UNIVERSITY OF HELSINKI

https://helda.helsinki.fi

Helda

Innocence over utilitarianism - Heightened Moral Standards for Robots in Rescue Dilemmas

Sundvall, Jukka

John Wiley and Sons Ltd
2023-06

Sundvall, J, Drosinou, M-A, Hannikainen, I, Elovaara, K M, Halonen, J P, Herzon, V, Kopecký, R, Jirout Košová, M, Koverola, M, Kunnari, A J O, Perander, S L L, Saikkonen, T J, Palomäki, J P & Laakasuo, M 2023, 'Innocence over utilitarianism - Heightened Moral Standards for Robots in Rescue Dilemmas', European Journal of Social Psychology, vol. 53, no. 4, pp. 779-804. https://doi.org/10.1002/ejsp.2936

http://hdl.handle.net/10138/572525
10.1002/ejsp.2936

cc_by
publishedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

Check for updates

Received: 25 October 2021 | Accepted: 12 January 2023

DOI: 10.1002/ejsp.2936

RESEARCH ARTICLE

EASP WILEY

Innocence over utilitarianism: Heightened moral standards for robots in rescue dilemmas

Jukka Sundvall[1] | Marianna Drosinou[2] | Ivar Hannikainen[3] | Kaisa Elovaara[1] | Juho Halonen[1] | Volo Herzon[2] | Robin Kopecký[4,5] | Michaela Jirout Košová[5] | Mika Koverola[1] | Anton Kunnari[1,2] | Silva Perander[6] | Teemu Saikkonen[7] | Jussi Palomäki[1] | Michael Laakasuo[1]

[1]Department of Digital Humanities, Cognitive Science, Faculty of Arts, University of Helsinki, Helsinki, Finland

[2]Department of Psychology and Logopedics, Faculty of Medicine, University of Helsinki, Helsinki, Finland

[3]Department of Philosophy I, Faculty of Psychology, Cartuja Campus, Universidad de Granada, Spain

[4]Institute of Philosophy, Czech Academy of Sciences, Prague, Czechia

[5]Department of Philosophy and History of Science, Faculty of Science, Charles University, Prague, Czechia

[6]Department of Computer Science, Faculty of Science, University of Helsinki, Helsinki, Finland

[7]Zoological Museum, Biodiversity Unit, University of Turku, Turku, Finland

Correspondence
Jukka Sundvall, Faculty of Arts, University of Helsinki, Vuorikatu 3A, H541, Helsinki 00014, Finland.
Email: jukka.sundvall@helsinki.fi

Funding information
Tiina and Antti Herlin Foundation; The Academy of Finland, Grant/Award Number: 323207; Weisell Foundation; Strategic Research Council, Grant/Award Number: 345186; Jane and Aatos Erkko Foundation, Grant/Award Number: 170112

Abstract

Research in moral psychology has found that robots, more than humans, are expected to make utilitarian decisions. This expectation is found specifically when contrasting utilitarian action to deontological inaction. In a series of eight experiments (total $N = 3752$), we compared judgments about robots' and humans' decisions in a rescue dilemma with no possibility of deontological inaction. A robot's decision to rescue an innocent victim of an accident was judged more positively than the decision to rescue two people culpable for the accident (Studies 1–2b). This pattern repeated in a large-scale web survey (Study 3, $N = \sim 19{,}000$) and reversed when all victims were equally culpable/innocent (Study 5). Differences in judgments about humans' and robots' decisions were largest for norm-violating decisions. In sum, robots are not always expected to make utilitarian decisions, and their decisions are judged differently from those of humans based on other moral standards as well.

KEYWORDS
folk ethics, folk justice, moral dilemma, rescue robotics, utilitarianism

1 | INTRODUCTION

In the film *I, Robot*, the protagonist, Detective Del Spooner, has grown distrustful of the robots ubiquitous in society. Spooner's aversion to robots is not caused by any hostile episode. Quite the contrary: a robot had *saved* Spooner's life, rescuing him from a sinking car. In saving Spooner, the robot chose to leave a 12-year-old child in the car to drown, based on the child's lower likelihood of survival.

