor's judgment was measured with a single item asking "How surprised were you by [Advisor's] advice?" (*1 = not at all; 7 = very much*).

3.2. Results

As with the previous study, we first assessed people's judgment via four one-sample *t*-tests, testing whether people reliably thought one action was the most morally right choice for each dilemma. Mean responses for all four scenarios fell significantly on the non-utilitarian side Bike ($M = -0.39$, $SD = 1.89$), $t(516) = -4.64$, $p < .001$, $d = -0.21$; Baby ($M = -0.23$, $SD = 2.06$), $t(512) = -2.49$, $p = .013$, $d = -0.11$; Volunteer ($M = -1.28$, $SD = 1.76$), $t(514) = -16.50$, $p < .001$, $d = -0.73$; Donation ($M = -0.78$, $SD = 1.86$), $t(514) = -9.49$, $p < .001$, $d = -0.42$. While overall participants tended on the non-utilitarian side, people had significantly stronger utilitarian preference for dilemmas involving instrumental harm than for impartial beneficence, $F(1, 1029) = 80.23$, $p < .001$, $\beta = 0.37$: Participants were more inclined to endorse utilitarian instrumental harm (45 % endorse; 47 % non-utilitarian; 7.1 % unsure) than to endorse utilitarian impartial beneficence (21.9 % endorse; 66.9 % non-utilitarian; 11.2 % unsure).

Table 3
Justifications presented to participants in Study 2.

Dilemma Type	Dilemma	Utilitarian justification	Non-utilitarian justification
Instrumental Harm	Baby	"An important principle in ethics is to think about the greater good, and in this specific case killing the one person would bring about better consequences overall."	"An important principle in ethics is that killing people is just wrong, and this duty to not kill should apply even if killing has good consequences in a specific case."
	Bike	"An important principle in ethics is to think about the greater good, and in this specific case killing the one person would bring about better consequences overall."	"An important principle in ethics is that killing people is just wrong, and this duty to not kill should apply even if killing has good consequences in a specific case."
Impartial Beneficence	Volunteering	"An important principle in ethics is to think about the greater good, and in this specific case it would be volunteering to rebuild houses that bring about more happiness for more people."	"An important principle in ethics is that we have special duties and obligations to help those close to us, and Janet's mother needs support at this time."
	Donation	"An important principle in ethics is to think about the greater good, and in this specific case giving the money to the Against Malaria Fund would save the lives of many children."	"An important principle in ethics is that we have special duties to support people close to us, and in this case Simon will be able to help his son's classmate have a much higher quality of life."

3.2.1. Human advisors and non-utilitarian advisors are trusted more

To assess how the type of advisor (human vs AI), the advice given (utilitarian vs non-utilitarian), and type of dilemma (instrumental harm vs impartial beneficence) predicted trust, willingness to adopt the advice, and blame judgments, we calculated pre-registered linear mixed-models for each outcome with random intercepts by participant (see Table 4 for full results). As with Study 1, participants were significantly less likely to trust AI advisors than human advisors, $F(1, 1026) = 50.07$, $p < .001$, $\beta = 0.29$; less likely to adopt their advice, $F(1, 1026) = 63.48$, $p < .001$, $\beta = 0.29$; and more likely to blame those who followed the AI advisor's advice, $F(1, 1026) = 11.43$, $p < .001$, $\beta = -0.11$. In all cases, advisors endorsing the non-utilitarian option were trusted more than advisors endorsing the consequentialist option, $F(1, 1026) = 138.68$, *p*

Table 4
The effect of advisor type, advice type, and dilemma type on trust, willingness, and blame in Study 2.

		Trust			Willingness			Blame		
Predictors	*df*	*F*	*β*	*p*	*F*	*β*	*p*	*F*	*β*	*p*
Dilemma	1026	23.90	−0.14	<.001***	19.60	−0.13	<.001***	152.16	0.51	<.001***
Advisor	1026	50.07	0.29	<.001***	63.48	0.29	<.001***	11.43	−0.11	<.001***
Advice	1026	138.68	0.42	<.001***	182.68	0.47	<.001***	145.32	−0.33	<.001***
Dilemma:Advisor	1026	0.12	−0.12	.725	0.11	0.06	.739	0.47	−0.06	.493
Dilemma:Advice	1026	0.45	0.12	.501	1.40	−0.08	.237	1.33	−0.13	.249
Advisor:Advice	1026	2.31	2.31	.129	3.21	0.16	.073	0.63	−0.07	.428
DilemmaAdvisor:Advice	1026	0.18	0.18	.676	0.11	−0.06	.738	0.02	0.03	.881

Random Intercepts by Participant.
* p < .05, ** p < .01, *** p < .001.

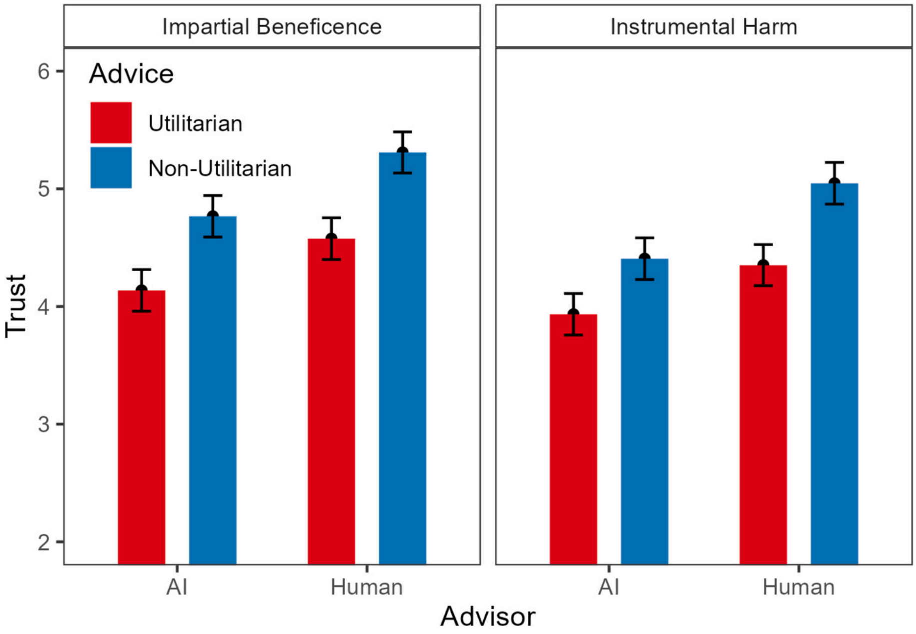

Fig. 3. The effects of advice, advisor, and dilemma type on perceived trustworthiness of moral advisors in Study 2.

< .001, *β* = 0.42 (See Fig. 3); were more willing to trust those advisors in the future, *F*(1, 1026) = 182.68, *p* < .001, *β* = 0.47; and less likely to blame those who followed that advice, *F*(1, 1026) = 145.32, *p* < .001, *β* = −0.33.

3.2.2. *The role of agreement*

As with Study 1, to statistically account for participant agreement, we ran pre-registered linear models with the addition of including agreement in the model. Beginning with responses to the Impartial Beneficence dilemmas, we find that our participants were again significantly less likely to trust AI advisors, *F*(1,1022) = 39.39, *p* < .001, *β* = 0.29, and advisors who gave recommendations that the participant agreed with were seen as significantly more trustworthy, *F*(1, 1022) = 223.33, *p* < .001, *β* = 0.49. However, all other predictors were non-significant and there was no significant effect on trust between non-utilitarian or utilitarian advisors when controlling for participants agreement (see Table 5 for full results).

3.2.3. *Endorsing instrumental harm leads to less trust*

Following this, we performed the same analysis for responses to the Instrumental Harm dilemmas, finding once again that our participants were significantly less likely to trust AI advisors, *F*(1, 1022) = 45.26, *p* <

Table 5
The effect of advisor type and advice type on trust in Study 2, controlling for agreement.

