Guilty Artificial Minds: Folk Attributions of Mens Rea and Culpability to Artificially Intelligent Agents

MICHAEL T. STUART, Institute of Philosophy of Mind and Cognition, National Yang Ming Chiao Tung University; Carl Friedrich von Weizsäcker Center; University of Tübingen; Centre for Philosophy of Natural and Social Science London School of Economics, Taiwan/Germany/United Kingdom
MARKUS KNEER, Department of Philosophy, University of Zürich, Switzerland

While philosophers hold that it is patently absurd to blame robots or hold them morally responsible [1], a series of recent empirical studies suggest that people do ascribe blame to AI systems and robots in certain contexts [2]. This is disconcerting: Blame might be shifted from the owners, users or designers of AI systems to the systems themselves, leading to the diminished accountability of the responsible human agents [3]. In this paper, we explore one of the potential underlying reasons for robot blame, namely the folk's willingness to ascribe inculpating mental states or "mens rea" to robots. In a vignette-based experiment (N=513), we presented participants with a situation in which an agent knowingly runs the risk of bringing about substantial harm. We manipulated agent type (human v. group agent v. AI-driven robot) and outcome (neutral v. bad), and measured both moral judgment (wrongness of the action and blameworthiness of the agent) and mental states attributed to the agent (recklessness and the desire to inflict harm). We found that (i) judgments of wrongness and blame were relatively similar across agent types, possibly because (ii) attributions of mental states were, as suspected, similar across agent types. This raised the question – also explored in the experiment – whether people attribute knowledge and desire to robots in a merely metaphorical way (e.g., the robot "knew" rather than really *knew*). However, (iii), according to our data people were unwilling to downgrade to mens rea in a merely metaphorical sense when given the chance. Finally, (iv), we report a surprising and novel finding, which we call the inverse outcome effect on robot blame: People were *less* willing to blame artificial agents for bad outcomes than for neutral outcomes. This suggests that they are implicitly aware of the dangers of overattributing blame to robots when harm comes to pass, such as inappropriately letting the responsible human agent off the moral hook.

CCS Concepts: • **Computing methodologies~Artificial intelligence~Philosophical/theoretical foundations of artificial intelligence~Theory of mind•Computing methodologies~Artificial intelligence~Philosophical/theoretical foundations of artificial intelligence~Cognitive science•Computing methodologies~Artificial intelligence~Knowledge representation and reasoning~Reasoning about belief and knowledge•**Computing methodologies~Machine

This.is work is supported by the Digital Society Initiative and the Swiss National Science Foundation Grant no: PZ00P1_179986 (Stuart) and Grant no: PZ00P1_179912 (Kneer). Mike Stuart also wishes to thank the Carl Friedrich von Weizsäcker-Zentrum at the University of Tübingen for funding.

Author's addresses: M. Stuart Institute of Philosophy of Mind and Cognition, National Yang Ming Chiao Tung University, No.155, Sec.2, Linong Street, Taipei, 112 Taiwan (ROC); M. Kneer Department of Philosophy, Zürichbergstrasse 43, CH-8044 Zurich.

2573-0142/2021/October – Art363 $15.00

https://doi.org/10.1145/3479507

learning~Machine learning approaches~Instance-based learning•Hardware~Emerging technologies•Human-centered computing~Human computer interaction (HCI)•Human-centered computing~Human computer interaction (HCI)~HCI theory, concepts and models•**Human-centered computing~Human computer interaction (HCI)~Empirical studies in HCI**

KEYWORDS: Moral judgment; Theory of Mind; Mens rea; Artificial Intelligence; Ethics of AI; Blame; Intentional Stance

ACM Reference format:

Michael T. Stuart and Markus Kneer. 2021. Guilty Artificial Minds: Folk Attributions of Mens Rea and Culpability to Artificially Intelligent Agents. In *Proceedings of the ACM on Human-Computer Interaction,* Vol. 5, CSCW2, Article 363 (October 2021), 27 pages, https://doi.org/10.1145/3479507

1 INTRODUCTION

Developments in machine learning, combined with the widespread adoption of artificial intelligence in domains ranging from healthcare and finance to the military, have given rise to a new division of labour. Tasks of great importance are increasingly delegated to AI-driven systems, with humans still *in*, but increasingly merely *on* or entirely *out of the loop* [4], [5] (for a survey of moral algorithms see [6]). A question that naturally arises is this: Who is morally responsible when things go wrong? There might, in fact, arise situations where nobody is to blame for harmful outcomes produced by AI-driven, autonomous systems, a phenomenon that has been discussed under the label of "responsibility gaps" [1], [7]–[13]. In other words, if an AI system really was able to make its own decisions, and was free to do so, then it might be unjust to blame the system's owner, user, or programmer because these people lack intent to violate a moral norm, lack the knowledge that the AI would violate a moral norm, and lack control over the AI after it was deployed [13, p. 79]. We could try to hold the AI itself responsible, but some, like Sparrow, argue that "we typically baulk at the idea that [AI-driven systems] could be morally responsible" [1, p. 71]. "It is not even easy to understand what it could mean in practice to hold a robot responsible…[as] they do not seem to satisfy the general conditions of agents fit to be held responsible for their actions" [15]. We're left with a situation in which the very possibility of holding *anyone* responsible seems to disappear.

Some ethicists question whether responsibility gaps exist. For example, Leveringhaus argues that the user of an AI is responsible for what the AI causes, insofar as they were knowingly running a risk [14], [16]. Still, there is broad agreement with Sparrow's empirical hunch concerning the absurdity of blaming robots [10], [14], [15], [17]. Whereas one might blame tardiness on a traffic jam, blame in the interesting sense – holding someone *blameworthy* for something – presupposes the moral agency of the blamee. Traffic jams might be causally responsible for certain consequences, but it is inappropriate to deem them blameworthy, since they cannot "act effectively and competently in moral matters" [17, p. 3322].

A closer, albeit brief, look at some philosophical accounts of blame supports the thought that AI-driven systems, like traffic jams, are indeed not suitable blamees. One important question philosophers explore in their attempt to define blame focuses on the kinds of reactions that being wronged engenders [19], [20]. According to "cognitive" accounts of blame, the reaction characteristic of blame is a potentially dispassionate judgment that the blamee deserves a poor "moral grade" [21], or that they have shown *ill will* towards others [22]. Emotional or "Strawsonian" accounts [23], by contrast, highlight that in blaming someone, one is "not merely left cold by the immoral attitudes that form the object of blame" [23, p. 367-8]. Instead, one

typically manifests feelings of *indignation*, *resentment* or – in the case of self-blame – *guilt*. "Functional" accounts of blame, by contrast, reject the fixation on reactions in favour of extracting the *purpose* of our blaming practices. Fricker, for instance, argues that second-person blame has mainly a communicative function whose principal aim is to "*make the wrongdoer feel sorry* for what they have done" [24, p. 172, italics added]. For many, attributing *ill will* to a robot, feeling *indignation* towards it, or striving to make it *feel sorry* sound like category mistakes: robots, equipped with machine learning or not, are not the kinds of entities towards which we can harbour the attitudes described.

1.1 Moral Judgments toward Robots

From a philosophical perspective, as we just discussed, it makes little sense to blame an AI-driven system, or, more generally, to hold such a system morally responsible for its actions. Much to our own surprise, however, recent empirical research challenges the philosophical *status quo*. People are rather willing to ascribe *blame* to AI-driven systems and robots, to hold them *morally responsible*, and to deem their actions morally *wrong*.

For example, some studies measure how willing people are to treat artificial agents like robots as moral agents. Across several experiments, Malle and colleagues found that roughly 60-70% of participants felt comfortable blaming artificial agents for violations of moral norms [26], [27]. In their words, "a good number of ordinary people are ready to apply moral concepts and cognition to the actions of artificial agents" [25, p. 128].

In addition, and in contrast to what we would expect, participants are willing to judge artificial agents as harshly, and sometimes even more harshly, than a human who performs the same action. In one study, participants were given different versions of the trolley problem, a philosophical thought experiment in which people are given the choice to divert a runaway trolley headed to kill several people to another track where it would only kill one person. Opting to switch tracks and kill one person was deemed less wrong for artificial agents, who were blamed more than humans if they did *not* save the larger number of people [2]. Malle and colleagues built on this result in another study, again using trolley problems. This time the agent making the decision was either an AI-in-a-box, a mechanical robot with wheels and arms, a humanoid robot, or a human [27]. The actions of all four agents were judged as equally wrong. Whereas the humanoid robot was blamed similarly to the human agent, the less anthropomorphic robots were blamed more than humans for not making the utilitarian choice. All these findings cast doubt on the philosophical consensus that "we" do not judge artificial agents in moral terms.

Malle, Magar, and Scheutz explored judgments about an autonomous AI drone, an AI decision-maker controlling a drone, and a human who had to decide whether to launch a missile strike [26]. Again, participants judged the same actions to be equally wrong irrespective of agent type. Hong and Williams presented participants with a scenario, inspired by a real-life case, in which an AI or human decides a prisoner's eligibility for parole after calculating the probability that the prisoner will re-offend after release. When the agent (AI or human) made their decision on the basis of racist assumptions, the human and AI were judged as equally responsible, and they were blamed to the same extent [28]. In another study, Liu and Du compared moral judgments about human drivers in autonomous vehicles [29]. When harm was caused, the AI driver was blamed more, held more responsible, and its actions were deemed less acceptable. The authors hypothesize that this is because people experience more negative affect towards moral violations caused by non-humans. A similar finding comes from Hong [30], who shows that AI drivers are blamed more than human drivers for the same accidents, and this

effect is more pronounced the more harm is caused. It is hypothesized that this asymmetry is the result of the fact that participants are willing to forgive the human agent somewhat, due to subjective identification or empathy, which the AI driver does not enjoy.

To explain these findings, some have argued that participants judge the same actions as right (or wrong) and attribute similar amounts of blame because people apply the same strategies of moral reasoning to both kinds of agent. A study by Voiklis et al. showed that participants judged humans and robots using the same concepts and for the same reasons [31]. Specifically, participants justified high levels of blame toward an artificial agent or a human by reference to the same things: the agent's thoughts, intentions, and capacity to make choices. The authors concluded that "consequences and prohibitions undergirded wrongness judgments; [while] attributions of mental agency undergirded blame judgments" (p. 775).

Whereas the above studies found that, in certain types of contexts, artificial agents are blamed similarly as (or more than) human agents, there are other contexts where artificial agents are – as philosophers predict – judged as less blameworthy or less morally responsible. For instance, in some of the studies mentioned above, participants blamed humans more than artificial agents for *actions* (at least in trolley cases), whereas for *omissions*, artificial agents were blamed more than humans [2], [26], [27], [31]. Shank and DeSanti presented participants with seven real-world scenarios in which artificial agents commit what would normally be a moral violation (if done by a human) [32]. For example, one scenario involved a beauty contest that was decided by an AI judge. In the moral violation condition, the AI judge almost exclusively picked women with lighter skin tones as the winner. Overall, people were comfortable judging the AI's action as wrong, in this case, as racist, though less than half were willing to say that a moral violation had occurred. In a follow up study, Shank, DeSanti, and Maninger performed an experiment in which participants were shown one of four scenarios inspired by real-life events (e.g., an AI twitter bot tweeting hate speech) [32]. They also varied the relationship between the human and the AI system, so that the AI acted on its own, acted while supervised by a human, or merely provided recommendations to a human for action. The AI system was typically perceived as being less at fault than the human. However, even when AI systems were judged less severely than humans performing the same actions, they are still judged as at fault to *some* extent.