Fictional as it may be, this scenario reflects important questions concerning the role of autonomous artificial agents in society

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
© 2023 The Authors. *European Journal of Social Psychology* published by John Wiley & Sons Ltd.

today. Artificial intelligence (AIs; software solutions with independent decision-making capacities) already make decisions that impact human lives, for instance, in the form of parole decisions (O'Neil, 2016). Robots—machines equipped with AI (i.e., decision-making software that has capacities to plan and execute goals and striving towards goals), built to serve different functions—may someday make rescue decisions, too. What expectations do we have of such robots, and the way they will make life-saving decisions? If we are to avoid people developing a Spooner-like disregard for robotic agents, what are the moral standards that these robots should support? As technological progress marches onwards (rescue robotics is a growing field; Walker, 2019), it becomes increasingly relevant to examine how the public reacts to automated decisions.

1.1 | Robots and AIs as moral agents?

Robots and AIs are anticipated to play a growing role in high-risk occupations: helping to perform surgeries, aiding in military and policing operations, disaster relief, and emergency response teams (Marcus & Davis, 2019). The transition towards an automated workforce in these areas promises to reduce casualties and human error. Furthermore, developments in AI may lead to robots not being expected simply to *enact* human decisions, but often to autonomously determine which course of action to pursue (Marcus & Davis, 2019; Mitchell, 2019). Rescue robots, currently mostly controlled by human users, could guide themselves (Walker, 2019); military robots could decide on when and where to strike (although it is uncertain whether this will ever be allowed; see Campaign to Stop Killer Robots, 2019; Davison, 2017).

By autonomously determining which course of action to pursue in contexts where their actions affect human lives, robots and AIs would be making decisions vested with moral relevance (Awad et al., 2018; Bigman & Gray, 2018; Bigman et al., 2019; Wallach & Allen, 2008). We will use the term 'artificial agent' (after Wallach & Allen, 2008) to refer collectively to robots and AIs in roles that allow them to make decisions that are morally relevant to humans.

Are people even willing to consider artificial agents as eligible agents for making moral decisions? Bigman and Gray (2018) found that people generally consider robots as less appropriate moral decision-makers than humans, regardless of the type of moral decision or outcome. If robots are viewed as inappropriate agents for moral decision-making to begin with, this could explain differences between moral judgments of decisions made by humans and robots. One cause for this may be the perceived 'lesser mindedness' of robots, whereby they are consistently perceived as having weaker emotional capabilities than humans (de Graaf & Malle, 2019; Gamez et al., 2020; Gray et al., 2007; Laakasuo et al., 2021a; Weisman et al., 2017).

1.2 | Artificial agents are not always worse

People do not necessarily judge moral decisions by artificial agents as overall *worse* than decisions made by humans regardless of what the decision was. Rather, a growing body of research demonstrates that some decisions can be viewed as morally better when made by an artificial agent than a human, while others are indeed viewed as worse. Additionally, when judging different decisions an agent could make, people may judge a given decision as clearly better than another for one type of agent, but less so for another.

For example, Malle et al. (2015) found that a human was *blamed* more for making a utilitarian sacrifice (killing one to save five) than taking the deontological option (do nothing, allow five people to die) in a version of the classic trolley dilemma (see Christensen & Gomila, 2012; Greene, 2007).[1] A robot was *blamed* to an equal degree for both options, but not to a greater degree than a human. That is, Malle et al. (2015) found that people assigned blame to humans depending on their decisions, but assigned blame to robots regardless of their decisions. Ratings of *wrongness* (in contrast to blameworthiness) showed that the robot's action was judged as more morally wrong when it took the deontological option, whereas the human's action was judged as more wrong when they took the utilitarian option. Starr et al. (2021) found, in a variant of the trolley dilemma assessing human and robot advisors, that a robot advisor was regarded as more *trustworthy* and *reliable* when it recommended taking the utilitarian option, while human advisors were regarded less *blameworthy* and more *ethical* than robot advisors when they recommended the deontological option. Komatsu et al. (2021) found that both Japanese and US samples judged a robot more negatively than a human specifically for deontological inaction.