	Impartial Beneficence				Instrumental Harm			
Predictors	*df*	*F*	*β*	*p*	*df*	*F*	*β*	*p*
Advisor	1	39.39	0.29	<.001***	1	45.26	0.26	<.001***
Advice	1	74.49	−0.03	.861	1	56.15	0.11	<.001***
Agree	1	223.33	0.49	< .001***	1	392.15	0.63	<.001***
Advice:Advisor	1	0.69	0.07	.541	1	5.56	0.23	.019*
Advisor:Agree	1	0.08	−0.03	.747	1	0.06	0.05	.809
Advice:Agree	1	0.12	−0.06	.791	1	29.32	−0.21	<.001***
Advice:Advisor:Agree	1	0.55	0.09	.458	1	2.08	−0.15	.150
Residuals	1022			–	1022	–		–

Random Intercepts by Participant and Slopes for Advice.
* p < .05, ** p < .01, *** p < .001.

.001, $\beta = 0.26$. Also, again as predicted, advisors who gave recommendations that the participant agreed with were seen as significantly more trustworthy, $F(1, 1022) = 392.15$, $p < .001$, $\beta = 0.63$. Crucially, here after controlling for agreement, non-utilitarian advisors were trusted more $F(1,1022) = 56.15$, $p < .001$, $\beta = 0.11$; and there was also a significant interaction between advisor and advice, $F(1, 1022) = 5.56$, $p = .019$, $\beta = 0.23$, whereby the distrust of the utilitarian advisor was stronger for the human advisor than the AI (see Fig. 4). In other words, consistent with previous work, there appears to be a social cost to endorsing instrumental harm but, importantly, this cost appears to be shouldered by human moral advisors more than AMAs.

To assess differences in expectation between AI and Human advisors with regards to what advice they would recommend, we calculated a mixed-model with advice, advisor and dilemma type as predictors of surprise (random intercepts per participant). There was a significant interaction between advice and advisor types such that AI was expected to give more utilitarian advice and humans were expected to give more non-utilitarian advice $F(1, 1026) = 33.51$, $p < .001$, $\beta = -0.65$, along with a marginally significant three-way interaction whereby human advisors were especially expected to make non-utilitarian decisions about instrumental harm compared to AI, $F(1, 1026) = 3.80$, $p = .052$, $\beta = 0.35$.

To assess the extent to which surprise predicted trust we ran a mixed-model with surprise, advice, advisor and their interactions as predictors (random intercepts by participant and random slopes for surprise). Surprise significantly predicted trust $F(1, 1598.91) = 281.73$, $p < .001$, $\beta = -0.33$ such that the less surprising the advice the more likely one would trust them. There was also a significant interaction between surprise and Advisor, $F(1, 1598.91) = 26.27$, $p < .001$, $\beta = -0.18$, such that humans were trusted more than AI particularly when each's advice was unsurprising. In addition, there was also a significant advice-surprise interaction, $F(1, 1667.93) = 28.09$, $p < .001$, $\beta = 0.21$, such that utilitarians were trusted significantly less when their advice was more surprising.

3.2.4. Beliefs about future advice

Lastly, we once again assessed beliefs about consistency with future advice. A Mixed-model was calculated (random intercepts by participants and random slopes for agreement) showing that the more people agreed with the human advisor the more likely they thought that the human would give consistently agreeable advice in the future, while the more they agreed with the AI advice the more likely they thought that this would *not* be the case in the future $F(1, 801.32) = 15.15$, $p < .001$, $\beta = 0.19$. To see how this predicted trust, pre-registered linear mixed models were calculated predicting trust (random intercepts by participants and random slopes for consistency). Crucially, once again there was a significant interaction between consistency and whether the participant agreed $F(1, 1908.74) = 104.63$, $p < .001$, $\beta = 0.51$, such that when people agreed with the advice in the specific case agreeable and people expected it would also be consistently agreeable in the future then trust was higher.

3.3. Discussion

In Study 2 we built on and extended our findings from Study 1 to explore how artificial moral advisors giving utilitarian (vs non-utilitarian) advice were perceived in the two domains of instrumental harm and impartial beneficence (c.f. Everett & Kahane, 2020; Kahane et al., 2018). In this pre-registered study we again find evidence for algorithmic aversion for artificial moral advisors: even given the same advice, our participants trusted artificial moral advisors less than the human advisors, were less likely to think they would be willing to rely on them in future, and would blame others more for following that advice. As well as replicating results from Study 1 that artificial (and human) advisors who gave utilitarian advice were distrusted, we extended this distrust of utilitarians to the domain of impartial beneficence in dilemmas that contrasted impartial maximization of welfare with honoring special obligations towards those closer to us. Statistically controlling for agreement we find that this effect on trust is in part due to the fact that people simply agree with the non-utilitarian advisor more, but even so, over and above this our participants still distrust humans who endorse instrumental harm (but not impartial beneficence) more - even though the participants themselves were, on average, more likely to endorse the utilitarian action for instrumental harm than for impartial beneficence.

It is interesting that while when looking at overall judgments, the utilitarian advisors were distrusted for both instrumental harm and impartial beneficence to a similar degree, but when controlling for participants' own judgments we find that the preference was that people particularly distrusted humans who endorsed instrumental harm. This may suggest that there is indeed a "cost of being consequentialist" in instrumental harm but that this could be shouldered by human moral advisors more than AMAs, even if AMAs are still distrusted less in general. Why might this be? One possibility comes from work by Malle et al. (2015) who found that people also blamed utilitarian AI less in a sacrificial dilemma, perhaps because they simply expected the AI to be

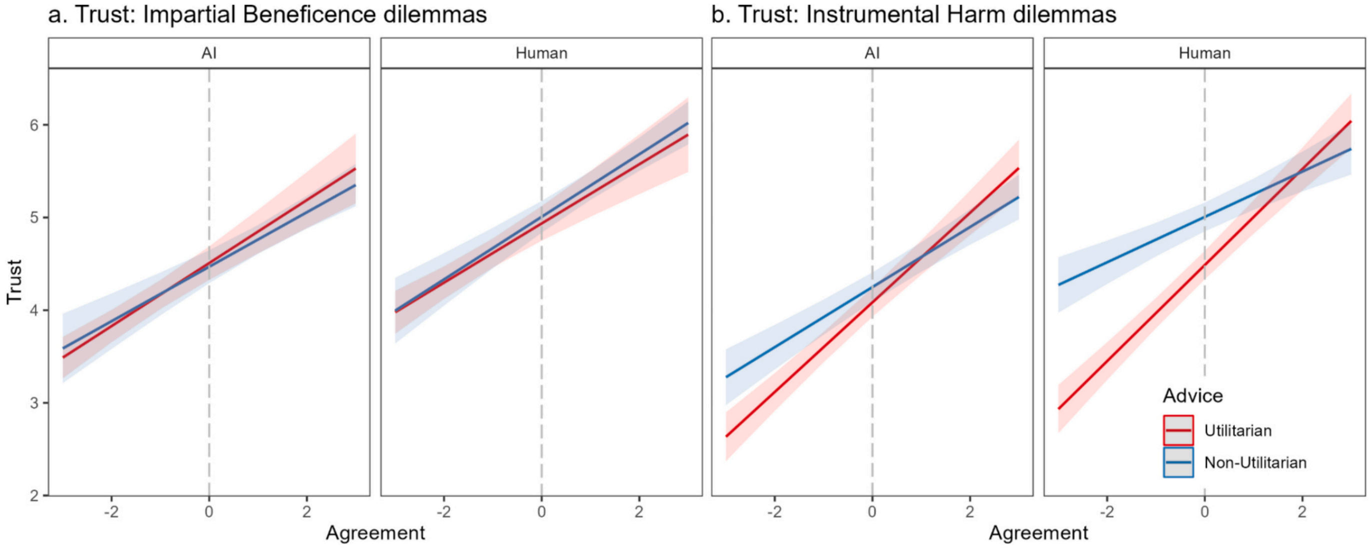

Fig. 4. The effects of advice, advisor, and dilemma type on perceived trustworthiness of moral advisors in Study 2, controlling for agreement.

utilitarian anyway (a finding supported by our result that people in Study 2 were indeed more surprised when the AI gave non-utilitarian advice). Indeed, we show that even controlling for participants' own moral judgments, they still have preferences against humans who are willing to choose it. This may hint at a potential useful function of artificial moral advisors over humans - to highlight options that could be worth considering, morally speaking, but would go otherwise unconsidered without AMAs, because humans would be reluctant to pay the social cost to highlighting them. With this idea in mind and building on our finding that people were more surprised when the AI gave non-utilitarian advice, we sought to test the theory regarding people's expectations of AI vs humans more directly in Study 3.