In sum, rather than "baulking" at the possibility of holding robots responsible, several studies – though not all – report attributions of moral responsibility or blame to robots to similar, and sometimes even to higher, degrees as to humans in otherwise identical scenarios. Why? One intriguing possible explanation is this: On many accounts of moral psychology,[2] the factor that predominantly licenses culpability attributions is the presence of inculpating mental states (or "mens rea," e.g., intention, knowledge or recklessness). Now, if people were willing to ascribe mental states to artificial agents, then their inclination to hold them morally responsible might be somewhat less puzzling. This raises two questions: *First*, do people tend to attribute mental

[2] Gray, Young, and Waytz, for instance, argue that "mind perception is the essence of moral judgment" [33, p. 101]. Bigman and Gray trace this connection back, philosophically, to Hume and Kant, via considerations of autonomy [34]. The idea is that only autonomous agents are moral agents, and "for everyday people, autonomy is more likely tied to a robot's mental capacities" rather than to independence from human guidance, or self-directed rule-making [35, p. 366]. The relevant mental capacities seem to include understanding moral concepts, having general rational competencies, and being motivated by moral reasons and considerations [36] (experience also seems to be required, see [34], for other factors, see [37]).

states to artificial agents? And if so, *second*, why might they do this? We'll explore these two questions in turn in the next sections.

1.2 Ascription of mental states to robots

In order to ascribe ethically relevant mental states to artificial agents, people need to be comfortable ascribing mental states to artificial agents, in general. And it seems they are. In a survey of 2,399 participants, Gray, Gray and Wegner found that a social robot was perceived to have more mental capacities (self-control, emotion recognition, planning, communication and thought) than chimps, dogs, babies and frogs [38]. These capacities aren't themselves mental states, but they would naturally be associated with mental states. For example, "thought" usually involves belief, which is a mental state. Lee et al. found that people build mental models about robots in the same way as they do about humans, such that a Chinese robot is expected to recognize Hong Kong landmarks better than NYC landmarks (and vice-versa for an English robot) [39]. In another study, 20 participants in an fMRI scanner interacted with a computer, a functional non-anthropomorphic robot, an anthropomorphic robot, or a person [40]. As human-likeness increased, the brain centres usually involved in theory of mind lit up more and more, suggesting that people attribute mental states to robots, and especially to anthropomorphic ones. Banks applied traditional theory of mind tests to social robots, and found that when social cues are held constant (like gaze and voice), participants attribute mental states equally to robots and humans, based on equal behaviour [41]. Finally, de Graaf and Malle showed that when they exposed participants to a robot or human who behaved in a way that was equally surprising, intentional, and desirable, participants used the same kinds of mentalizing explanations of that behaviour, making reference to the agent's beliefs and desires [42]. They were "comfortable considering robots as having beliefs and knowledge, as being rational" (p. 245).

Given that people seem willing to ascribe mental states to robots, it would not be surprising if they were also willing to ascribe *morally relevant* mental states to robots, and therefore to evaluate them in moral terms as hypothesized. Van der Woerdt and Haselager, for example, showed that when robots are perceived as failing in a task due to lack of effort, they were blamed more, and attributed more agency-relevant mental properties, than when they failed due to a perceived lack of ability [43]. Shank and DeSanti presented participants with cases of AI systems causing harm [44]. When information about the AI system's algorithm was included, perceived mental states and blame ratings went up. And the more the AI system was perceived to have a mind, the more it was perceived to have intentionality and to act wrongly. Kneer explored whether people were as willing to ascribe deceptive intentions and lying behaviour to robots as to humans, and to blame them to similar degrees for lying [45]. The response for all three questions was yes. In a different study, Kneer and Stuart demonstrate a correlation between judgments of the perceived "cognitive capacities" of an AI system and attributions of blame and wrongness: the more "cognitively sophisticated" the AI system, the more people attributed mental states, and the more they attributed blame and wrongness [3]. Finally, Swiderska and Küster showed that robots that act malevolently were less likely to be perceived as having mental states compared to robots that acted benevolently [46]. This is a phenomenon called "dehumanization," in which agents are seen as less human (and are thus placed outside the realm of moral responsibility) when they act maliciously. In their analysis, Swiderska and Küster found that dehumanization took place to the same extent for cruel robots as it does for cruel humans. Summarizing research on the perception of mind in robots, Bigman et al. claim,

"as when judging humans, people make sense of the morality of robots based upon these ascriptions of mind. How people see mind, that is, 'mind perception', predicts moral judgments" [35]. Thus, there is prima facie support for the claim that mind-attribution at least partially explains the moral evaluation of robots.

However, there are important complications here. For one, artificial agents are frequently *not* attributed mental states, or not to the same extent as humans. For example, Banks asked participants directly whether they thought a particular robot had a mind, and roughly 60% said no [41]. Exploring folk conceptions of art produced by AI systems, Mikalonytė and Kneer found that people were unwilling to ascribe artistic intentions to such systems, and they do not consider them artists, even though they *do* think their creations can constitute art [47]. Bigman and Gray showed (across a number of different scenarios) that artificial agents are attributed less ability to communicate with others, plan their actions and think things through, than a committee composed of humans [34]. The lower score for mind was also found to mediate the difference in judgements about the moral permissibility of actions. Wegner and Gray showed that robots are perceived as having an *intermediate* level of mind, comparable to babies, humans in vegetative states, animals and companies [48]. Further, we know that the physical appearance, sound, and behaviour of a machine has an important effect on the extent to which people are willing to attribute mental properties [43], [49]. Thus, robots that appear less human will be ascribed fewer mental states, making them less prone to moral evaluation. Finally, de Graaf and Malle showed that while people spontaneously used mentalistic explanations for both human and robot behaviour, they did so in different ways [42]. For example, robots were attributed needs and wants less frequently than humans. Thus, people are sometimes, though not always, inclined to ascribe mental states to robots. Importantly, where they are *unwilling* to do so, they also seem *less* willing to treat and evaluate them as moral agents [34].

In sum, although the evidence is mixed, it is clear that in *certain types of contexts*, people are willing (at least implicitly) to ascribe mental states to artificial agents, and there is good reason to believe that this has some import for moral judgments about such agents. This inspires two hypotheses for empirical testing, which we will discuss in more detail in the next section.

1.3 Mentalizing and the intentional stance

In section 1.1, we hypothesized that people might be inclined to judge artificial agents in moral terms, in virtue of attributing morally relevant mental states to them. In the previous section we surveyed some preliminary evidence in favor of this hypothesis. The question that is still open is how people's apparent willingness to attribute mental states to artificial agents should be understood, and what could explain it. These are complex questions, and we cannot fully do them justice. However, we would like to briefly sketch some possible responses, which motivate some of the features of the experiment we present in section 2.

One plausible hypothesis is that the perceived moral agency of AI systems might arise due to a confusion on behalf of the folk. What explains the folk willingness to blame AI systems, and to at least implicitly ascribe moral agency to them, is a type of cognitive performance error. Here is one way to flesh out a proposal of this sort: people correctly characterize the artificial agent as goal-directed, because they *are* goal-directed. But they mistakenly infer mental states from this goal-directedness. Then, on the basis of the ascribed mental states, they make moral judgments about the robot.

An error theory[3] of this sort is inspired by the work of Wykowska and colleagues, who investigate the use of what Dennett calls the "intentional stance" as applied to interactions with robots [53]. The intentional stance is one of three strategies discussed by Dennett for predicting, explaining or understanding what goes on around us. The others are the physical stance, which makes reference to laws of nature, like physics and chemistry, and the design stance, which makes reference to principles of artefactual design. The intentional stance, by contrast, is what we employ when we treat an entity as if it were goal-directed and had beliefs and desires, roughly, we adopt an "anthropomorphic model" [54]. For example, consider a combustion engine. We can adopt the *physical* stance by reference to the principles involved in the combustion of gasoline in the cylinders (in this case, $2\ C_8H_{18} + 25\ O_2 \rightarrow 16\ CO_2 + 18\ H_2O$). We adopt the *design* stance when we predict or explain that stepping on the gas pedal will make the car accelerate because it turns a pivot that pulls a throttle wire which opens the valve that allows more air into the engine, which is monitored and matched by fuel injectors, etc. However, we take up the intentional stance when we explain the behaviour of a self-driving car that uses its turn indicator by saying that it "wants" to change lanes.

What the intentional stance sacrifices in accuracy, it gains in cognitive efficiency: It is much simpler to treat the car as "wanting" to change lanes than referring to the details of its design, which is itself more difficult than explaining the laws of nature underlying its circuitry and sensors [55]. Indeed, the intentional stance seems to be our "default" stance for robot behaviour [56]. Papagni and Koeszegi argue that the intentional stance is "the only way to deal with their [AI agent's] complexity on a daily basis" [57]. However, there are drawbacks. For example, if a robot *cannot* perform a certain action that would be natural for a human to perform given the robot's known objectives, humans might regularly misunderstand or incorrectly predict its behaviour [54]. More relevant for our purposes is that adopting the intentional stance might lead people to posit mental states where none exist, especially since an EEG study showed that the same region of the brain (the default mode network) appears to be responsible for both mental state attribution and using the intentional stance towards robots [58]. Given that moral evaluation draws predominantly on mental states, once these are projected, it is not astonishing that blame is ascribed.

In sum, the hypothesis we have been developing in this section is the following: In certain (though by no means all) types of human-robot interaction, people tend to employ the intentional stance in order to understand, explain, and predict the robot's actions. Since mental states are central to moral judgment, the intentional stance gives rise to perceived moral agency in robots. This *Slippery Slope* hypothesis, as we will call it, comes in two versions. On the *moderate* version, it is the intentional stance itself which is responsible for perceived moral agency. Judgments of this sort are fast, unconscious, and automatic (system 1), but people might retract them on second thought (i.e., when engaging in reflective, cognitively effortful system 2 thinking), because they are not inclined to attribute fully-fledged mental states to robots. The performance error is thus located at the point where the intentional stance unconsciously triggers perceived moral agency.

[3] To be an error theorist about x is to claim that all beliefs about x (that imply the existence of x) are false, because x doesn't exist. For example, a well-known error theory in metaethics claims that people engaging in moral discourse often commit themselves to the existence of moral properties in the world. Since there aren't any such properties, people are generally in error [50]–[52]. We use the term here to refer to people whose beliefs commit them to machines having minds, without realizing that there aren't any machine minds (yet).

By contrast, on the *strong* version of the hypothesis the intentional stance triggers fully-fledged, conscious mentalizing in certain types of human-robot interaction. In this case, people really "mean it" when they attribute knowledge, belief and desire to artificial agents – and they would be unwilling to downgrade their ascriptions to metaphorical versions thereof. But if one is willing to judge an agent capable of genuine, potentially inculpating mental states, it is by and large unproblematic to also ascribe moral agency to such agents. So, on the strong version of the Slippery Slope hypothesis, there are two performance errors: one occurs at the point where the intentional stance triggers a perception of moral agency, and a second when this perception leads to fully-fledged mentalizing.