Other studies have specifically compared robot and human agents occupying specialized professional roles. Malle et al. (2019) asked participants to provide judgments of a decision in a military context: a human pilot or an AI is given an order to launch a missile strike that could collaterally kill bystanders, and decides to either obey or disobey the order. Participants gave higher ratings of blame to the human pilot for disobeying than for obeying, whereas the AI agent was blamed equally for both decisions. Critically, the AI was blamed less for disobeying than the human was—that is, an artificial agent's decision was judged more positively than the same decision by a human. In the domain of medical care, Laakasuo et al. (2022) found that robot nurses were judged more *negatively* than human nurses for forcibly medicating a patient. However, when the robot nurse refused to forcibly medicate the patient, it was judged equally to or more *positively* than a human nurse reaching the same decision. Collectively, the aforementioned studies suggest that the decisions of artificial agents are not always evaluated more negatively, but rather that they are held to different moral standards than humans.

1.3 | Anthropomorphism and the mindedness of artificial agents

A basic question in the moral psychology of artificial agents concerns how people would judge the actions of entities that differ from the

[1] Both here and in discussing our own studies we will use the word 'utilitarianism' as a shorthand for 'utilitarianism as it is commonly operationalized in psychological studies'. We do not claim that the participants who preferred a utilitarian option in the studies we cite or our own studies necessarily held a comprehensive utilitarian moral philosophy. What we do claim is that they exhibited a utilitarian-like preference.

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

prototypical moral agent, an adult human. A logical follow-up question is whether the human-likeness of an agent matters for moral judgment: if the actions of a human and a non-human are judged differently, what happens if the non-human appears human in one way or another? Does it matter which aspect of the non-human agent becomes more human-like?

People often attribute human-like qualities to artificial agents, that is, they may anthropomorphize such agents. Złotowski et al. (2015), in a review of anthropomorphism in human–robot-interaction (HRI), argued that the extent to which people attribute 'humanness' to robots depends on the combination of the human-likeness of the robot's appearance and the human-likeness of its behaviour when interacting with humans. According to Złotowski et al. (2015), the latter may be more important than appearance. Factors such as verbal and non-verbal communication, a robot's autonomy, perceived emotionality or intelligence, and predictability may all contribute to perceptions of a robot's 'humanness', more so than human-like physical features. The tasks that robots are intended to carry out are also a factor: people attribute more human-like qualities to robots as a function of the seriousness of their tasks (Hancock et al., 2011). This does not mean that physical features are irrelevant, however. Minor alterations in the shape of a humanoid robot influence trait attribution (Trovato et al., 2018), and simply the colour of a robot is associated with implicit properties projected onto it (Addison et al., 2019). To further complicate this, the associations are not necessarily linear: the so-called Uncanny Valley effect is a commonly experienced unsettling feeling people have when humanoid robots and/or audio visual simulations closely resemble humans but are not entirely convincing (Palomäki et al., 2018).

Bigman et al. (2019) suggested that we should expect people to attribute more moral responsibility to robots as their human-likeness in appearance or behaviour increases. This would match with findings suggesting that a robot's human-like appearance can increase expectations of human-like behaviour in non-moral domains, such as following human norms related to personal space (Syrdal et al., 2008) or being able to empathize and understand requests (Kwon et al., 2016). There are few studies on anthropomorphism's effects on moral judgment, with somewhat conflicting results. Malle et al. (2016) found that a previously reported (Malle et al., 2015) human–robot asymmetry in judgments about decisions in a variation of the trolley dilemma only held for non-humanoid robots. A non-humanoid robot was blamed more for a deontological than a utilitarian decision, whereas the trend was the opposite for the decisions of both humans and humanoid robots. In contrast to this finding, Laakasuo et al. (2021a) found, using variations of the trolley dilemma, that decisions by robots with human-like facial features were judged as less moral than similar decisions by either humans or robots without human-like facial features. Given many methodological differences between Laakasuo et al. (2021a) and Malle et al. (2016), it may be that the difference in results stems from, for example, nuances in how a human-like robot is depicted (face vs. overall human-like shape), or differences between judgments of morality and blame.