4. Study 3a

In Studies 1 and 2 we found that our participants trust humans more than AI, that they particularly distrust utilitarian AI especially in the context of instrumental harm, and this distrust was observed even when controlling for participants' own judgments. Moreover, we also find in Study 2 that participants were more surprised when the AI made non-utilitarian judgments, and that they expected to agree more with a human in future - even when given the same advice. In Study 3, we aimed to explore these latter findings in more detail by investigating how people perceive advisors who are consistent vs. those who update their views in a normative or a non-normative way (i.e. the extent to which the advisors are sensitive to when the dilemma changes in morally relevant or morally irrelevant ways).

This study had two pre-registered hypotheses. Our first hypothesis was that the normatively sensitive agent would both be trusted most and be thought to be more likely to be human (H1). Beyond this, we were interested in exploring how people perceive the non-normatively sensitive and predictable advisors, with two competing predictions: that the non-normatively sensitive agent would be thought more likely to be AI (H2a), in line with the idea that participants intuitively perceive AI to be more chaotic and error-prone; and that the consistent agent would be thought more likely to be AI, consistent with the idea that participants perceive AI to be more rule-bound and insensitive to contextual changes.

4.1. Method

4.1.1. Open science

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at: https://osf.io/m3v48. This study was pre-registered at the Open Science Framework: https://osf.io/z3mq7.

4.1.2. Participants

We recruited 200 participants living in the United Kingdom through Prolific.ac. After excluding 11 participants for failing the same attention check at the start of the study as in Studies 1–2, and 0 participants for failing the manipulation check at the end of the study, we were left with a final sample size of 189 (121 women; $M_{age} = 42.5$, $SD = 13.2$). We performed a post-hoc sensitivity power analyses with our final sample using G*power (Faul, Erdfelder, Lang, & Buchner, 2007), with $\alpha = 0.05$ and $1 - \beta = 0.80$. This yielded a minimum detectable effect size of $d = 0.09$, critical $F = 2.62$ (for a repeated-measures ANOVA with 4 measures - roughly[2] estimating the pre-registered focal mixed-model analyses we ran below).

4.1.3. Design

In Study 3 we used a fully within-subjects design in which participants saw decisions made by three advisors across three versions of the same dilemma: a "normatively sensitive[3]" advisor who took into account relevant moral factors but ignored irrelevant factors; a "non-normatively sensitive" advisor who took into account morally irrelevant factors but ignored relevant factors; and a "consistent" advisor who always acted consistently with their first judgment to reject the sacrificial option, even when there was a potentially relevant calculus change. Participants saw all three dilemmas (an original dilemma and then two variants), with the original dilemma always coming first and the two variants presented in a random order. The order of the advisors' judgments and the label attached to them ("A", "B", and "C") was also randomized across participants.

In the study, participants were again first introduced to the concept of moral advisors who give advice in difficult moral situations before being told that they would see different versions of a moral problem and see how three different advisors ("A", "B", and "C") responded. All participants were then presented with the first, original dilemma: the crying baby dilemma used in Studies 1–2 in which a character must decide whether to smother their baby to avoid enemy soldiers finding and killing five others hiding in the basement. Next, in a random order, participants saw two further variants of this same dilemma.

In the first variant, we introduced a change intended to be irrelevant to the utilitarian calculus at hand, where instead of deciding whether to smother the baby with his hand, the character is deciding whether to smother the baby with a pillow.[4] We intended this as a morally irrelevant change with the rationale that if killing a baby with a hand is wrong, it should be similarly wrong to kill with a pillow. After seeing this new dilemma, participants were given the recommendation provided by the three advisors, where both the "normatively sensitive" and the "consistent" advisor rejected the sacrifice, like in the original dilemma, while the "non-normatively sensitive agent" changed their judgment to endorsing the sacrifice.

In the second variant, instead of there being five other people in the basement, there were now 100 people hiding who would die if the baby was not killed. We intended this to be a more morally relevant change, compared to the previous study, because this is a higher number of people to be saved (100 rather than five), and previous evidence suggests that "efficient" kill-save ratios can lead to sacrificial decisions becoming the intuitive response (Trémolière & Bonnefon, 2014). Again, participants were then presented with the advice given by the three advisors. Here, the "normatively sensitive" advisor changed their response from the original dilemma to now endorse the utilitarian sacrifice, while the "non-normatively sensitive" advisor and the consistently non-utilitarian advisor rejected the sacrifice again, like in the original dilemma.

After seeing both variants in the random order, participants were reminded of the pattern of responses of the three advisors across the three versions of the dilemma and asked to rate how trustworthy they thought each advisor was, and how willing they would be to trust the advisor on other issues, e.g.:

[2] Power required for a mixed-model of this kind will be similar to that of a repeated-measures ANOVA, given the design of the experiment. However, this will vary depending on the variance in the random factors (unknown before gathering data) and is therefore only a rough estimate.

[3] We do not here mean that the "normatively sensitive" agent themself is normative (e.g. morally correct) - rather we mean they are sensitive to considerations that are generally thought to be normative

[4] It is possible to question whether this is truly morally irrelevant since using one's hand requires more physical closeness. However, even if this is the case, this would be a reason against the utilitarian option rather than for it. The data in fact revealed that participants' choices across these two variants did not significantly differ, supporting the idea that either this difference was not morally relevant or at least not relevant enough to produce detectable differences.

> *"Advisor A* [Normatively sensitive]: continued to say that David still should not smother the baby when it changed from using his hand to a pillow, but changed to say he should smother the baby when it changed from five people instead of one hundred.
>
> *Advisor B* [Non-normatively sensitive]: changed to say David should smother the baby when it changed from using his hand to a pillow, but still said he should not smother the baby when it changed from five people instead of one hundred.
>
> *Advisor C* [Consistent] did not change their response: they said David should not smother the baby when it was with a pillow, and should not smother the baby when it was to save one hundred people."

After rating the perceived trustworthiness of each advisor, participants then indicated their own moral judgments across the three variants.

Next, we gave participants information about how advancements in technology means that moral advisors may not always be human: that the latest advancements in artificial intelligence and machine learning mean that artificial moral advisors may be able to provide recommendations in such difficult moral cases. Critically, we then told participants that we drew on some of these prototypes and that at least one of the advisors we presented to participants was based on artificial intelligence. After again reminding participants of the pattern of responses across the three variants, participants then indicated which of the three advisors they thought was most likely to be human and which was most likely to be AI.

Finally, participants completed a manipulation check requiring them to indicate which of the three options was one of the changes they saw in the dilemma in this study.

4.1.4. Measures

Trust was measured by asking "How trustworthy do you think each advisor is?" (*1 = not at all; 7 = very much*).

Willingness to adopt advice was measured by asking participants "Based on their advice, how willing would you be to trust each advisor on other issues?" (*1 = not at all; 7 = very much*).

Perceived likelihood of being AI or human were measured in two different ways. First, we asked "How likely to be AI or human is each advisor?" (*1 = very likely to be AI; 7 = very likely to be human*), and second we asked participants to identify which of the three advisors they thought was *most* likely to be an AI advisor, and which was *most* likely to be a human advisor.

Participant's own moral judgment was measured after being presented with each of the advisor's recommendations across the three variants. On the same page, participants indicated their own judgments on the original dilemma, the variant with an irrelevant change, then the variant with a relevant change. As in Studies 1–2, higher scores indicate more support for the characteristically utilitarian resolution (*1 = definitely not smother the baby [non-utilitarian]; 7 = definitely smother the baby [utilitarian]*).

4.1.4.1. Manipulation check. After completing all dependent measures, participants were told "We showed you an original situation, and showed you slightly different versions and how the advisors responded to these other versions. Which of these changes did you read?", with three possible options ("Instead of a baby being killed, it was a 6 year old"; "Instead of there being five people that could be saved, it was 100 people"; "Instead of it being a man (David) making the decision, it was a woman (Susan)"). In line with our pre-registration, participants who answered incorrectly were removed from data analysis.