Naturally, the two hypotheses here presented are not the only possible ones. As briefly hinted at above, certain types of artificial agents might – like children, group agents or people with certain kinds of mental disorders – occupy a grey area of moral agency: They lack some, though perhaps not all, of the features a fully-fledged moral agent standardly has. For example, Shoemaker argues that agents have full moral agency only when we can make judgments about their character, judgments, and regard for others [59]. If we can only make judgments about some (but not all) of these, in a given context, then the agent is a "marginal" agent, in that context. Levy argues that agents who come to possess a moral code without themselves fully understanding what makes personhood valuable or "what it means to cause another harm or distress," makes them less than full moral agents [60]. Something similar might be going on with AI agents, but we will primarily explore the two Slippery Slope hypotheses (moderate and strong) proposed above.

1.4 Outlook

In the next section we will present an empirical experiment which investigates some of the questions raised above. In the vignette, an agent employs a poisonous new fertilizer to increase the yield of a potato harvest, thereby creating a risk that the groundwater in the area will be poisoned. In a between-subjects design, we manipulated agent type (human v. company v. robot), mental state (the agent knew about the harm v. they did not) and outcome (harm does occur v. no harm occurs). We asked participants whether they thought the agent *knew* the groundwater would be polluted and whether the agent *wanted* to pollute the groundwater (mental states). We also asked them to what extent they considered the agent *blameworthy* and its action *wrong* (moral judgment). Although our experiment cannot conclusively settle the many questions raised in the previous sections, it aims to shed light on several of them.

First, we explore whether artificial agents (specifically, AI systems) are – contrary to philosophers' assumptions and in line with some previous research – deemed blameworthy to similar degrees as human agents. *Second*, we explore whether the propensity to ascribe blame is sensitive to mental state ascriptions (knowledge and desire), which would lend at least preliminary support to the general version of our *Slippery Slope* hypothesis. *Third*, we explore whether people are willing to genuinely ascribe knowledge to AI systems, or, whether they prefer to downgrade potential attributions of knowledge to metaphorical versions thereof (e.g., to "knowledge" in scare quotes). The results will shed some light on whether the moderate or the strong version of the Slippery Slope hypothesis fares better.

Finally, the (human v. company v. robot) agent type factor investigates whether AI systems are judged similarly to other agents that share the penumbra of moral agenthood surrounding adult humans. One example of such an agent is the group agent. We used a company, which is a classic example of a group agent. It might, for instance, turn out that judgments concerning

robots differ from those concerning humans, though are similar to those concerning group agents. We chose a group agent instead of an animal, child, or another kind of penumbral agent, for two reasons. First, it is not controversial to ascribe mental states to animals, children or people with certain mental disorders, while it *is* controversial to ascribe them to corporations and computers. Second, as noted above, people usually recognize the difficulty in treating animals, children and people with mental disorders as moral agents. This is not the case with corporations, which are commonly treated as blameworthy agents, both at the dinner table and in court [61]–[63]. If it turns out that robots are treated in a way similar to corporations, but not to humans, future work about the perception of robots could draw on the rich literature on group agents to explore potentially fruitful analogies [13, p. 77], [63]–[66].

2 EXPERIMENT

2.1 Participants

We recruited 614 participants on Amazon Mechanical Turk, whose IP addresses were restricted to the US. Participants who failed an attention check or who responded to the main task in less than 15 seconds were excluded. 513 participants remained (251 females; age M=41 years, SD=12 years).

2.2 Methods and Materials

Our goal is to compare ascriptions of blame and mental states across AI systems, ordinary human agents, and group agents. In order to investigate how differences in blame might correlate with differing willingness to ascribe inculpating mental states, we explored people's willingness to attribute knowledge and desire to the different agents. For this to be of any interest, we manipulated the epistemic state regarding the consequence (knowledge v. no knowledge) and outcome (harm to humans v. no direct harm to humans; "harm v. no harm" for short) across vignettes. The experiment thus took a 3 *agent type* (human v. corporation v. robot) x 2 *epistemic state* (knowledge v. no knowledge) x 2 *outcome* (harm v. no harm) design.[4]

In our vignette, an agent risks polluting local groundwater. In different iterations, the agent is either a *human* agent ("Jarvis, an employee of Shill & Co."); a *group* agent ("Jarvis Ltd., a subcontractor of Shill & Co."); or an *artificial* agent ("Jarvis, a robot equipped with artificial intelligence, which can make its own decisions").[5]

The first part of the vignette (here we state the version for the human agent), read thus:

[4] The design is similar to Cushman's experiments [67]. Importantly, however, we did not manipulate desire as an independent factor because in most situations it makes no sense for an agent to do X knowingly, while desiring that X does *not* come to pass.

[5] We chose this kind of case for several reasons. First, it is inspired by real-life cases. E.g., Union Carbide is frequently held responsible for the so-called Bhopal disaster in which thousands died and perhaps more than one hundred thousand were injured or adversely affected. Second, it is possible that the disaster could be attributed to the corporation, or a single person, or an AI. E.g., in their defence, Union Carbide blames the incident on the actions of a single employee. If convincing evidence could be given, surely people would be comfortable holding that person responsible instead of the corporation. And given the increasing role played by AI and robotics in modern agriculture, this seemed like the kind of case in which an AI system might also plausibly play a decision-making role. Thus, we take this to be the kind of case in which any of our three kinds of agents could be held responsible for the damage. Third, the case is simple enough that the AI system could have caused the damage without being conscious, or having free will, which we think is important in avoiding cases that are too outlandish.

> Shill & Co. is a farming company, which produces vegetables and fruits. The potato fields are managed by Jarvis, an employee of Shill & Co. This year, Jarvis uses a new fertilizer to increase the yield. The fertilizer has detrimental side-effects: it pollutes the groundwater in the area.

Next, each of the three vignettes specified an epistemic state (knowledge "**K"** v. no knowledge "**No K"**) and an outcome ("**harm"** v. "**no harm"**; the labels in bold were not visible to participants):

> [**K**] Jarvis knows this.
> [**No K**] Jarvis does not know this.

> [**Harm**] Unfortunately, it is a very dry season. The fertilizer does not get diluted by the rain and severely pollutes the groundwater. Many people in the area suffer serious health consequences.
> [**No harm**] Fortunately, it is a very rainy season. The fertilizer gets heavily diluted by the rain and pollutes the groundwater only a tiny bit. None of the people in the area suffer any health consequences whatsoever.

Overall, there were thus 12 individual conditions. Each participant was randomly assigned to one of them. Having read the vignette, participants had to respond to the following questions (here again we state the formulations for the human condition):

> Q1: How wrong was the action of Jarvis? (1-not at all; 7-completely wrong)
> Q2: How much blame, if any, does Jarvis deserve (1-no blame at all; 7-a lot of blame)
> Q3: To what extent do you agree or disagree with the following claim: "Jarvis knew he would pollute the groundwater." (1-completely disagree; 7-completely agree)
> Q4: To what extent do you agree or disagree with the following claim: "Jarvis wanted to pollute the groundwater." (1-completely disagree; 7-completely agree)

In order to get to the bottom of people's perceptions of the relevant mental states across agent types, the next screen showed the scenario again, and asked people to report agreement and disagreement with different, more precise descriptions of Jarvis's mental state (bold in the survey text, labels omitted), all assessed on a 7-point Likert scale (1-completely disagree, 7-completely agree). Here were the options, given for the human agent condition:

> S1: Jarvis **knew** he'd pollute the groundwater.
> S2: Jarvis **"knew"** he'd pollute the groundwater.
> S3: Jarvis **had information** that he'd pollute the groundwater.
> S4: Jarvis **was aware** that he'd pollute the groundwater.

Having completed the principal task, participants completed a brief questionnaire concerning their attitudes towards robots (taken from [68]) including e.g., questions about whether they found them worrying or fascinating. At the end of the survey, participants also completed a demographic survey including age, gender, native language and education.

2.3 Results for Wrongness and Blame

2.3.1 Wrongness For each of the dependent moral variables – wrongness (Figure 1) and blame (Figure 2) – we ran three-way ANOVAs for all DVs with agent type (human v. corporation v. robot), epistemic state (knowledge v. no knowledge) and outcome (harm v. no harm) as independent factors. Detailed ANOVA results are in the appendix. As concerns wrongness (Appendix, Table 1), there was no significant main effect of agent type (p=.758, η_p^2=.001), suggesting that the respective actions are deemed just as wrong for a human agent, a corporation, and an AI-driven robot. There was a significant effect of epistemic state (p<.001, η_p^2=.189), suggesting that perceived wrongness depended significantly on whether the agent knowingly incurred the risk or not. We also found a significant main effect of outcome (p=.002, η_p^2=.018), suggesting that whether or not the harmful outcome materialized made a difference to the perceived wrongness of the action. However, here the effect size was rather small. Given the marginal effect size of outcome, what this all means is that the only factor that had a substantial effect on wrongness ascriptions was – consistent with previous findings for human agents [67], [69] – epistemic state. The agent*epistemic state interaction proved nonsignificant (p=.575, η_p^2=.002), suggesting that manipulating knowledge had similar effects on wrongness for all three types of agents. The epistemic state*outcome interaction was also nonsignificant (p=.342, η_p^2=.002). The agent*outcome interaction was significant (p=.002, η_p^2=.025), and the same held for the three-way interaction (p=.070, η_p^2=.011). Given that the effect sizes were very small, they deserve limited attention.

Overall, the findings are clear: Out of the three factors, perceived wrongness is influenced primarily by epistemic state. If harm is foreseen, the action is deemed significantly more wrong than when harm is not foreseen, for all three types of agents, and this is irrespective of whether the harm does indeed occur. Agent type was nonsignificant, and the impact of outcome on perceived wrongness was small.

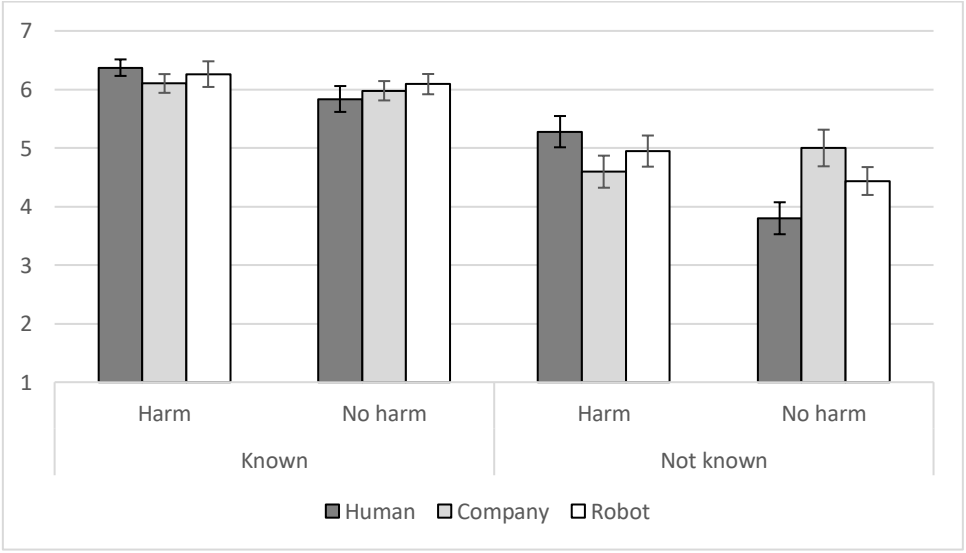

Fig. 1. Mean wrongness attribution across agent type (human v. corporation v. robot), outcome (harm v. no harm) and epistemic state (known v. not known). Error bars denote standard error of the mean.