The human-likeness of robots, either in appearance, behaviour, or cognitive capacities, may also increase aversion towards them. Yogeeswaran et al. (2016) presented participants with staged video interviews of either a robot that resembled a human very closely, or a robot with a humanoid shape that was still clearly recognizable as a robot. In the videos, it was stated either that the robot could perform various physical and mental tasks, or that it could outperform humans in those tasks. In the latter case only, participants reported perceiving robots in general as a threat significantly more if they had seen the video of the more human-like robot than if they had seen the less human-like robot. Additionally, informing participants about a robot's ability to outperform humans did not increase perceived threat if the robot's appearance was less human-like. Similarly, Złotowski et al. (2017) showed participants videos of various kinds of robots, combined with either a statement that the robots are capable of autonomous decision-making, or a statement that the robots are only able to follow human commands. Again, participants perceived robots as more of a threat when they were told about the robots' capacity for autonomous action—something that would make them more human-like than only being able to follow orders.

To summarize, people project human-like qualities onto non-human artificial agents based on appearance and behaviour. Because of this anthropomorphization, people may come to expect seemingly human-like artificial agents to behave in ways that align with their expectations of human behaviour. However, human-like artificial agents can also feel more threatening to people. Anthropomorphism seems to have an effect on moral judgment, but its exact nature is unclear: there are results that align with the idea of more human-like expectations for human-like artificial agents, and results that align with the idea of aversion towards human-like artificial agents ('uncanny valley'; see Palomäki et al., 2018). To contribute to understanding this effect, we included descriptions of robots that had a human-like appearance and/or a human-like capacity for independent decision-making in our Studies 2a–2c. We wanted to examine the effects of both appearance and cognitive capacities on moral judgment, and to control for potential default assumptions about how a robot looks (Phillips et al., 2017).

1.4 | Focus on utilitarianism

Perhaps the most popular research topic within moral psychology has been people's responses to variations of the classic trolley dilemma: whether to sacrifice one human life to save many others from being run over and killed by a runaway trolley. This question has also been examined in psychological studies of robot and AI ethics, for instance, using adaptations of the dilemma to pending questions concerning the design and programming of autonomous vehicles. Awad et al. (2018) showed that people across the world prefer autonomous vehicles to enact utilitarian principles. Specifically, when considering inevitable accidents, people prefer autonomous vehicles to save the largest number of lives, even if doing so 'sacrifices' the passengers inside the vehicle. Similarly, while a human's utilitarian decision in a trolley dilemma may be judged more negatively than a deontological one, judgments of robots' decisions may have the opposite trend (Malle et al., 2015; Voiklis et al., 2016). Taken together, prior research emphasizes people's tendency to

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

impose utilitarian moral standards on robots, even when they would not hold humans to the same moral expectations.

Yet a recurring criticism of trolley dilemma studies is that they pit utilitarian *action* against deontological *inaction* (Gawronski et al., 2017; Kahane et al., 2015). As such, the measurement of deontology—an ethical stance emphasizing that the ends do not justify the means—is conflated with a preference for inaction. This opens up the possibility that differences in the moral evaluation of robots versus humans may not only be driven by greater expectations of utilitarian reasoning for robots (or deontological reasoning for humans). Instead, they may also be driven by a more general tolerance for human inaction relative to robot inaction in difficult situations (see omission bias in Baron & Ritov, 2004; preference for inaction in Gawronski & Beer, 2016; Gawronski et al., 2017). Indeed, a robot that remains inactive when facing a moral dilemma may make people question why the robot was given the task in the first place (Komatsu et al., 2021, argued similarly: human inaction is easier to understand than robot inaction). The studies in this article sidestep this ambiguity by presenting dilemmas that contrast two courses of *action*, and eliminating the possibility of *inaction*.