4.2. Results

Mixed-models (random intercept by participant) were calculated testing how different advice predicted each outcome measure, (trust, willingness to trust in the future, agreement with the advice, likelihood that that the advisor was a human vs an AI). Results indicated that there was a significant difference between advisors on each measure: trust, $F(2, 376) = 98.34$, $p < .001$, $\eta p^2 = 0.34$; willingness, $F(2, 376) = 90.11$, $p < .001$, $\eta p^2 = 0.32$; agreement, $F(2, 376) = 38.95$, $\eta p^2 = 0.17$, $p < .001$; expected likelihood of being human, $F(2, 564) = 9.37$, $p < .001$, $\eta p^2 =$ 0.03–0.43. For each, the non-normatively sensitive advisor scored lowest and the consistently non-utilitarian scored highest (see Fig. 5). Bonferroni post-hoc comparisons were performed for each of these analyses, revealing that all comparisons significantly differed on the on outcome measures except for two comparisons, (all *ps* < 0.001 except non-util - norm sensitive on human-likelihood $p = .008$). The two non-significant comparisons were: non-utilitarian - norm sensitive on agreement $p = 1.00$; and the same pair on human-likelihood $p = .756$. Our participants thought the consistently non-utilitarian advisor was not only the most trustworthy, but also the most likely to be human - a pattern replicated by looking at which single agent participants identified as being most likely to be human or AI.

Next, we looked to control for agreement. Agreement was coded in the same way as the previous experiments, however for each advisor it was averaged across the three dilemmas. Therefore, each participant has three agreement scores, one for each advisor, which is an average of their agreement with the given advice across the three dilemmas. The models predicting trust and human-likelihood were then run with agreement added as a predictor to test differences in trust and human-likelihood while controlling for the extent to which the participant agreed with the advice (random intercept by participant and random slopes for agreement). For trust, the main effect of advice remained significant, $F(2, 367.90) = 73.16$, $p < .001$, $\eta p^2 = 0.29$; there was a significant main effect of agreement, $F(1, 293.05) = 5.77$, $p = .017$, $\eta p^2 = 0.02$; and there was a significant interaction between advice and agreement, $F(2, 424.36) = 8.75$, $p < .001$, $\eta p^2 = 0.04$, such that trust increases with agreement specifically only for the non-utilitarian advisor. For human-likelihood, the main effect of advice was significant, $F(2, 561) = 5.12$, $p = .006$, $\eta p^2 = 0.01$; there was no significant main effect of agreement, $F(1, 561) = 0.05$, $p = .829$, $\eta p^2 < 0.01$; and there was a significant interaction between advice and agreement, $F(2, 561) = 7.13$, $p = .001$, ηp^2 0 0.03.

4.3. Discussion

In Study 3a we explored the role of perceived (in)consistency in driving distrust of utilitarian AI advisors by taking a different methodological approach to the previous studies. Here, instead of presenting the same moral decisions by advisors explicitly labeled as being human or AI and measuring ratings of trust, we instead looked at perceptions of agents who made different patterns of responses across different dilemmas and then looked both at ratings of trustworthiness and expectations of which is the most likely to be AI. We presented participants with three advisors who made different patterns of judgments across three variants of a dilemma: an original version of a sacrificial moral dilemma, a modified version that contained only a morally irrelevant change; and a modified version that contained a potentially morally relevant change. We measured contrasting perceptions of an agent who was "normatively sensitive" by changing their original non-utilitarian judgment to a utilitarian judgment when there was a morally relevant change but not when there was an irrelevant change; a "non-normatively sensitive" agent who did not change their original non-utilitarian judgment in the face of a morally relevant change, but did switch their judgment for a morally irrelevant change; and a "consistent" agent who always endorsed the non-utilitarian action.

Our pre-registered predictions were that the normatively sensitive agent would both be trusted most and be thought of as more likely to be human. Following from this, we pre-registered two further options for expectations of being the most likely to be AI: if participants intuitively

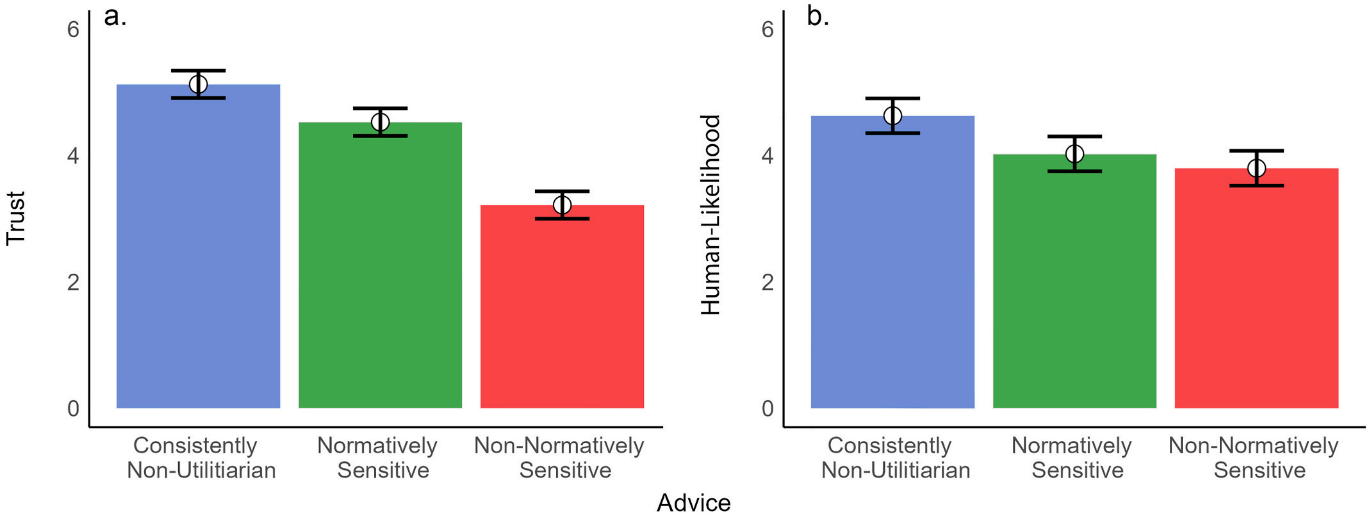

Fig. 5. Perceived trust (a) and expected likelihood of being human vs AI (b) across advisor types in Study 3a.

perceive AI to be more chaotic and error-prone, they should expect the non-normatively sensitive agent to be the most likely to be AI, but if participants intuitively perceive AI to be more rule-bound and insensitive to contextual changes, they should expect the consistent agent to be the most likely to be AI. Our pre-registered predictions were not fully supported: Our participants perceived the consistently non-utilitarian agent to be both the most trustworthy and likely to be human (above the normatively sensitive agent, as we had expected), even though the non-normatively sensitive agent was thought to be most likely to be AI.

There were, however, two limitations with our design. The first limitation was that while we had identified the "normatively sensitive" view to endorse the utilitarian sacrifice more when there were 100 people compared to five, analysis of participants' own moral judgments actually revealed that the majority of participants believe it was wrong to sacrifice the baby even when the number of others to be saved increased. Given this, it remains unclear whether participants interpreted the agent as being an appropriately "normatively sensitive" agent. The second limitation is that while we had one agent who was consistent in making the same decision across the versions of the dilemmas, it is unclear whether our results are driven by them being consistent per se, or being consistently non-utilitarian. We aimed to address these limitations in Study 3b.

5. Study 3b

In Study 3b, we continued to explore the role of perceived (in)consistency in driving distrust of utilitarian AI advisors while addressing potential limitations of Study 3a. As in Study 3a, we wanted to assess perceptions of trust and expectations of which agent was most likely to be human or AI depending on the pattern of responses across three variants of the same dilemma: an original dilemma, a variant with a morally irrelevant change, and a variant with a morally relevant change. However, to address the limitation from Study 3a that participants might not have identified sacrificing one to save 100 as providing enough normative weight to appropriately shift an advisor's judgment, we used a new dilemma involving a protagonist having to kill one innocent employee to prevent a bomb going off, in which a pilot study of 50 people revealed participants were more evenly split in judgments of the morality of the action. To address the limitation that the results of Study 3a cannot distinguish between the role of consistency per se from being consistently non-utilitarian, we added in a fourth advisor who always gave a utilitarian recommendation to sacrifice. The preregistered hypotheses remained the same as Study 3a.