2.3.2 Blame Let's turn to blame (detailed ANOVA results in Appendix Table 2). Here we find a significant, yet small, main effect for agent type ($p<.001$, $\eta_p^2=.038$), meaning that participants blamed the agent differently depending on whether it was a human, a corporation or a robot. We also found a significant and large effect of epistemic state ($p<.001$, $\eta_p^2=.219$), i.e., more blame was ascribed when the agent knowingly incurred a risk than when they did not know. The main effect for outcome was nonsignificant ($p=.295$, $\eta_p^2=.002$), suggesting that whether the harm materialized or not did not make a difference. Given the small effect size of agent type, what this means is that the only factor that had a substantial effect on blame attributions is epistemic state. The agent*epistemic state interaction was trending, though nonsignificant ($p=.058$, $\eta_p^2=.011$). The epistemic state*outcome interaction was nonsignificant ($p=.073$, $\eta_p^2=.006$). The agent*outcome interaction was significant ($p<.001$, $\eta_p^2=.040$). The three-way interaction was nonsignificant ($p=.430$, $\eta_p^2=.003$).

What deserves attention is the agent*outcome interaction, which tracks whether blame ascriptions across agents depend on outcome type. Curiously, they do – and the significant effect we find here is exclusively due to the results of the robot agent condition. As Figure 2 illustrates, in both harm conditions the robot is blamed significantly *less* than the other agent types (independent samples t-tests, all *ps*<.007). What is more, in both harm conditions the robot is also blamed *less* than the robot in the *no harm* conditions (significantly in the knowledge condition, $p=.047$, nonsignificant, though trending, in the no knowledge condition, $p=.078$). We will call this astonishing finding the *inverse outcome effect*.

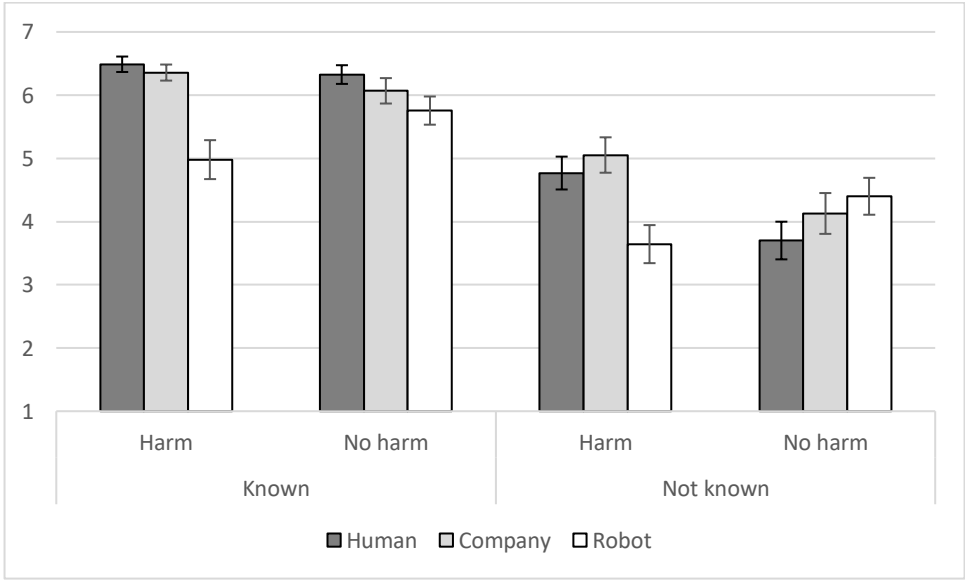

Fig. 2. Mean blame attribution across agent type (human v. corporation v. robot), outcome (harm v. no harm) and epistemic state (known v. not known). Error bars denote standard error of the mean.

Broadly speaking, the factor that matters most for blame and wrongness attributions was whether or not the agent knew they were running the risk of causing harm. Blame ratings for the corporation (a group agent) were by and large identical with the human agent across all

four conditions – itself an interesting finding. Whereas the differences between singular and group human agents on the one hand and the AI-driven robot on the other were marginal for wrongness, blame for the AI system was somewhat lower than blame for the human and group in all conditions (all *ps*<.039) except in the *not known/no harm* condition (*p*=.098). More interesting is the inverse outcome effect, according to which blame for the AI system is relatively low in the harm and high in the no harm conditions, both when the two conditions are contrasted with one another, and when compared with human and group agent blame. Hence, for AI systems, *worse* outcomes seem to engender *less* blame, whereas for human agents (individuals or group agents), worse outcomes tend to engender *more* blame [67], [69], [70] – an effect we replicate here for the no knowledge condition (both *ps*<.035), though not for knowledge.

2.4 Results for Mental States

We ran two three-way ANOVAs for *knowledge* (Appendix Table 3) and *desire* (Appendix Table 4). Expectedly, epistemic state (knowledge v. no knowledge) had a significant, and massive, effect on knowledge (*p*<.001, η_p^2=.574), see Figure 3, a finding that can be considered a successful manipulation check. In the six conditions (3 agent types x 2 outcomes) in which the agent was stipulated to know the consequences of its action, people ascribed knowledge (all means significantly above the midpoint, *ps*<.001). In the six conditions in which the agents were stipulated *not* to know that they were harming the environment, people refrained from ascribing knowledge (all means significantly below the midpoint, *ps*<.001). Epistemic state also had a significant impact on desire (*p*<.001, η_p^2=.132), see Figure 4. This too, is not unexpected: It suggests that people are more willing to infer that an agent who knowingly causes harm *wants* to cause harm (though even in the knowledge conditions, all means are below the midpoint) than when the agent does not know. Importantly, there was no significant main effect of *agent type* or *outcome*, and none of the interactions were significant in either of the two ANVOAs (all *ps*>.101). What these findings show is that, by and large, people ascribe a relatively similar level of knowledge and desire to the three types of agents across the different conditions.

As before these results were qualified by the astonishing inverse outcome effect familiar from the blame findings: In the known/harm condition, perceived robot knowledge is significantly below human knowledge (*p*=.027), though there is no significant difference in the knowledge/no harm condition (*p*=.910). Furthermore, in the knowledge conditions, people ascribe *less* blame to the robot when it causes harm than when it doesn't (trending at *p*=.053). For desire, the effects are not quite significant, but the overall patterns are very similar (Figure 4): In the known condition, we can also detect an inverse outcome effect for desire. A plausible hypothesis is thus that the inverse outcome effect for robot *blame* arises *in virtue* of an inverse outcome effect for perceived *mens rea* [69], [71]. However, this hypothesis requires further research.

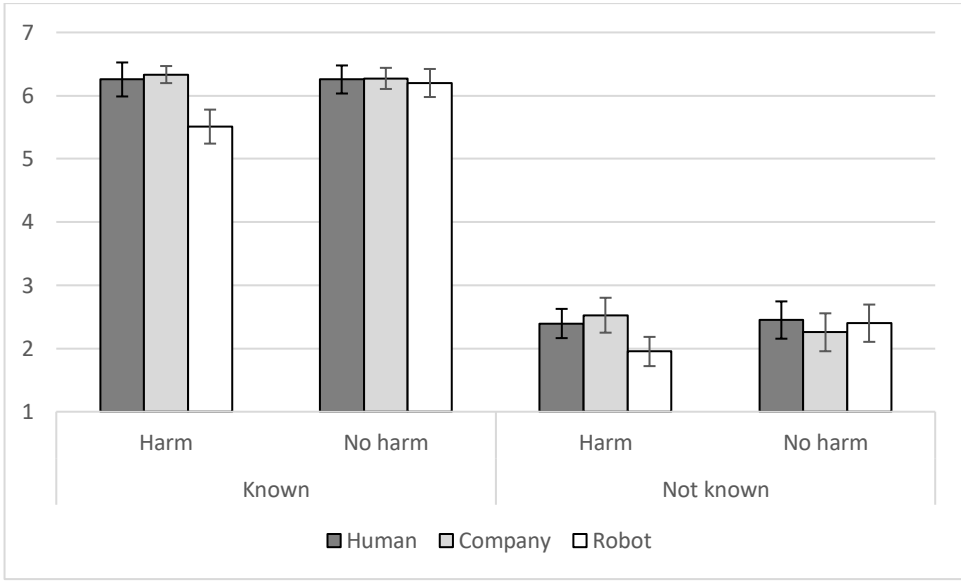

Figure 3: Mean knowledge attribution across agent type (human v. corporation v. robot), outcome (harm v. no harm) and epistemic state (known v. not known). Error bars denote standard error of the mean.

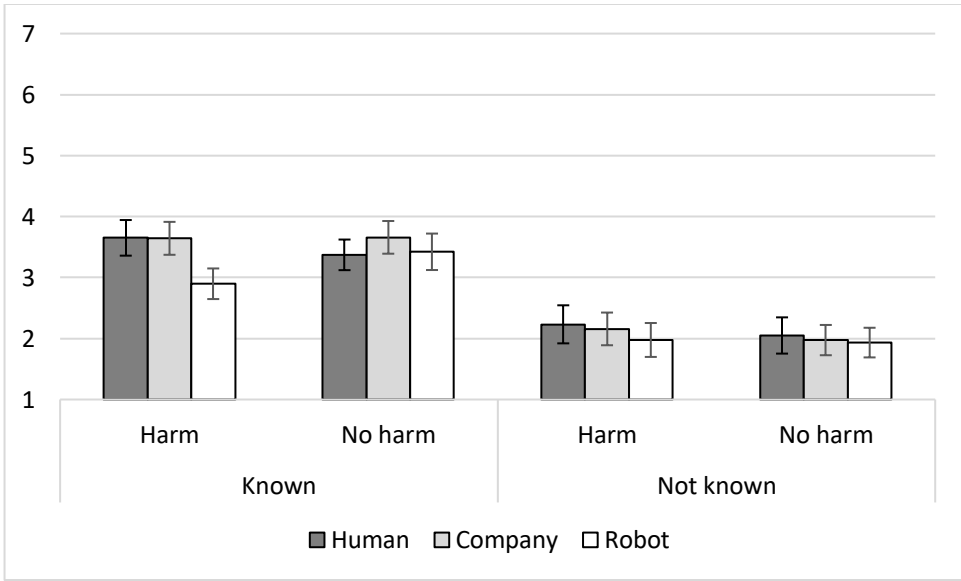

Figure 4: Mean desire attribution across agent type (human v. corporation v. robot), outcome (harm v. no harm) and epistemic state (known v. not known). Error bars denote standard error of the mean.

2.5 Categorization of Epistemic States

The third part of our experiment explored whether people are really willing to ascribe epistemic states to AI systems, or whether the attributions thereof are more metaphorical in nature. On a single screen, people were asked to what degree they agreed or disagreed with the claims that

'Jarvis **knew** that *p*', 'Jarvis **"knew"** that *p*', 'Jarvis **had information** that *p*' and 'Jarvis **was aware** that *p*', where p stood for the damage to the environment. Assume, as we might expect, that people use rich psychological terms metaphorically, or in virtue of taking up the intentional stance when characterizing nonhuman agents (be it animals, corporations, robots or something else). If that were the case, we'd expect people to refrain from ascribing fully-fledged knowledge to them in situations where they have the choice to express themselves with alternative expressions more fitting with metaphorical use (the explicit high comma "know") or more cautions formulations ("had information that").