1.5 | Current studies

We focused on a vignette in which an accidental situation leads to a moral dilemma. In the original version of the dilemma, two drunken motorboaters crash into a fisherman's boat, the damaged boats drift far apart and all parties fall into the sea. The situation is observed by either a human or a robot lifeguard who must decide which party to rescue: the two motorboaters whose behaviour caused the accident, or the innocent fisherman. The agent only has time to rescue one of the parties. As a short summary of our results, in several variations of the moral dilemma, we found that robot agents were expected to adhere more strictly to both utilitarian and folk justice norms than humans, depending on which norm was the most salient in each variation.

Instead of contrasting utilitarian action with deontological inaction, we wanted to contrast utilitarianism with an underexplored moral concern: the innocence or culpability of the party to be saved in the dilemma. Arguably, according to utilitarian calculus, it is better to save the larger number of human lives (i.e., the drunken motorboaters), regardless of their innocence or culpability. However, people also have moral intuitions relating to fairness (Curry et al., 2019; Haidt & Graham, 2007) and 'just deserts' (Carlsmith & Darley, 2008). Is it fair to leave the fisherman to drown, given that he was not at fault for the accident? Conversely, should the people most obviously responsible for the accident 'get what they deserve'? It is not obvious whether punitive motives or a wish to protect the innocent would play the bigger role, but in our scenario both lines of thinking point in the same direction: prioritize innocent lives. In order to avoid bias towards interpreting results in terms of protecting the innocent or punishing the guilty, we will refer to this idea more generally as a folk sense of justice or simply folk justice.

We chose our particular vignette setting firstly because both rescue options are active decisions which are motivated by the role and duties of the rescue agent. Contrary to standard trolley dilemmas, this decision dilemma does not juxtapose omissions (passive decisions, e.g. 'do not pull the lever, allow five people to be run over') and commissions (active decisions, 'pull the lever to redirect the trolley to run over one person'), which are also confounded with deontological ('it is not okay to sacrifice one to save five') and utilitarian preferences ('it is okay to sacrifice one to save five'), respectively. In our vignette, both decisions are commissions, and only one of them is a utilitarian decision (to save the two motorboaters). The second motive for this dilemma choice was that while studies with deontological *action* options have been conducted (e.g., Gawronski et al., 2017), studies contrasting utilitarianism with moral concerns other than deontology (refraining from doing harm) are scarcer. Our studies examined whether robots are expected to consider such non-utilitarian standards to a different extent than humans are.

Throughout our studies, we investigated the extent to which utilitarian and non-utilitarian concerns in a rescue dilemma shape moral evaluations of robotic versus human lifeguards. In Study 1, we set the baseline for our subsequent studies by examining whether participants judged the decisions of a robot lifeguard differently from those of a human. The study pitted utilitarianism and folk justice against each other. We found that participants judged the robot, but not the human, more negatively for making the utilitarian choice of rescuing the culpable party. In Studies 2a–2c, we examined factors that may have contributed to differing judgments: the descriptions of the robot agent (appearance and cognitive capacities) and the culpable party (socio-economic status). We found that participants were more likely to judge the decisions of specifically human-like robots differently from the decisions of humans. We also found that the socio-economic status of the culpable party had an effect on judgments, but this was overshadowed by the effect of culpability in itself. In Studies 3–5, we manipulated the number of culpable people and whether or not the utilitarian choice aligned with folk justice or not. In sum, we found again that the effect of culpability was stronger than the effect of utilitarian concerns, and that participants consistently judged the decisions of robot agents more negatively than those of human agents specifically when the decision was the 'worst possible' moral choice in a scenario.

2 | STUDY 1

The purpose of this study was to explore the conflict between saving an innocent life versus a larger number of culpable lives. The practical aim was to test the materials and explore possible candidates for dependent variables.