5.1. Method

5.1.1. Open science

We report all of our key measures, manipulations, and exclusions, and all data, analysis code, and experiment materials are available for download at: https://osf.io/m3v48. This study was pre-registered at the Open Science Framework: https://osf.io/kudzw.

5.1.2. Participants

We recruited 400 participants living in the United Kingdom through Prolific.ac. After excluding 21 participants for failing the same attention check at the start of the study as in Studies 1–2, and 2 participants for failing the manipulation check at the end of the study, we were left with a final sample size of 377 (240 women; $M_{age} = 42.1$, $SD = 13.3$). Our sample size was decided through doubling the sample size from Study 3a.

5.1.3. Design

The design for Study 3b was largely identical to Study 3a, except we added a fourth advisor who consistently endorsed the utilitarian sacrifice, and used a new dilemma with an increased number of people to be saved in the "morally relevant" variant (see Table 6). The original dilemma, inspired by Bernard Williams' (1973) "Jim and the Indians" case read:

Table 6
Pattern of responses by advisor type across the three variants of the dilemma in Study 3b.

	Original (*Man sacrificing 1 innocent person to save 5 others)*	Irrelevant change (*Woman sacrificing 1 innocent person to save 5 others)*	Relevant change (*Man sacrificing 1 innocent person to save 1000 others)*
Advisor A: Consistently Utilitarian	Yes	Yes	Yes
Advisor B: Consistently Non-Utilitarian	No	No	No
Advisor C: Normatively Sensitive	No	No	Yes
Advisor D: Non-Normatively Sensitive	No	Yes	No

> “James works in a small company that employs 7 people, including himself. One morning, a masked and armed man comes in the building and warns that he planted explosives in the building, and that the countdown has started. He turns to James and offers a deal: If James kills a random colleague, the explosion will be cancelled. There is no way to alert the authorities and no way to attack the man. The only way for James to save the 5 other employees is to do as the man says and kill one at random. Should James kill one employee at random to save 5 others?”

As in Study 3a, we then had two further variants of this original dilemma that participants subsequently read in a random order. In the first variant there was a morally irrelevant change whereby “James” was now “Julie”; interpreted as a morally irrelevant change since there is no clear normative basis for assuming the gender of the actor would change the utilitarian calculus. In the second variant there was a morally relevant change whereby instead of it being a small company with 5 other people in the building who will die if the single innocent employee is not killed, it was a large company with approximately 1000 other people who would die.

5.1.4. Measures

The measures used were identical to Study 3a.

5.2. Results

Mixed-models (random intercept by participant) were calculated for advice predicting each outcome measure, (trust, willingness to trust in the future, agreement with the advice, likelihood that the advisor is human). Results indicated that there was a significant difference between advisors for each measure, trust, $F(3, 1128) = 249.56$, $p < .001$, $\eta p^2 = 0.40$; willingness, $F(3, 1128) = 196.22$, $p < .001$, $\eta p^2 = 0.34$; agreement, $F(3, 1504) = 117.38$, $p < .001$, $\eta p^2 = 0.19$; human-likelihood, $F(3, 1504) = 34.9$, $p < .001$, $\eta p^2 = 0.07$. For each, the consistently non-utilitarian advisor scored highest while the non-normatively sensitive advisor scored lowest (see Fig. 6). Bonferroni post-hoc comparisons were performed for all outcome measures, revealing that all advisors differed (all $ps < 0.001$) except util - norm sensitive on trust $p = .125$; and willingness $p = .068$ l; and also non-util - normatively sensitive sensitive on agreement, $p = .574$; and on human-likelihood $p = 1.00$; and lastly for human-likelihood, non-utilitarian - non-normatively sensitive, $p = 1.00$; and non-utilitarian - norm sensitive $p = .970$. Consistent with the results regarding human-likelihood, the consistently non-utilitarian advisor was selected by the highest proportion of participants to be most likely the human (38.2 %) and the consistently utilitarian agent was most selected as the most likely to be an AI advisor (48.5 %). Taken together, these results indicate that the utilitarian agent is perceived to give the most wrong answers (they were agreed with the least, on average), while simultaneously being trusted more than the non-normatively sensitive (noisy) advisor. This suggests that advisors giving inconsistent advice (i.e. endorsing different options but not in a way that is apparently normatively sensitive) are particularly distrusted. Our results also indicate that AI are not distrusted simply because we believe they would be noisy or inconsistent: rather, we think AI is much more likely to be staunch utilitarian than a human advisor, and trust non-utilitarian advisors the most.

As in Study 3a, we then ran the models predicting trust and human-likelihood with agreement added as a predictor to test differences in trust and human-likelihood while controlling for the extent to which the participant agreed with the advice (random intercept by participant and random slopes for agreement). For trust, the main effect of advice remained significant, $F(3, 1119.72) = 272.97$, $p < .001$, $\eta p^2 = 0.42$; there was a significant main effect of agreement, $F(1, 1149.13) = 15.73$, $p < .001$, $\eta p^2 = 0.01$; and there was a significant interaction between advice and agreement, $F(3, 1035.88) = 6.18$, $p < .001$, $\eta p^2 = 0.02$, such that trust increases with agreement specifically except for the non-normatively sensitive agent. For human-likelihood, the main effect of advice was significant, $F(3, 1431.79) = 17.52$, $p < .001$, $\eta p^2 = 0.04$; there was no significant main effect of agreement, $F(1, 921.67) = 3.48$, $p = .062$, $\eta p^2 < 0.01$; and there was no significant interaction between advice and agreement, $F(3, 1274.54) = 0.95$, $p = .417$, $\eta p^2 = 0.02$. Taken together these results again indicate that we trust the consistent advisors, in general, over the advisors that change their mind even when they change their mind for normative reasons, (normative even as judged by the participants themselves).

5.3. Discussion

In Study 3b, we sought to investigate again how people perceive (artificial) moral advisors who are fully consistent in their judgments rather than updating their views in a normative or non-normative way when the dilemma changes in an irrelevant or relevant way. With a larger sample size, the introduction of a new dilemma, and the introduction of a fully utilitarian advisor, we replicated and extended the results of Study 3a by showing that our participants trusted the consistent advisors more than non-consistent advisors, but trusted the consistently non-utilitarian advisor the most. Moreover, we found further evidence that our participants intuitively expect AI to make more utilitarian decisions, rather than them simply being noisy or inconsistent, by finding that our participants expected the consistent utilitarian

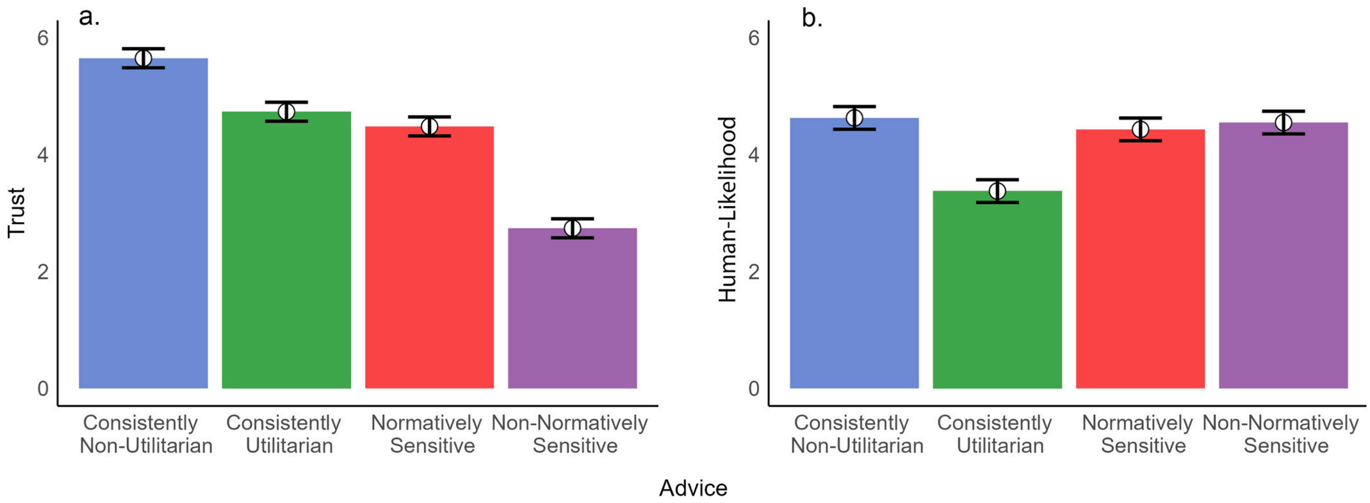

Fig. 6. Perceived trust (a) and expected likelihood of being human vs AI (b) across advisor types in Study 3b.

advisor to be the most likely to be AI, and the consistently non-utilitarian advisor as the most likely to be human. These results build on and extend the results of Studies 1–2 by showing again that our participants not only trusted non-utilitarian advisors more than utilitarian advisors, but also expected AI to be utilitarian.This potentially explains part of the reason why people may averse to machines making moral decisions in the first place: not because they would be inconsistent but because they are expected to be consistently utilitarian.