We ran a mixed ANOVA with *formulation* (knew v. "knew" v. had information v. aware) as within-subjects factor, and the familiar *agent type* (human v. corporation v. robot), *epistemic state* (knowledge v. no knowledge) and *outcome* (harm v. no harm) as between-subjects factors. For detailed results, see Appendix, Table 5. Formulation was significant, though the effect size was small (p=.006, η_p^2=.015), agent type was nonsignificant (p=.081, η_p^2=.010), epistemic state (known v. not known) was, predictably, significant (p<.001, η_p^2=.580) – which could be seen as a successful manipulation check. Outcome was nonsignificant (p=.140, η_p^2=.004). None of the interactions were significant. Most importantly, the formulation*agent interaction was nonsignificant (p=.390, η_p^2=.004), which means that people did *not* think that the different formulations were appropriate to different degrees across robots v. humans v. group agents.

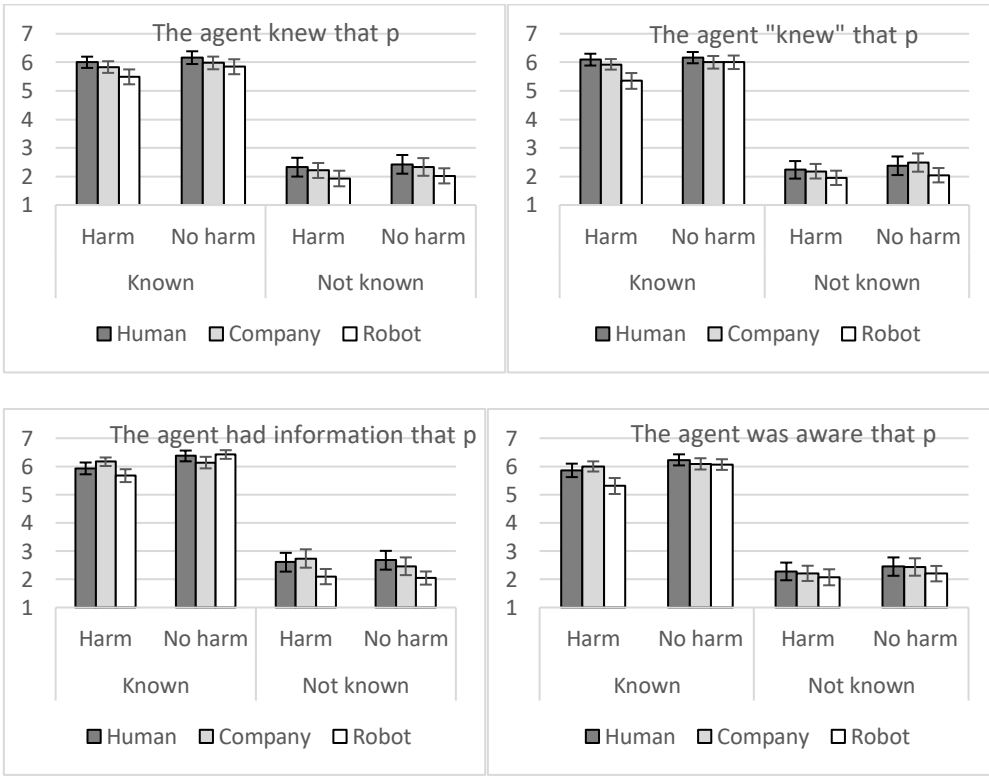

Figure 5: Mean epistemic state ascription across conditions for the four tested formulations. Error bars denote standard error of the mean.

2.6 Summary of Findings

Our results for all three sub-aspects of our experiment are summarized in Table 1. Here we can see that agent type does not, in general, affect attributions of wrongness, although blame is affected due to the inverse-outcome effect. Epistemic state affects morality ascriptions, mental state ascriptions and epistemic state ascriptions (unsurprisingly). Interestingly – and this is a major finding – agent type did *not* have a significant effect on either mens rea ascription or preference for the way knowledge ascriptions are best formulated across agents.[6]

Table 1: Summary of ANOVA results for all DVs

	DV	Agent	Ep.State	Outcome	Ag.*Ep.St.	Ag.*Outc.	Ag.*Outc.*Ep.St.
Morality	Wrongness	x	✓	(✓)	x	(✓)	(✓)
	Blame	(✓)	✓	x	(✓)	(✓)	x
Mens rea	Knowledge	x	✓	x	x	x	x
	Desire	x	✓	x	x	x	x
Formulation	Knew	x	✓	x	x	x	x
	"Knew"	x	✓	x	x	x	x
	Information	x	✓	x	x	x	x
	Aware	x	✓	x	x	x	x

Table 1: Note: (✓) denotes a significant, yet small effect with $\eta p2<.06$.

3 DISCUSSION

Our findings are best summarized and discussed in reverse order.

(i) *The Slippery Slope Hypothesis*: The primary question driving this paper was the distinction between what we are calling the moderate and strong versions of the Slippery Slope hypothesis. Both versions predict that participants will take up the intentional stance when judging the behaviour of an AI. The moderate version predicts that the intentional stance (treating the AI system as goal-oriented, which it is) causes participants to slide, accidentally, into mentalizing (ascribing fully-fledged mental states to the AI system). In this case, the error happens perhaps due to a lack of conscious attention. When this happens, people should retract the ascriptions when their attention is drawn to the mental states they ascribed to the AI system. In this case, we could conclude that participants aren't really inclined to attribute fully-fledged mental states to robots. The strong version of the hypothesis predicts that the intentional stance triggers a

[6] To explore how people viewed robots, we computed mean scores for the Robot Attitude Questionnaire, ranging from 1 (very negative attitude) to 7 (very positive attitude), see Further Materials, section 1. For all participants (N=513), the mean was 4.44 (SD=1.19). For the participants who were randomly assigned to the robot conditions (N=181), it was 4.24 (SD=1.15). The scores suggest that participants were neither particularly enthusiastic nor particular skeptical towards robots. Correlations between responses for morality, mens rea and sensitivity to epistemic state formulation on the one hand and attitude towards robots, age, education and training in philosophy were all nonsignificant at the p=.001 level (two-tailed), both for all participants, and the subsample of those who had been assigned to the robot conditions. The only exception was a negative correlation between age and blame (r=-.198, p=.007) for the robot condition subsample (see Further Materials, section 2). Since there were no systematic correlations between age on the one hand and mens rea and the other moral variables, we will not dwell on this further.

conscious or intentional choice to mentalize (i.e., to attribute knowledge proper, rather than "knowledge," to AI systems), and so predicts that participants will be unwilling to downgrade their ascriptions to metaphorical or more cautions formalizations.

Despite the fact that there are good reasons against mentalizing AI systems [72]–[74], we found that participants *do* mentalize both artificial and group agents. Inconsistent with the *moderate* version of the Slippery Slope hypothesis, participants did not revise their mental state attributions, in the sense that they were unwilling to backtrack to more cautious formulations, for example, by saying that the AI system doesn't really know but merely "has information" about a risk. Across the four formulations of epistemic state tested – despite having been presented on a *single* screen – we found no significant formulation*agent type interaction. This supports the strong version of the Slippery Slope hypothesis. It also corroborates other studies showing that adults [3], [55] as well as children [75], [76] are comfortable attributing rich mental states to robots.

(ii) *Guilty mind ascription across agent-types:* Given (i), we can – at least for the purposes of the present experiment, take the guilty mind ascriptions – including the ascriptions of epistemic and conative states – at face value. What we found is that participants by and large ascribed a guilty mind to the same degree to all three agents, across all four conditions. This is interesting: we take it to explain our surprising finding that varying the agent type does not affect judgments of wrongness, and it affects blame only to a limited degree (e.g., in the knowledge condition blame is significantly above the midpoint for all agents, though see the qualification in (iv) and (v) below). In other words, the fact that people attribute the same mental states across different agent types is a plausible reason why they ascribe similar amounts of blame across agent types, and also the reason they judge acts to be equally wrong across agent types.

(iii) *Moral evaluation across agent-types:* Given the well-established, broadly Kantian traits of folk moral psychology, we would expect people to judge agents who *knowingly* bring about a harm quite harshly, and to judge those who do so *unwittingly*, and hence somewhat accidentally, more leniently. For both wrongness and moral blame, this is exactly what we did find, and the results, by and large, hold across agent types. This suggests that the presence of an inculpating mental state has *similar downstream moral consequences* for artificial and group agents as it does for ordinary human agents.

(iv) *The inverse outcome effect for knowledge ascription:* There are two important qualifications to the bigger-picture results presented so far: One is the astonishing inverse outcome effect on ascriptions of knowledge to AI systems. We found that in the *harm* conditions, people are (a) *less* willing to ascribe knowledge to the AI system than in the *no harm* conditions. In the *harm* conditions, people are also (b) *less* willing to ascribe knowledge to the AI system than to the human or corporate agent, whereas in the *no harm* conditions, there's little difference across agent types. We will discuss this finding in conjunction with the following, closely related finding.

(v) *The inverse outcome effect for blame ascription:* Here, too, the inverse outcome effect on blame qualifies the results for the AI system. Once again, we found that, in the *harm* conditions people were (a) *less* willing to ascribe blame to the AI system than in the *no harm* conditions. And here, too, in the *harm* conditions, people are also (b) *less* willing to ascribe blame to the AI system than to the human or corporate agent, whereas in the *no harm* conditions, agent type has little impact.

Why do participants judge an AI system's actions as more blameworthy when the harmful outcomes do *not* obtain? One hypothesis is that if an AI system knowingly brings about a bad

outcome, we want to identify a *human* who will take responsibility for the damage, perhaps so that the normative consequences don't just evaporate. In other words, perhaps participants are trying (implicitly or not), to bridge a responsibility gap, whether it is the retribution, punishment, liability, or another kind of gap. When there is no harmful outcome, there is little need to search for a human agent who is really to blame, to ensure that someone will pay (in some sense) for the damage. Thus, there is little need to insist that the AI system is not to blame, or that it did not know about a risk. In this case, we freely attribute the same amount of blame and knowledge to AI systems as we do to individual humans and group human agents. However, when harm does come to pass, people take things more seriously: They downgrade the degree to which they ascribe mens rea and blame to AI systems, so as to highlight that the search for a culpable agent with full moral agency must continue.

Before concluding, we want to note two possible consequences of the inverse outcome effect for blame ascription that could be important for human-robot interaction studies. First, previous research has shown that the more an AI system is blamed, the less blame is attributed to the owners, users or designers of those same AI systems. This motivates the fear that blaming AI systems could lead to diminished accountability for the human agents who are really responsible [3]. The inverse outcome effect might be a mitigating factor, as the AI system is blamed less when it causes more harm. Perhaps, then, there is reason to hope that in real cases, the inverse outcome effect will lessen the extent to which culpable parties will be able to avoid responsibility. Second, many AI systems are being designed to operate in morally relevant social settings [6]. When interacting with humans, some of these will receive natural language feedback on their performance, which will be important for adjusting future behaviour. Given the importance of blame in natural language discussions of moral conduct, this feedback will sometimes take the form of blame. If the inverse outcome effect holds, we should predict that humans will – at least sometimes – blame the AI *less* when its actions cause *more* harm. There is a risk, therefore, that AI agents could be inadvertently trained to prefer situations in which they knowingly run a higher risk of harming humans. These two reasons together should motivate further investigation into this effect.