2.1 | Method

2.1.1 | Participants and design

We recruited 216 Finnish-speaking adults to take part in an online experiment through announcements on University of Helsinki student

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 1 Sensitivity power analyses for main effects

		Effect size		Power		
Study	Manipulation	Observed	Needed for 80% power	Small effect	Medium effect	Large effect
Study 1	Agent	0.17	0.20	0.28	0.93	0.99
(*N* = 217)	Rescued	0.28	0.20	0.28	0.93	0.99
Study 2a	Agent	0.13	0.26	0.17	0.76	0.99
(*N* = 186)	Rescued	0.45	0.24	0.22	0.84	0.99
Study 2b	Agent	0.14	0.20	0.27	0.95	0.99
(*N* = 266)	Rescued	0.36	0.18	0.35	0.97	0.99
Study 2c	Agent	0.03	0.15	0.42	0.99	0.99
(*N* = 446)	Rescued	0.92	0.14	0.52	0.99	0.99
	SES	0.03	0.14	0.52	0.99	0.99
Study 4a	Agent	0.08	0.07	0.98	0.99	0.99
(*N* = 912)	Rescued	0.13	0.07	0.98	0.99	0.99
	Culpable	0.03	0.07	0.98	0.99	0.99
Study 4b	Agent	0.12	0.07	0.97	0.99	0.99
(*N* = 811)	Rescued	0.24	0.07	0.97	0.99	0.99
	Culpable	0.03	0.07	0.97	0.99	0.99
Study 5	Agent	0.10	0.07	0.99	0.99	0.99
(*N* = 915)	Rescued	0.44	0.07	0.99	0.99	0.99
	Culpable	0.08	0.07	0.99	0.99	0.99

Note: Effect sizes are Cohen's *f*s. Effect size benchmarks are: small effect $f = 0.10$, medium effect $f = 0.25$, and large effect $f = 0.40$.

mailing lists. In a 2 (lifeguard: human, robot) × 2 (rescue decision: fisherman, motorboaters) between-subjects design, participants were randomly assigned to one of four conditions. At the time, sample size was determined based on the APA 2012 recommendation of at least 30 participants per cell plus a 10% margin for exclusions (see also Van-Voorhis & Morgan, 2007). See Table 1 for the results of a sensitivity power analysis on the main effects and Appendix F in Supplementary Materials for a simulation-based sensitivity power analysis on the interaction effects of Study 1.

Participants had a mean age of 26.96 (SD = 7.89; Range = 19–65). Of the participants, 159 were women, 36 men, 8 others, and 13 preferred not to state their gender. Altogether 53% of the participants had a Bachelor's degree or higher, and 82% self-reported having an income that was less than the average income in their country.[2]

2.1.2 | Procedure

Participants filled a questionnaire on the university survey platform. After reading an informed consent information screen and giving their consent, the participants gave their responses to exploratory individual differences measures. Then they proceeded to read the vignette, where either a human lifeguard or a rescue robot needs to make a decision whether to rescue two people who caused a boating accident or one innocent person. After reading the vignette, the participants gave their responses to the dependent variables. Finally, they provided their demographics, were debriefed and thanked. As compensation, participants were invited to participate in a raffle with a total of five movie vouchers each of a value of ~10€.

2.1.3 | Materials

Vignette

Our vignette took place in a future where robots are used in many domains including rescue operations. During a cold autumn day, a rescue robot or a human lifeguard witnesses a situation where two drunken people in a motorboat crash into a fisherman's boat. The fisherman and the two motorboaters fall in the water and drift some distance apart. The agent will not have time to rescue both the fisherman *and* the motorboaters, so the agent needs to decide which party to rescue. The agent then rescues either the fisherman or the two motorboaters. See Appendix A in Supplementary Materials for the base form of the vignette.

Dependent variable: Moral approval of the decision

Participants evaluated the agent's decision by indicating their agreement or disagreement with five items: (1) 'The decision is morally acceptable'; (2) 'The decision is morally wrong'; (3) 'The decision is

[2] Income was self-reported on a 10-point ladder scale in relation to average income in the participants' home country. The options were 'I belong to the wealthiest 5%', 'I belong to the wealthiest 10%', 'I belong to the wealthiest 20%', 'I belong to the wealthiest 35%', 'My income is slightly above average', 'My income is average', 'My income is almost average', 'I belong to the poorest 35%', 'I belong to the poorest 20%', and 'I belong to the poorest 10%'. In 2017, when the data were collected, average annual wages in the country were $45,339 (OECD, 2022).