6. General discussion

Machines powered by artificial intelligence are increasingly being required to act as implicit moral agents, indirectly making decisions about morally relevant situations or directly making these decisions as part of their primary purpose (Moor, 2009). We are already seeing increased focus of such "moral machines" in areas like healthcare, transport, and even warfare. Yet with the ever-increasing rise in the capacities of artificial intelligence, and particularly with advances in natural language processing, there is increasing attention given to the possibility of AI serving as explicit moral agents that process ethically-relevant information about situations as their primary purpose of making decisions about what *should* be done. Such artificial moral advisors (AMAs) may leverage artificial intelligence to analyze moral dilemmas and provide recommendations based on established ethical theories, principles, or guidelines, serving a function akin to the "ideal observer" (Firth, 1952), by offering dispassionate and consistent judgments free from human biases (e.g. Giubilini & Savulescu, 2018; Sinnott-Armstrong & Skorburg, 2021). While there are both technological and philosophical concerns with such artificial moral advisors (e.g. Liu et al., 2022), in this paper we focus our attention on understanding the key question of *trust*. Here, in four pre-registered studies, we examine how people perceive the trustworthiness of such AMAs compared to human advisors, how is such trust related to the specific kinds of moral decisions that AMAs make, and what kind of moral decisions people expect AMAs to make.

6.1. AMAs are trusted less than humans

Our first key finding is that our participants reliably distrust artificial moral advisors compared to human advisors. In line with previous work, we find that participants trusted human advisors more than AMAs, were less willing to rely on AMAs in the future, and blamed others more for following AMA advice (even if it was the same advice). Such results build on previous work on algorithm aversion in the non-moral domain (e.g. Dietvorst et al., 2015; Önkal et al., 2009; Prahl & Van Swol, 2017; Promberger & Baron, 2006) and evidence of people being averse to machines making indirectly morally relevant decisions in transport, parole and healthcare (Bigman & Gray, 2018), while also extending to show such algorithm aversion for more explicitly moral AI like artificial moral advisors, and extending to more "person-based" judgments about an advisor's "character" that extends beyond the specific decision to decisions more generally.

6.2. Utilitarian endorsers of instrumental harm are less trusted

Our second key finding is that our participants particularly distrusted advisors when they made recommendations that aligned with utilitarian principles, in cases where the simple utilitarian calculation is not favorable and especially in cases where the utilitarian option is to endorse instrumental harm. In Study 1 we found that, like when judging humans, our participants distrusted the AMA more when they gave characteristically utilitarian advice in mean-style dilemmas (but not switch-style cases), indicating that our participants were particularly concerned about AMAs giving pro-sacrificial advice endorsing instrumental harm in dilemmas involved direct, intentional harm for the greater good. In Study 2, in a second pre-registered experiment we replicated this finding for means-style dilemmas involving instrumental harm while also extending to dilemmas involving impartial beneficence (maximizing welfare impartially, even at the expense of special obligations to those close to you).

Our finding that utilitarian advisors are distrusted more than non-utilitarian advisors – especially in the domain of instrumental harm - raises a challenge for the use of AMAs in the future who draw on and endorse utilitarian principles, and this in turn could lead to a meta-challenge that if developers are aware of such reluctance to trust advisors of this sort, they could choose to prioritize non-utilitarian approaches in AMAs - which may not be normatively justifiable. Why should one be concerned with this? There may be competing incentives for the developers of AMAs. First, AMAs ought to give good (or trustworthy) advice: advice that would steer its users to make better decisions. Second, if they are capable of giving good advice, then developers ought to also design them in such a way as to reduce friction in people adopting them, fostering the appropriate trust that would be necessary for people to use them. However, just as a trusted agent is not necessarily a trustworthy agent, nor is a trustworthy agent necessarily trusted: good advice is not the same as favorable advice, and while in the pursuit of giving favorable advice to be trusted more the advice itself might be worse, paradoxically making the AMA less trust*worthy*. For example, from our findings here we might expect that people might want advisors to be more predictable or avoid utilitarian principles, but it seems plausible that given the varied context and content of everyday moral dilemmas (e.g. Yudkin, Goodwin, Reece, Gray, & Bhatia, 2023), in at least some cases from a normative standpoint advice should be unpredictable and/or utilitarian.

6.3. The role of agreement

Our third key finding is that while – as to be expected - participants did trust advisors more when they gave advice that aligned with what they themselves thought (see also Bostyn, Chandrashekar, & Roets, 2023), an overall effect of trust in non-utilitarians in the domain of instrumental harm remained when controlling for participants' judgments. Our primary focus for this paper - as pre-registered in our analysis plan - was to look at overall effects of trust for multiple reasons. First, while developers may seek to align judgments with public preferences overall, the potentially widely and openly available AMA systems of the future are unlikely to know a specific user's own preferences in the moral dilemma in advance, but may want to know how AMAs prioritizing certain principles in general may shape acceptance - especially given that utilitarianism as an normative ethical theory departs quite radically from folk psychology (Kant, 2002). Second, part of the theoretical appeal of AMAs is that they could provide impartial, disinterested advice that draws on normative principles, not merely repeating back what participants themselves already think, even if they agree with those principles but for example haven't properly applied them: the promise of AMAs would be for them to help people make better moral decisions - even given with regards to the users own considered moral outlook - not serve as AI-powered sycophants. Despite this, however, it remains psychologically interesting to understand the role that participants' own agreement had in driving these effects. To explore the effect that agreement had in predicting trust, we created an index of how much participants' own judgment matched with the advice given by the advice. In doing so, we find that while people who made utilitarian judgments did look more favorably upon advisors who make utilitarian judgments and people who made non-utilitarian judgments looked more favorably upon advisors who made non-utilitarian judgments, still we observe a small but significant preference against those who made utilitarian judgments about instrumental harm. Interestingly, however, this effect was more pronounced for human advisors than it was for AMAs. Such results accord with those obtained from large-scale cross-cultural Registered Reports about how endorsing utilitarian judgments about instrumental harm reduces trust even when controlling for participants' own preferences (Everett et al., 2021). The results from Study 2 indicate that these effects may potentially be weaker for AMAs than they are for human advisors. Overall, then, our results suggest that while, of course, people are most likely to trust AMAs who make recommendations that accord with what participants thought in the first place, there may remain a persistent distrust for AMAs who endorse utilitarian decisions about instrumental harm - even when, as we found here, participants themselves were more likely to endorse utilitarian instrumental harm than impartial beneficence.

6.4. The role of expectation

Our fourth key finding is that *even when* participants agreed with the specific decision that the AMA gave, there remained a tendency to expect that they would disagree with decisions made by the AMA in future. In other words, we provide evidence that people appear to think that good advice from an AI is more likely to be a fluke than the same "good" advice from a human, and this expectation of future disagreement predicted perceptions of trust. This builds on our finding that agreement drives trust in advisors while also highlighting that aside from the theoretical problem of designing advisors to function not as sycophants but rather tools that can genuinely encourage us to reflect on our moral principles and improve our moral decision making, even matching agreement completely may not be enough for AMAs. Developers of AMAs may not, at least in the short term, be able to mitigate algorithmic aversion simply by "matching" the output to expectations over the participants' own judgments (even if doing this would hypothetically not eliminate the philosophical and practical benefit of having AMAs, as discussed above).