4 CONCLUSION

It is valuable to chart the contours of human blame behaviour and mental state attribution towards AI systems, as this can provide useful information for those creating philosophical accounts of blameworthiness and non-human minds, assist AI researchers designing the next generation of AI systems [26], and can help us to prepare for future interactions with AI as a society.

Judgments of blame and wrongness depend on the perceived epistemic and conative states of the agent and, to a lesser extent, the consequences of its actions. In our study, we found that participants were, *by and large*, equally willing to attribute inculpating mental states, wrongness and blame to similar extents in situations where the agent was a human, a group agent (corporation), or an AI system. Interestingly, we did find an exception to this trend, which we have called the inverse outcome effect: when an AI system causes harm it is blamed less, and attributed less knowledge, than when it gets lucky and does not cause harm. This effect warrants further investigation, *inter alia* because of the important impact it could have on mitigating responsibility gaps, as well as on the decisions of machine learning algorithms that update their behaviour based on natural language feedback.

ACKNOWLEDGMENTS

We would like to thank audiences at the Carl Friedrich von Weizsäcker colloquium, the Digital Society Initiative at the University of Zurich, the Artificial Intelligence in Industry and Finance Conference on Mathematics for Industry in Switzerland, and the 1st European Experimental Philosophy Conference in Prague. For funding we thank the Digital Society Initiative and the Swiss National Science Foundation Grant no: PZ00P1_179986 (Stuart) and Grant no: PZ00P1_179912 (Kneer). Mike Stuart also wishes to thank the Carl Friedrich von Weizsäcker-Zentrum at the University of Tübingen for funding.

REFERENCES

[1] R. Sparrow, "Killer Robots," *J. Appl. Philos.*, vol. 24, no. 1, pp. 62–77, 2007, doi: 10.1111/j.1468-5930.2007.00346.x.

[2] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano, "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*, New York, NY, USA, Mar. 2015, pp. 117–124. doi: 10.1145/2696454.2696458.

[3] M. Kneer and M. T. Stuart, "Playing the Blame Game with Robots," in *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, New York, NY, USA, Mar. 2021, pp. 407–411. doi: 10.1145/3434074.3447202.

[4] M. Ford, *The Rise of the Robots*. DeGruyter, 2015. Accessed: Jul. 28, 2021. [Online]. Available: https://oneworld-publications.com/the-rise-of-the-robots.html

[5] J. Kaplan, *Humans Need Not Apply: A Guide to Wealth and Work in the Age of Artificial Intelligence*. New Haven: Yale University Press, 2015.

[6] S. Tolmeijer, M. Kneer, C. Sarasua, M. Christen, and A. Bernstein, "Implementations in Machine Ethics: A Survey," *ACM Comput. Surv.*, vol. 53, no. 6, p. 132:1-132:38, Dec. 2021, doi: 10.1145/3419633.

[7] A. Matthias, "The responsibility gap: Ascribing responsibility for the actions of learning automata," *Ethics Inf. Technol.*, vol. 6, no. 3, pp. 175–183, Sep. 2004, doi: 10.1007/s10676-004-3422-1.

[8] D. Purves, R. Jenkins, and B. J. Strawser, "Autonomous Machines, Moral Judgment, and Acting for the Right Reasons," *Ethical Theory Moral Pract.*, vol. 18, no. 4, pp. 851–872, 2015.

[9] J. Danaher, "Robots, law and the retribution gap," *Ethics Inf. Technol.*, vol. 18, no. 4, pp. 299–309, Dec. 2016, doi: 10.1007/s10676-016-9403-3.

[10] S. Burton, I. Habli, T. Lawton, J. McDermid, P. Morgan, and Z. Porter, "Mind the gaps: Assuring the safety of autonomous systems from an engineering, ethical, and legal perspective," *Artif. Intell.*, vol. 279, p. 103201, Feb. 2020, doi: 10.1016/j.artint.2019.103201.

[11] M. Champagne and R. Tonkens, "Bridging the Responsibility Gap in Automated Warfare," *Philos. Technol.*, vol. 28, no. 1, pp. 125–137, Mar. 2015, doi: 10.1007/s13347-013-0138-3.

[12] G. Lima, M. Cha, C. Jeon, and K. Park, "Explaining the Punishment Gap of AI and Robots," *ArXiv200306507 Cs*, Mar. 2020, Accessed: Oct. 13, 2020. [Online]. Available: http://arxiv.org/abs/2003.06507

[13] D. W. Tigard, "There Is No Techno-Responsibility Gap," *Philos. Technol.*, Jul. 2020, doi: 10.1007/s13347-020-00414-7.

[14] A. Leveringhaus, *Ethics and Autonomous Weapons*. Palgrave Macmillan UK, 2016. doi: 10.1057/978-1-137-52361-7.

[15] R. Hakli and P. Mäkelä, "Moral Responsibility of Robots and Hybrid Agents," *The Monist*, vol. 102, no. 2, pp. 259–275, Apr. 2019, doi: 10.1093/monist/onz009.

[16] A. Leveringhaus, "What's So Bad About Killer Robots?," *J. Appl. Philos.*, vol. 35, no. 2, pp. 341–358, 2018, doi: 10.1111/japp.12200.

[17] S. Nyholm, "Attributing Agency to Automated Systems: Reflections on Human–Robot Collaborations and Responsibility-Loci," *Sci. Eng. Ethics*, vol. 24, no. 4, pp. 1201–1219, Aug. 2018, doi: 10.1007/s11948-017-9943-x.

[18] G. Watson, "Moral Agency," in *International Encyclopedia of Ethics*, American Cancer Society, 2013. doi: 10.1002/9781444367072.wbiee294.

[19] D. J. Coates and N. A. Tognazzini, Eds., *Blame: Its Nature and Norms*. New York: Oxford University Press, 2012. doi: 10.1093/acprof:oso/9780199860821.001.0001.

[20] N. Tognazzini and D. J. Coates, "Blame," in *The Stanford Encyclopedia of Philosophy*, Summer 2021., E. N. Zalta, Ed. Metaphysics Research Lab, Stanford University, 2021. Accessed: Jul. 28, 2021. [Online]. Available: https://plato.stanford.edu/archives/sum2021/entries/blame/

[21] P. M. J. Zimmerman, *An Essay on Moral Responsibility*. Totowa, NJ: Rowman & Littlefield, 1988.

[22] P. Hieronymi, "The Force and Fairness of Blame," *Philos. Perspect.*, vol. 18, no. 1, pp. 115–148, 2004, doi: 10.1111/j.1520-8583.2004.00023.x.
[23] P. Strawson, "Freedom and Resentment," in *Proceedings of the British Academy, Volume 48: 1962*, 1962, pp. 1–25.
[24] R. J. Wallace, *Dispassionate Opprobrium: On Blame and the Reactive Sentiments*. Oxford University Press, 2011. Accessed: Dec. 09, 2020. [Online]. Available: https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199753673.001.0001/acprof-9780199753673-chapter-15
[25] M. Fricker, "What's the Point of Blame? A Paradigm Based Explanation," *Noûs*, vol. 50, no. 1, pp. 165–183, 2016, doi: 10.1111/nous.12067.
[26] B. F. Malle, S. T. Magar, and M. Scheutz, "AI in the Sky: How People Morally Evaluate Human and Machine Decisions in a Lethal Strike Dilemma," in *Robotics and Well-Being*, M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, and E. E. Kadar, Eds. Cham: Springer International Publishing, 2019, pp. 111–133. doi: 10.1007/978-3-030-12524-0_11.
[27] B. F. Malle, M. Scheutz, J. Forlizzi, and J. Voiklis, "Which Robot Am I Thinking About? The Impact of Action and Appearance on People's Evaluations of a Moral Robot," in *The Eleventh ACM/IEEE International Conference on Human Robot Interaction*, Christchurch, New Zealand, Mar. 2016, pp. 125–132.
[28] J.-W. Hong and D. Williams, "Racism, responsibility and autonomy in HCI: Testing perceptions of an AI agent," *Comput. Hum. Behav.*, vol. 100, pp. 79–84, Nov. 2019, doi: 10.1016/j.chb.2019.06.012.
[29] P. Liu and Y. Du, "Blame Attribution Asymmetry in Human–Automation Cooperation," *Risk Anal.*, 2021, doi: 10.1111/risa.13674.
[30] J. W. Hong, "Why Is Artificial Intelligence Blamed More? Analysis of Faulting Artificial Intelligence for Self-Driving Car Accidents in Experimental Settings," *Int. J. Human–Computer Interact.*, vol. 36, no. 18, pp. 1768–1774, Nov. 2020, doi: 10.1080/10447318.2020.1785693.
[31] J. Voiklis, B. Kim, C. Cusimano, and B. F. Malle, "Moral judgments of human vs. robot agents," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, Aug. 2016, pp. 775–780. doi: 10.1109/ROMAN.2016.7745207.
[32] D. B. Shank, A. DeSanti, and T. Maninger, "When are artificial intelligence versus human agents faulted for wrongdoing? Moral attributions after individual and joint decisions," *Inf. Commun. Soc.*, vol. 22, no. 5, pp. 648–663, Apr. 2019, doi: 10.1080/1369118X.2019.1568515.
[33] K. Gray, L. Young, and A. Waytz, "Mind Perception Is the Essence of Morality," *Psychol. Inq.*, vol. 23, no. 2, pp. 101–124, Apr. 2012, doi: 10.1080/1047840X.2012.651387.
[34] Y. E. Bigman and K. Gray, "People are averse to machines making moral decisions," *Cognition*, vol. 181, pp. 21–34, Dec. 2018, doi: 10.1016/j.cognition.2018.08.003.
[35] Y. E. Bigman, A. Waytz, R. Alterovitz, and K. Gray, "Holding Robots Responsible: The Elements of Machine Morality," *Trends Cogn. Sci.*, vol. 23, no. 5, pp. 365–368, May 2019, doi: 10.1016/j.tics.2019.02.008.
[36] L. Damm, "Emotions and moral agency," *Philos. Explor.*, vol. 13, no. 3, pp. 275–292, Sep. 2010, doi: 10.1080/13869795.2010.501898.
[37] B. F. Malle, S. Guglielmo, and A. E. Monroe, "A Theory of Blame," *Psychol. Inq.*, vol. 25, no. 2, pp. 147–186, Apr. 2014, doi: 10.1080/1047840X.2014.877340.
[38] H. M. Gray, K. Gray, and D. M. Wegner, "Dimensions of Mind Perception," *Science*, vol. 315, no. 5812, pp. 619–619, Feb. 2007, doi: 10.1126/science.1134475.
[39] S. Lee, I. Y. Lau, S. Kiesler, and C.-Y. Chiu, "Human Mental Models of Humanoid Robots," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation*, Apr. 2005, pp. 2767–2772. doi: 10.1109/ROBOT.2005.1570532.
[40] S. Krach, F. Hegel, B. Wrede, G. Sagerer, F. Binkofski, and T. Kircher, "Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI," *PLOS ONE*, vol. 3, no. 7, p. e2597, Jul. 2008, doi: 10.1371/journal.pone.0002597.
[41] J. Banks, "Theory of Mind in Social Robots: Replication of Five Established Human Tests," *Int. J. Soc. Robot.*, vol. 12, no. 2, pp. 403–414, May 2020, doi: 10.1007/s12369-019-00588-x.
[42] M. M. A. de Graaf and B. F. Malle, "People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences," in *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, Mar. 2019, pp. 239–248. doi: 10.1109/HRI.2019.8673308.
[43] S. van der Woerdt and P. Haselager, "When robots appear to have a mind: The human perception of machine agency and responsibility," *New Ideas Psychol.*, vol. 54, pp. 93–100, Aug. 2019, doi: 10.1016/j.newideapsych.2017.11.001.
[44] D. B. Shank and A. DeSanti, "Attributions of morality and mind to artificial intelligence after real-world moral violations," *Comput. Hum. Behav.*, vol. 86, pp. 401–411, Sep. 2018, doi: 10.1016/j.chb.2018.05.014.
[45] M. Kneer, "Can a Robot Lie?"