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 2 Full statistics of two- and three-Way ANOVAs for Studies 1–2c

Study number	Factor name	*F*	*p*	η^2_p
Study 1	Agent	6.63	.010*	.028
	Decision	14.95	<.001***	.071
	Agent × Decision	4.24	.040*	.020
Study 2a	Agent	1.64	.196	.016
	Decision	35.53	<.001***	.167
	Agent × Decision	3.21	.042*	.034
Study 2b	Agent	2.54	.080	.018
	Decision	34.67	<.001***	.114
	Agent × Decision	3.30	.038*	.025
Study 2c	Agent	0.23	.797	.001
	Decision	364.21	<.001***	.459
	SES	0.44	.509	.001
	Agent × Decision	1.11	.329	.006
	Agent × SES	1.19	.305	.005
	Decision × SES	9.48	.002**	.021
	Agent × Decision × SES	0.25	.778	.001

Note: DV: moral approval (mean score of five items); Agent: human vs. robot in Study 1; human vs. simple drone vs. smart android in Studies 2a and 2c; human vs. smart drone vs. simple android in Study 2b; Decision: rescue innocent fisherman vs. rescue culpable motorboaters in Studies 1–2b; rescue innocent fisherman vs. rescue culpable motorboater in Study 2c; SES: motorboater's SES is high vs. low. * = significant at $p < .05$; ** = significant at $p < .01$; *** = significant at $p < .001$.

unethical'; (4) 'The decision should not be considered morally right'; (5) 'The decision is irresponsible'. Items were assessed on 7-point Likert scales (anchored at 1: 'Totally disagree' and 7: 'Totally agree') and were averaged, after reverse-scoring items 2 through 5, to form an index of moral approval (Cronbach's $\alpha = 0.86$). These items were chosen from an original list of 15 items based on face validity and exploratory factor analysis; see Appendix C in Supplementary Materials for full list of items. Higher scores indicated greater moral approval of the rescue agent's decision.

2.2 | Results and discussion of Study 1

A two-way ANOVA revealed main effects of both rescue agent (human vs. rescue robot) and decision (rescue the fisherman vs. the motorboaters), which were qualified by a significant two-way interaction [$F(1, 212) = 4.24$, $p = .041$, $\eta^2_p = 0.028$] (see Table 2 for full statistics). Specifically, as shown in Figure 1, the decision to rescue the drunken motorboaters was deemed worse than rescuing the innocent fisherman, especially for the robot. Contrast analyses confirmed that for the robot, a decision to rescue the fisherman was approved more than a decision to rescue the motorboaters ($B = 1.01$, 95% CI = [0.55; 1.46], $t(212) = 4.34$, $p < .001$, $d = 0.60$), but for the human agent, there was no significant difference ($B = 0.31$, 95% CI = [−0.18; 0.80], $t(212) = 1.24$, $p = .217$, $d = 0.17$). Thus, it seems that our participants specifically preferred for the robot to rescue the innocent party rather than make the utilitarian decision.

Previous research (Malle et al., 2015; Voiklis et al., 2016) has emphasized people's expectation for robots to enforce utilitarian moral standards in sacrificial dilemmas. In Study 1, utilizing a different dilemma design than in most prior studies on utilitarian intuitions, we found that participants preferred it when a robot made the non-utilitarian decision of rescuing one innocent person rather than two culpable people. Due to differences between Study 1 and prior research, it is not obvious what aspect of the study caused people to judge the robot's decisions differently from the human's, and differently from prior studies. In Studies 2a, 2b, and 2c, we conducted a series of robustness checks, examining whether features of the robot's appearance and mind (Studies 2a–2b) or of the victims' social identity (Study 2c) drove the discrepancy in judgments.