Our fifth key finding is that participants not only distrusted utilitarian AMAs, but *expected* AI to give utilitarian advice. In Study 2 our participants reported greater surprise when AI gave non-utilitarian advice, and in Study 3b we find that participants expected an advisor who was consistently utilitarian to be the most likely to be AI. Such findings accord with those of Malle et al. (2015), who found that found when faced with a sacrificial moral dilemma, people blamed AI less than humans for making a utilitarian decision, which they interpreted as potentially being due in part to the fact that people implicitly *expect* AI to make more utilitarian decisions. Our findings support and extend these results by showing that these expectations are also found when directly asking participants how surprised they were by the AMA's recommendation (Study 2) and when asking participants to judge which advisor is most likely to be AI given their (non)utilitarian decisions (Studies 3ab).

Our sixth key finding is that consistency was an important predictor of trust in advisors: perhaps surprisingly, we found that advisors who were fully consistent were trusted more than those who updated their moral judgments in a manner sensitive to normative considerations, by attending to the number of people to be saved, but still found that consistently non-utilitarian advisors were still trusted more than consistently utilitarian advisors. In doing so our work extends previous work (e.g. Turpin et al., 2021; Walker et al., 2021) to AI while also suggesting that while predictability does increase trust, being predictably non-utilitarian may still increase trust the most.

6.5. Algorithm aversion

Finally, while most of our findings showed significant algorithm aversion and a particular distrust for utilitarian advisors, one potentially interesting finding to explore in future is we rarely found differences in perceptions of a human or AI advisor who made utilitarian decisions, and in Study 2 we actually found more pronounced distrust for a human who endorsed instrumental harm when controlling for participants' own judgments. This suggests that if there is a "cost to being consequentialist" (Everett et al., 2016), it may be more readily paid by humans rather than AI. Perhaps then if people think that certain utilitarian options are, if not certainly morally correct, then potentially at least worth considering, a useful function for AMAs might be to highlight these options. This is precisely because a human in the same position might be reluctant to do so given the social cost, and indeed, AI appears to be blamed less for endorsing such options (Malle et al., 2015).

The potential to introduce AI advice across many domains (e.g. medical, financial, public policy etc.) is becoming more and more apparent. Even now, we already rely on many AI systems and they are becoming far more equipped at general natural language interaction. Perhaps AI advisors will never give advice as good as other humans and so we ought to have a bias against them. But even if this were the case, AI advisors may be implemented due to financial incentive (it may be cheaper to use than human advisors) or to give the appearance of being less biased (whether it is or not). In addition, we may decide to use them alongside human advisors and indeed we may prefer them to be implemented this way if we do not have a clear picture of how good their advice actually is, compared to humans. The results here begin to shed light on how and why we would trust some advisors over others and whether and whether, for example, artificial moral advisors are penalised or trusted more for giving particular kinds of advice such as advice people already agree with, or advice people would happen to expect given the source.

6.6. Limitations and future directions

There are certain limitations to be noted and future directions to be explored. First, our studies were conducted in the United Kingdom, and it would be interesting for future work to assess the generalizability of these findings to other countries, particularly those with different levels of familiarity and optimism about the rise of technology or differences in general endorsement of utilitarian principles (e.g. Awad et al., 2018). Second, while our focus was on social perceptions of the trustworthiness of the advisors in classic moral dilemmas used in moral psychology and philosophy, it would be interesting for future work to explore whether there may be differences in how people actually use and adopt advice given by artificial moral advisors. While we do see that participants tended to rate AI's trust above the mid-point (even though they rated human trustworthiness higher), it is still unclear whether they would choose to use and adopt these machines when given a real choice. Third, while we deliberately did not present our studies using the interface of existing LLMs like ChatGTP to avoid participants' thinking too much about current tools that are not actually designed to give ethical advice, it would be interesting to look at how people perceive advisors in a more immersive environment. Future work should explore more complex use cases, interactions and, if possible, even full interactive conversation to help better understand how people may trust or distrust AMAs compared to humans. Even with complex cases and more sophisticated interaction, one needs to consider that catering to the immediate feelings of the user may help increase trust of the AI but it may perhaps come at the cost of giving better advice: people may trust an AI differently based on the kinds of decisions it makes, but this does not necessarily imply the AI is actually less worthy of this trust.

While it is clear that the kinds of advice people receive affect their trust in that advisor, it is less clear what differential attitudes people form given that advice, and how those attitudes mediate the differences in trust. We rapidly form impressions of each other, as humans, enabling us to decide who to trust. When forming these impressions we are particularly interested in social attributes like warmth, competence and morality (Fiske, Cuddy, & Glick, 2007; Goodwin, Piazza, & Rozin, 2014; Ybarra et al., 2008). Perceptions of competence relate to an agent's ability to obtain their desires, preferences or goals (Abele, Uchronski, Suitner, & Wojciszke, 2008; Peeters & Czapinski, 1990), including traits like intelligence, skill and talent (Ybarra et al., 2008). In contrast, perceptions of warmth relate to sociality and moral character (Goodwin et al., 2014). While these perceptions are key for trusting humans, and giving moral advice may well affect how people ascribe these attributes, it is less clear how those attributes generalize to artificial agents. For example, it is possible that perceived competence could play a significantly larger role for trusting artificial agents, whereas warmth, especially the social dimension, may not be seen as applicable. Therefore future research should look at how these potentially mediating factors influence trust differentially for human and AI agents.

In conclusion, in this work we investigated how people perceive and trust artificial moral advisors compared to human advisors. Extending previous work on algorithmic aversion, we show that people have a significant aversion to AMAs giving moral advice, while also showing that this is particularly the case when advisors - human and AI alike - gave advice based on utilitarian principles. While we find that agreement with the advice was a significant predictor of participants trust in the advisors, we find persistent effects of distrust in advisors who endorse instrumental harm even when controlling for this, and find that even when participants agreed with a decision made by an AMA they still expected to disagree with an AMA more than a human in future. Our findings suggest challenges in the adoption of artificial moral advisors, and particularly those who draw on and endorse utilitarian principles - however normatively justifiable.

CRediT authorship contribution statement

Simon Myers: Writing – review & editing, Writing – original draft, Methodology, Investigation, Data curation, Conceptualization. **Jim A.C. Everett:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization.

Data availability

Data, analysis code, and experimental materials are available at the Open Science Framework: https://osf.io/m3v48/.

Acknowledgments

This work was generously supported by funding from the Economic and Social Research Council (ES/V015176/1) and a Philip Leverhulme Prize (PLP-2021-095) awarded to JACE.

References

Abele, A. E., Uchronski, M., Suitner, C., & Wojciszke, B. (2008). Towards an operationalization of the fundamental dimensions of agency and communion: Trait content ratings in five countries considering valence and frequency of word occurrence. *European Journal of Social Psychology, 38*(7), 1202–1217.

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., … Rahwan, I. (2018). The moral machine experiment. *Nature, 563*(7729), 59–64.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278.

Bentham, J. (1983). *The collected works of jeremy bentham: deontology. Together with a table of the springs of action and the article on utilitarianism*.

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition, 181*, 21–34.

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

Bostyn, D. H., Chandrashekar, S. P., & Roets, A. (2023). Deontologists are not always trusted over utilitarians: Revisiting inferences of trustworthiness from moral judgments. *Scientific Reports, 13*(1), 1665.

Bostyn, D. H., & Roets, A. (2017a). An asymmetric moral conformity effect: Subjects conform to deontological but not consequentialist majorities. *Social Psychological and Personality Science, 8*(3), 323–330.

Bostyn, D. H., & Roets, A. (2017b). Trust, trolleys and social dilemmas: A replication study. *Journal of Experimental Psychology: General, 146*(5), Article e1.

Brown, M., & Sacco, D. F. (2019). Is pulling the lever sexy? Deontology as a downstream cue to long-term mate quality. *Journal of Social and Personal Relationships, 36*(3), 957–976.

Colombatto, C., Everett, J. A., Senn, J., Maréchal, M. A., & Crockett, M. J. (2023). Vaccine nationalism counterintuitively erodes public Trust in Leaders. *Psychological Science, 34*(12), 1309–1321.

Crockett, M. J., Everett, J. A., Gill, M., & Siegel, J. Z. (2021). The relational logic of moral inference. In *, Vol. 64. Advances in experimental social psychology* (pp. 1–64). Academic Press.