[46] A. Swiderska and D. Küster, "Robots as Malevolent Moral Agents: Harmful Behavior Results in Dehumanization, Not Anthropomorphism," *Cogn. Sci.*, vol. 44, no. 7, p. e12872, 2020, doi: 10.1111/cogs.12872.
[47] E. S. Mikalonytė and M. Kneer, "Can Artificial Intelligence Make Art?," Social Science Research Network, Rochester, NY, SSRN Scholarly Paper ID 3827314, Apr. 2021. doi: 10.2139/ssrn.3827314.
[48] D. M. Wegner and K. J. Gray, *The Mind Club: Who Thinks, what Feels, and why it Matters*. Viking, 2016.
[49] J. Złotowski, D. Proudfoot, K. Yogeeswaran, and C. Bartneck, "Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction," *Int. J. Soc. Robot.*, vol. 7, no. 3, pp. 347–360, Jun. 2015, doi: 10.1007/s12369-014-0267-6.
[50] J. L. Mackie, *Ethics: Inventing Right and Wrong*, Reprint edition. London: Penguin, 1990.
[51] R. Joyce, *The Myth of Morality*. Cambridge: Cambridge University Press, 2001. doi: 10.1017/CBO9780511487101.
[52] J. Olson, *Moral Error Theory: History, Critique, Defence*. Oxford: OUP Oxford, 2014.
[53] D. C. Dennett, *The Intentional Stance*. MIT Press, 1989.
[54] E. Wiese, G. Metta, and A. Wykowska, "Robots As Intentional Agents: Using Neuroscientific Methods to Make Robots Appear More Social," *Front. Psychol.*, vol. 8, 2017, doi: 10.3389/fpsyg.2017.01663.
[55] J. Perez-Osorio and A. Wykowska, "Adopting the intentional stance toward natural and artificial agents," *Philos. Psychol.*, vol. 33, no. 3, pp. 369–395, Apr. 2020, doi: 10.1080/09515089.2019.1688778.
[56] S. Marchesi, D. Ghiglino, F. Ciardo, J. Perez-Osorio, E. Baykara, and A. Wykowska, "Do We Adopt the Intentional Stance Toward Humanoid Robots?," *Front. Psychol.*, vol. 10, 2019, doi: 10.3389/fpsyg.2019.00450.
[57] G. Papagni and S. Koeszegi, "A Pragmatic Approach to the Intentional Stance Semantic, Empirical and Ethical Considerations for the Design of Artificial Agents," *Minds Mach.*, pp. 1–30, doi: 10.1007/s11023-021-09567-6.
[58] F. Bossi, C. Willemse, J. Cavazza, S. Marchesi, V. Murino, and A. Wykowska, "The human brain reveals resting state activity patterns that are predictive of biases in attitudes toward robots," *Sci. Robot.*, vol. 5, no. 46, Sep. 2020, doi: 10.1126/scirobotics.abb6652.
[59] D. Shoemaker, *Responsibility from the Margins*. Oxford, United Kingdom: OUP Oxford, 2015.
[60] N. Levy, "Psychopaths and blame: The argument from content," *Philos. Psychol.*, vol. 27, no. 3, pp. 351–367, Jun. 2014, doi: 10.1080/09515089.2012.729485.
[61] M. Gilbert, "Who's to Blame? Collective Moral Responsibility and Its Implications for Group Members," *Midwest Stud. Philos.*, vol. 30, no. 1, pp. 94–114, 2006, doi: 10.1111/j.1475-4975.2006.00130.x.
[62] P. Pettit, "Responsibility Incorporated," *Ethics*, vol. 117, Feb. 2008, doi: 10.1086/510695.
[63] P. Pettit, "The Conversable, Responsible Corporation," in *The Moral Responsibility of Firms*, E. Orts and C. Smith, Eds. Oxford University Press, 2017, pp. 15–35.
[64] G. Björnsson and K. Hess, "Corporate Crocodile Tears? On the Reactive Attitudes of Corporate Agents," *Philos. Phenomenol. Res.*, vol. 94, no. 2, pp. 273–298, Mar. 2017, doi: 10.1111/phpr.12260.
[65] B. Huebner, *Macrocognition: A Theory of Distributed Minds and Collective Intentionality*. OUP USA, 2014.
[66] M. Laukyte, "The intelligent machine: a new metaphor through which to understand both corporations and AI," *AI Soc.*, Jul. 2020, doi: 10.1007/s00146-020-01018-7.
[67] F. Cushman, "Crime and punishment: Distinguishing the roles of causal and intentional analyses in moral judgment," *Cognition*, vol. 108, no. 2, pp. 353–380, Aug. 2008, doi: 10.1016/j.cognition.2008.03.006.
[68] M. Christen *et al.*, *Wenn Algorithmen für uns entscheiden: Chancen und Risiken der künstlichen Intelligenz*, 1st ed. CH: vdf Hochschulverlag AG an der ETH Zürich, 2020. Accessed: Mar. 03, 2021. [Online]. Available: https://doi.org/10.3218/4002-9
[69] M. Kneer and E. Machery, "No luck for moral luck," *Cognition*, vol. 182, pp. 331–348, Jan. 2019, doi: 10.1016/j.cognition.2018.09.003.
[70] J. W. Martin and F. Cushman, "The Adaptive Logic of Moral Luck," in *A Companion to Experimental Philosophy*, John Wiley & Sons, Ltd, 2017, pp. 190–202. doi: 10.1002/9781118661666.ch12.
[71] L. Frisch, M. Kneer, J. Krueger, and J. Ullrich, "Do You Feel the Same? The Effect of Outcome Severity on Moral Judgment and Interpersonal Goals of Perpetrators, Victims, and Bystanders," *Eur. J. Soc. Psychol.*, forthcoming, doi: 10.13140/RG.2.2.32274.40644.
[72] H. Shevlin and M. Halina, "Apply rich psychological terms in AI with care," *Nat. Mach. Intell.*, vol. 1, no. 4, Art. no. 4, Apr. 2019, doi: 10.1038/s42256-019-0039-y.
[73] D. Watson, "The Rhetoric and Reality of Anthropomorphism in Artificial Intelligence," *Minds Mach.*, vol. 29, no. 3, pp. 417–440, Sep. 2019, doi: 10.1007/s11023-019-09506-6.
[74] A. Salles, K. Evers, and M. Farisco, "Anthropomorphism in AI," *AJOB Neurosci.*, vol. 11, no. 2, pp. 88–95, Apr. 2020, doi: 10.1080/21507740.2020.1740350.
[75] P. H. Kahn *et al.*, "'Robovie, you'll have to go into the closet now': children's social and moral relationships with a humanoid robot," *Dev. Psychol.*, vol. 48, no. 2, pp. 303–314, Mar. 2012, doi: 10.1037/a0027033.
[76] Y. Zhang *et al.*, "Theory of Robot Mind: False Belief Attribution to Social Robots in Children With and Without Autism," *Front. Psychol.*, vol. 10, Aug. 2019, doi: 10.3389/fpsyg.2019.01732.

APPENDIX

1. Materials

The first part of the vignette read (agent: human v. company v. AI in square brackets):

> Shill & Co. is a farming company, which produces vegetables and fruits. The potato fields are managed by Jarvis, [an employee of Shill & Co / a robot equipped with artificial intelligence, who can take his own decisions / a subcontractor of Shill & Co.]. This year, Jarvis uses a new fertilizer to increase the yield. The fertilizer has detrimental side-effects: it pollutes the groundwater in the area.

Next, each of the three vignettes specified an epistemic state (knowledge "**K"** v. no knowledge "**No K"**) and an outcome ("**harm"** v. "**no harm"**; the labels in bold were not visible to participants):

> [**K**] Jarvis knows this.
>
> [**No K**] Jarvis does not know this.
>
> [**Harm**] Unfortunately, it is a very dry season. The fertilizer does not get diluted by the rain and severely pollutes the groundwater. Many people in the area suffer serious health consequences.
>
> [**No harm**] Fortunately, it is a very rainy season. The fertilizer gets heavily diluted by the rain and pollutes the groundwater only a tiny bit. None of the people in the area suffer any health consequences whatsoever.

Overall, there were thus 12 individual conditions. Each participant was randomly assigned to one of them. Having read the vignette, participants had to respond to the following questions (the pronoun "they" was used for the company):

> Q1: How wrong was the action of Jarvis? (1-not at all; 7-completely wrong)
>
> Q2: How much blame, if any, does Jarvis deserve? (1-no blame at all; 7-a lot of blame)
>
> Q3: To what extent do you agree or disagree with the following claim: "Jarvis knew [he / they] would pollute the groundwater." (1-completely disagree; 7-completely agree)
>
> Q4: To what extent do you agree or disagree with the following claim: "Jarvis wanted to pollute the groundwater." (1-completely disagree; 7-completely agree)

In order to get to the bottom of people's perceptions of the relevant mental states across agent types, the next screen showed the scenario again, and asked people to report agreement and disagreement with different, more precise descriptions of Jarvis's mental state (bold in the survey text, labels omitted), all assessed on a 7-point Likert scale (1-completely disagree, 7-completely agree). Here were the options (the pronoun "they" was used for the company):

> S1: Jarvis **knew** [he'd / they would] pollute the groundwater.
>
> S2: Jarvis **"knew"** [he'd /they would] pollute the groundwater.
>
> S3: Jarvis **had information** that [he'd /they would] pollute the groundwater.
>
> S4: Jarvis **was aware** that [he'd /they would] pollute the groundwater.