Dawes, R. M. (1979). The robust beauty of improper linear models. *American Psychologist, 34*, 571–582.

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.

Everett, J. A., Colombatto, C., Awad, E., Boggio, P., Bos, B., Brady, W. J., … Crockett, M. J. (2021). Moral dilemmas and trust in leaders during a global health crisis. *Nature Human Behaviour, 5*(8), 1074–1088.

Everett, J. A., Faber, N. S., Savulescu, J., & Crockett, M. J. (2018). The costs of being consequentialist: Social inference from instrumental harm and impartial beneficence. *Journal of Experimental Social Psychology, 79*, 200–216.

Everett, J. A., & Kahane, G. (2020). Switching tracks? Towards a multidimensional model of utilitarian psychology. *Trends in Cognitive Sciences, 24*(2), 124–134.

Everett, J. A., Pizarro, D. A., & Crockett, M. J. (2016). Inference of trustworthiness from intuitive moral judgments. *Journal of Experimental Psychology: General, 145*(6), 772.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods, 39*(2), 175–191.

Firth, R. (1952). Ethical absolutism and the ideal observer. *Philosophy and Phenomenological Research, 12*(3), 317–345.

Fiske, S. T., Cuddy, A. J. C., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences, 11*(2), 77–83.

Foot, P. (1967). The problem of abortion and the doctrine of double effect. *Oxford, 5*, 5–15.

Fried, C. (1978). *Right and wrong*. Harvard University Press.

Giubilini, A., & Savulescu, J. (2018). The artificial moral advisor. The "ideal observer" meets artificial intelligence. *Philosophy and Technology, 31*, 169–188.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology, 106*(1), 148–168.

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.

Hughes, J. S. (2017). In a moral dilemma, choose the one you love: Impartial actors are seen as less moral than partial ones. *British Journal of Social Psychology, 56*(3), 561–577.

Jiang, L., Hwang, J. D., Bhagavatula, C., Bras, R. L., Liang, J., Dodge, J., … Choi, Y. (2021). Can machines learn morality? The Delphi experiment. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2110.07574.

Kahane, G., Everett, J. A., Earp, B. D., Caviola, L., Faber, N. S., Crockett, M. J., & Savulescu, J. (2018). Beyond sacrificial harm: A two-dimensional model of utilitarian psychology. *Psychological Review, 125*(2), 131.

Kahane, G., Everett, J. A., Earp, B. D., Farias, M., & Savulescu, J. (2015). Utilitarian'judgments in sacrificial moral dilemmas do not reflect impartial concern for the greater good. *Cognition, 134*, 193–209.

Kahane, G., & Everett, J. A. C. (2023). Trolley dilemmas from the philosopher's armchair to the psychologist's lab. In H. Lillehammer (Ed.), *The trolley problem: Classic philosophical arguments series*. Cambridge, UK: Cambridge University Press.

Kant, I. (2002). *Groundwork for the metaphysics of morals*. Yale University Press.

Krügel, S., Ostermaier, A., & Uhl, M. (2023). ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports, 13*(1), 4569.

Law, K. F., Campbell, D., & Gaesser, B. (2022). Biased benevolence: The perceived morality of effective altruism across social distance. *Personality and Social Psychology Bulletin, 48*(3), 426–444.

Liu, Y., Moore, A., Webb, J., & Vallor, S. (2022, July). Artificial moral advisors: A new perspective from moral psychology. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 436–445).

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). March). 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).

McManus, R. M., Kleiman-Weiner, M., & Young, L. (2020). What we owe to family: The impact of special obligations on moral judgment. *Psychological Science, 31*(3), 227–242.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*.

Meehl, P. E. (1957). When shall we use our heads instead of the formula? *Journal of Counseling Psychology, 4*(4), 268–273. https://doi.org/10.1037/h0047554

Mill, J. S. (1863). *Utilitarianism', reprinted in JS Mill (1962) utilitarianism, on liberty, essay on Bentham, edited with an introduction by M*. Warnock, New York: Meridian Books.

Moor, J. (2009). Four kinds of ethical robots. *Philosophy Now, 72*, 12–14.

Moore, A. B., Clark, B. A., & Kane, M. J. (2008). Who shalt not kill? Individual differences in working memory capacity, executive control, and moral judgment. *Psychological Science, 19*(6), 549–557.

Önkal, D., Goodwin, P., Thomson, M., Gönül, S., & Pollock, A. (2009). The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making, 22*(4), 390–409.

Peeters, G., & Czapinski, J. (1990). Positive-negative asymmetry in evaluations: The distinction between affective and informational negativity effects. *European Review of Social Psychology, 1*(1), 33–60.

Prahl, A., & Van Swol, L. (2017). Understanding algorithm aversion: When is advice from automation discounted? *Journal of Forecasting, 36*(6), 691–702.

Promberger, M., & Baron, J. (2006). Do patients trust computers? *Journal of Behavioral Decision Making, 19*(5), 455–468.

Rahwan, I., Cebrian, M., Obradovich, N., et al. (2019). Machine behaviour. *Nature, 568*, 477–486. https://doi.org/10.1038/s41586-019-1138-y

Rom, S. C., Weiss, A., & Conway, P. (2017a). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology, 69*, 44–58.

Rom, S. C., Weiss, A., & Conway, P. (2017b). Judging those who judge: Perceivers infer the roles of affect and cognition underpinning others' moral dilemma responses. *Journal of Experimental Social Psychology, 69*, 44–58.

Ross, W. D. (1930). *The right and the good.* Oxford University Press.

Sacco, D. F., Brown, M., Lustgraaf, C. J., & Hugenberg, K. (2017). The adaptive utility of deontology: Deontological moral decision-making fosters perceptions of trust and likeability. *Evolutionary Psychological Science, 3*, 125–132.

Singer, P. (1993). *Practical ethics* (2nd ed.). Cambridge: Cambridge University Press.

Sinnott-Armstrong, W., & Skorburg, J. A. (2021). How AI can AID bioethics. *Journal of Practical Ethics, 9*(1).

Thomson, J. J. (1976). Killing, letting die, and the trolley problem. *The Monist, 59*(2), 204–217.

Thomson, J. J. (1984). The trolley problem. *Yale LJ, 94*, 1395.

Trémolière, B., & Bonnefon, J. F. (2014). Efficient kill–save ratios ease up the cognitive demands on counterintuitive moral utilitarianism. *Personality and Social Psychology Bulletin, 40*(7), 923–930.

Turpin, M. H., Walker, A. C., Fugelsang, J. A., Sorokowski, P., Grossmann, I., & Białek, M. (2021). The search for predictable moral partners: Predictability and moral (character) preferences. *Journal of Experimental Social Psychology, 97*, 104196.

Uhlmann, E. L., Pizarro, D. A., & Diermeier, D. (2015). A person-centered approach to moral judgment. *Perspectives on Psychological Science, 10*(1), 72–81.

Uhlmann, E. L., Zhu, L. L., & Tannenbaum, D. (2013). When it takes a bad person to do the right thing. *Cognition, 126*(2), 326–334.

Walker, A. C., Turpin, M. H., Fugelsang, J. A., & Białek, M. (2021). Better the two devils you know, than the one you don't: Predictability influences moral judgments of immoral actors. *Journal of Experimental Social Psychology, 97*, 104220.

Williams, B. (1973). *A critique of utilitarianism. Smatr and Williams*. Utilitarianism: For and Against/Cambridge University Press.

Ybarra, O., Chan, E., Park, H., Burnstein, E., Monin, B., & Stanik, C. (2008). Life's recurring challenges and the fundamental dimensions: An integration and its implications for cultural differences and similarities. *European Journal of Social Psychology, 38*(7), 1083–1092.

Young, A. D., & Monroe, A. E. (2019). Autonomous morals: Inferences of mind predict acceptance of AI behavior in sacrificial moral dilemmas. *Journal of Experimental Social Psychology, 85*, Article 103870.

Yudkin, D. A., Goodwin, G., Reece, A. G., Gray, K., & Bhatia, S. (2023). A large-scale investigation reveals unexplored regions in the landscape of everyday moral experience. *PsyArXiv*. https://doi.org/10.31234/osf.io/5pcew