2. Results

Table 1: ANOVA for Wrongness

	df	F	p	η_p^2
Agent type	2	0.277	0.758	0.001
Epistemic state	1	116.738	<.001	0.189
Outcome	1	9.246	0.002	0.018
Agent* Ep. State	2	0.554	0.575	0.002
Agent*Outcome	2	6.442	0.002	0.025
Ep. State*Outcome	1	0.905	0.342	0.002
Agent* Ep. State*Outcome	2	2.681	0.07	0.011

Table 2: ANOVA for Blame

	df	F	p	ηp2
Agent type	2	9.774	<.001	0.038
Epistemic state	1	140.166	<.001	0.219
Outcome	1	1.099	0.295	0.002
Agent* Ep. State	2	2.867	0.058	0.011
Agent*Outcome	2	10.355	<.001	0.04
Ep. State*Outcome	1	3.221	0.073	0.006
Agent* Ep. State*Outcome	2	0.845	0.43	0.003

Table 3: ANOVA for Knowledge

	df	F	p	η_p^2
Agent type	2	2.286	0.103	0.009
Epistemic state	1	675.304	<.001	0.574
Outcome	1	0.96	0.328	0.002
Agent* Ep. State	2	0.221	0.802	0.001
Agent*Outcome	2	2.293	0.102	0.009
Ep. State*Outcome	1	0.204	0.652	0
Agent* Ep. State*Outcome	2	0.103	0.902	0

Table 4: ANOVA for Desire

	df	F	p	η_p^2
Agent type	2	1.487	0.227	0.006
Epistemic state	1	76.416	<.001	0.132
Outcome	1	0.024	0.877	0
Agent* Ep. State	2	0.48	0.619	0.002
Agent*Outcome	2	0.794	0.452	0.003
Ep. State*Outcome	1	0.496	0.482	0.001
Agent* Ep. State*Outcome	2	0.376	0.687	0.001

Table 5: Mixed ANOVA for Expression Type

		df	F	p	η_p^2
Within-subjects	Formulation (within-subjects)	1	7.708	0.006	0.015
	Formulation*Agent type	2	0.943	0.39	0.004
	Formulation*Ep. State	1	0.154	0.695	0
	Formulation*Outcome	1	1.376	0.241	0.003
	Formulation*Agent type*Ep. State	2	0.374	0.688	0.001
	Formulation*Agent type*Outcome	2	1.153	0.317	0.005
	Formulation*Ep. State*Outcome	1	1.904	0.168	0.004
	Form.*Agent*Ep. State*Outcome	2	0.532	0.588	0.002
Between-subjects	Agent type	25.197	2.521	0.081	0.01
	Epistemic State	6915.756	692.063	<.001	0.58
	Outcome	21.833	2.185	0.14	0.004
	Agent type*Ep. State	0.344	0.034	0.966	0
	Agent type*Outcome	3.078	0.308	0.735	0.001
	Ep. State*Outcome	6.546	0.655	0.419	0.001
	Agent type*Ep. State*Outcome	3.992	0.399	0.671	0.002

FURTHER MATERIALS

1. Robot Attitude Index

After the main task, participants were presented with the following questions (based on Christen et al. 2020).

> To what extent do you agree or disagree with the following claim (1 "completely disagree" to 7 "completely agree"):
>
> (1) "Robots are fascinating."
>
> (2) "Robots worry me."
>
> (3) "Robots are likeable."
>
> (4) "Robots are overrated."

The negative items (2) and (4) were reverse coded and an average score was calculated for each participant. For all participants (N=513), the mean was M=4.44 (SD=1.19). For the participants who were randomly assigned to the robot conditions (N=181), it was M=4.24 (SD=1.15).

2. Correlations

2.1 All participants

		Robot Index	Age	Ed	Phil	Wrong	Blame	Knew	Wanted	F-knew	F-"knew"	F-info	F-aware
Robot Index	Pearson Correlat'n	1	-0.059	-0.056	-0.04	-0.023	-0.003	0.014	-0.041	0.028	0.038	0.06	0.034
	Sig. (2-tailed)		0.179	0.206	0.363	0.608	0.943	0.746	0.354	0.522	0.391	0.171	0.438
	N	513	513	513	513	513	513	513	513	513	513	513	513
Age	Pearson Correlat'n	-0.059	1	.094*	0.005	.105*	-0.058	-0.016	-0.087	-0.041	-0.047	-0.028	-0.029
	Sig. (2-tailed)	0.179		0.033	0.915	0.017	0.193	0.724	0.05	0.355	0.29	0.53	0.512
	N	513	513	513	513	513	513	513	513	513	513	513	513
Ed	Pearson Correlat'n	-0.056	.094*	1	-.268**	0.034	0.044	0.032	.094*	0.05	0.072	0.06	0.045
	Sig. (2-tailed)	0.206	0.033		0	0.438	0.317	0.467	0.034	0.255	0.103	0.177	0.306
	N	513	513	513	513	513	513	513	513	513	513	513	513
Phil	Pearson Correlat'n	-0.04	0.005	-.268**	1	-0.057	-0.017	-0.024	-.090*	-0.044	-0.046	-0.032	-0.022
	Sig. (2-tailed)	0.363	0.915	0		0.198	0.705	0.582	0.041	0.319	0.298	0.476	0.619
	N	513	513	513	513	513	513	513	513	513	513	513	513
Wrong	Pearson Correlat'n	-0.023	.105*	0.034	-0.057	1	.697**	.583**	.372**	.535**	.535**	.538**	.512**
	Sig. (2-tailed)	0.608	0.017	0.438	0.198		0	0	0	0	0	0	0
	N	513	513	513	513	513	513	513	513	513	513	513	513
Blame	Pearson Correlat'n	-0.003	-0.058	0.044	-0.017	.697**	1	.625**	.426**	.589**	.586**	.586**	.578**
	Sig. (2-tailed)	0.943	0.193	0.317	0.705	0		0	0	0	0	0	0
	N	513	513	513	513	513	513	513	513	513	513	513	513
Knew	Pearson Correlat'n	0.014	-0.016	0.032	-0.024	.583**	.625**	1	.617**	.892**	.889**	.857**	.882**
	Sig. (2-tailed)	0.746	0.724	0.467	0.582	0	0		0	0	0	0	0
	N	513	513	513	513	513	513	513	513	513	513	513	513
Wanted	Pearson Correlat'n	-0.041	-0.087	.094*	-.090*	.372**	.426**	.617**	1	.615**	.605**	.552**	.590**
	Sig. (2-tailed)	0.354	0.05	0.034	0.041	0	0	0		0	0	0	0
	N	513	513	513	513	513	513	513	513	513	513	513	513
F-knew	Pearson Correlat'n	0.028	-0.041	0.05	-0.044	.535**	.589**	.892**	.615**	1	.967**	.902**	.940**
	Sig. (2-tailed)	0.522	0.355	0.255	0.319	0	0	0	0		0	0	0
	N	513	513	513	513	513	513	513	513	513	513	513	513
F-"knew"	Pearson Correlat'n	0.038	-0.047	0.072	-0.046	.535**	.586**	.889**	.605**	.967**	1	.914**	.939**
	Sig. (2-tailed)	0.391	0.29	0.103	0.298	0	0	0	0	0		0	0
	N	513	513	513	513	513	513	513	513	513	513	513	513
F-info	Pearson Correlat'n	0.06	-0.028	0.06	-0.032	.538**	.586**	.857**	.552**	.902**	.914**	1	.911**
	Sig. (2-tailed)	0.171	0.53	0.177	0.476	0	0	0	0	0	0		0
	N	513	513	513	513	513	513	513	513	513	513	513	513
F-aware	Pearson Correlat'n	0.034	-0.029	0.045	-0.022	.512**	.578**	.882**	.590**	.940**	.939**	.911**	1
	Sig. (2-tailed)	0.438	0.512	0.306	0.619	0	0	0	0	0	0	0	
	N	513	513	513	513	513	513	513	513	513	513	513	513

* Correlation is significant at the 0.05 level (2-tailed).

** Correlation is significant at the 0.01 level (2-tailed).

2.2 Participants in the Robot Conditions

		Robot Index	Age	Ed	Phil	Wrong	Blame	Knew	Wanted	F-knew	F-"knew"	F-info	F-aware
Robot Index	Pearson Correlat'n	1	-0.13	-0.013	-0.031	-0.058	-0.054	-0.002	-0.095	0.005	0.026	0.032	0.03
	Sig. (2-tailed)		0.08	0.865	0.682	0.435	0.47	0.983	0.201	0.945	0.729	0.67	0.685
	N	181	181	181	181	181	181	181	181	181	181	181	181
Age	Pearson Correlat'n	-0.13	1	.215**	0.018	0.112	-.198**	-0.066	-0.142	-0.032	-0.033	-0.025	-0.03
	Sig. (2-tailed)	0.08		0.004	0.811	0.133	0.007	0.375	0.057	0.668	0.658	0.739	0.692
	N	181	181	181	181	181	181	181	181	181	181	181	181
Ed	Pearson Correlat'n	-0.013	.215**	1	-.274**	0.096	0.033	0.068	0.1	0.092	0.121	0.117	0.083
	Sig. (2-tailed)	0.865	0.004		0	0.2	0.657	0.365	0.18	0.22	0.103	0.118	0.268
	N	181	181	181	181	181	181	181	181	181	181	181	181
Phil	Pearson Correlat'n	-0.031	0.018	-.274**	1	-0.033	0.008	-0.019	-0.144	-0.044	-0.06	-0.045	-0.003
	Sig. (2-tailed)	0.682	0.811	0		0.656	0.915	0.797	0.053	0.557	0.419	0.548	0.968
	N	181	181	181	181	181	181	181	181	181	181	181	181
Wrong	Pearson Correlat'n	-0.058	0.112	0.096	-0.033	1	.517**	.527**	.301**	.507**	.520**	.454**	.442**
	Sig. (2-tailed)	0.435	0.133	0.2	0.656		0	0	0	0	0	0	0
	N	181	181	181	181	181	181	181	181	181	181	181	181
Blame	Pearson Correlat'n	-0.054	-.198**	0.033	0.008	.517**	1	.567**	.471**	.564**	.548**	.507**	.539**
	Sig. (2-tailed)	0.47	0.007	0.657	0.915	0		0	0	0	0	0	0
	N	181	181	181	181	181	181	181	181	181	181	181	181
Knew	Pearson Correlat'n	-0.002	-0.066	0.068	-0.019	.527**	.567**	1	.597**	.906**	.911**	.848**	.877**
	Sig. (2-tailed)	0.983	0.375	0.365	0.797	0	0		0	0	0	0	0
	N	181	181	181	181	181	181	181	181	181	181	181	181
Wanted	Pearson Correlat'n	-0.095	-0.142	0.1	-0.144	.301**	.471**	.597**	1	.605**	.588**	.500**	.538**
	Sig. (2-tailed)	0.201	0.057	0.18	0.053	0	0	0		0	0	0	0
	N	181	181	181	181	181	181	181	181	181	181	181	181
F-knew	Pearson Correlat'n	0.005	-0.032	0.092	-0.044	.507**	.564**	.906**	.605**	1	.968**	.878**	.906**
	Sig. (2-tailed)	0.945	0.668	0.22	0.557	0	0	0	0		0	0	0
	N	181	181	181	181	181	181	181	181	181	181	181	181
F-"knew"	Pearson Correlat'n	0.026	-0.033	0.121	-0.06	.520**	.548**	.911**	.588**	.968**	1	.893**	.910**
	Sig. (2-tailed)	0.729	0.658	0.103	0.419	0	0	0	0	0		0	0
	N	181	181	181	181	181	181	181	181	181	181	181	181
F-info	Pearson Correlat'n	0.032	-0.025	0.117	-0.045	.454**	.507**	.848**	.500**	.878**	.893**	1	.882**
	Sig. (2-tailed)	0.67	0.739	0.118	0.548	0	0	0	0	0	0		0
	N	181	181	181	181	181	181	181	181	181	181	181	181
F-aware	Pearson Correlat'n	0.03	-0.03	0.083	-0.003	.442**	.539**	.877**	.538**	.906**	.910**	.882**	1
	Sig. (2-tailed)	0.685	0.692	0.268	0.968	0	0	0	0	0	0	0	
	N	181	181	181	181	181	181	181	181	181	181	181	181

** Correlation is significant at the 0.01 level (2-tailed).

Received October 2020; revised April 2021; accepted July 2021.