3 | STUDY 2A

In Study 1, we did not describe the physical or mental properties of the robot agent. In this study, we intended to test whether participants' assumptions about the robot's mental qualities or appearance may have influenced their moral judgments (Gamez et al., 2020; Laakasuo et al., 2021a; Waytz et al., 2014). That is, participants may judge a robot's decisions differently based on whether they imagine the robot as human-like or not, which we did not control for in Study 1. In Study 2a, we split the robot agent condition into two separate conditions with different descriptions, a floater drone and an android, resulting in a total of three agent conditions. The floater drone was described

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

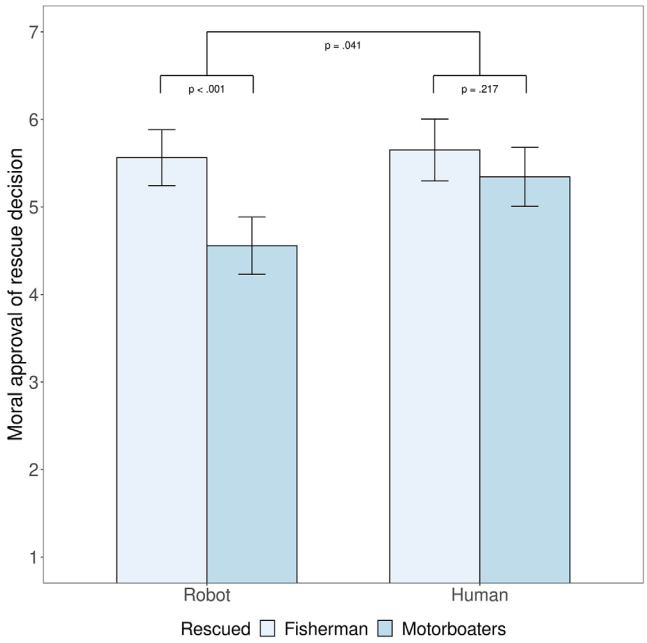

FIGURE 1 Estimated marginal means for the rated approval of decisions in Study 1. The robot agent, but not the human agent, was judged more harshly for saving the two motorboaters culpable for the accident. Error bars are 95% CIs. The dependent variable is a mean score of five items.

as lacking human-like independent thought, only acting based on its programming, while the android was described as having human-like independent thought. Since we were not certain whether the human-likeness of the robot's described appearance or the human-likeness of the robot's stated cognitive capacities would be a more important factor, we decided to manipulate both. Our intention here was to create conditions that described a robot that was clearly human-like or clearly not so. We aimed to investigate whether the human-likeness of the robot would affect participants' moral judgments of the robot's decisions. In addition, we wanted to observe potential differences between moral judgements of the robots' (floater drone and android) and the human lifeguard's decisions. The results of this study indicated that the least morally approved situation was when a human-like robot rescued the culpable party.

3.1 | Method

3.1.1 | Participants and design

This study was conducted in a laboratory environment. In total, 213 adult participants proficient in Finnish were non-intrusively recruited from a large public library in the city centre of Helsinki (details below). Participants were automatically randomized into one of six conditions in a 3 (agent: android, floater drone, human) × 2 (rescue decision: fisherman, motorboaters) between-subjects design. At the time, sample size was determined based on the APA 2012 recommendation of at least 30 participants per cell plus a 10% margin for exclusions (see also VanVoorhis & Morgan, 2007). The sample size was also constrained by practical and economic concerns stemming from running a physical laboratory and coordinating a large number of participants. See Table 1 for the results of a sensitivity power analysis on the main effects and Appendix F in Supplementary Materials for a simulation-based sensitivity power analysis on the interaction effects of Study 2a.

Participants who failed one or more attention checks or control questions were excluded. The final sample size was 186 participants, still within the parameters of at least 30 participants per cell, with mean age 31.42 (SD = 10.13; Range = 18–67). Of these, 103 were women and 83 were men; 67% had a Bachelor's degree or higher, and 76% reported having an income that was less than the average income in their country.[3]

3.1.2 | Procedure

We collected the data at a large public university library in the city centre of Helsinki. We recruited our participants non-invasively by having a table in the foyer with a sign stating: 'Participate in Psychological Research'. All recruited participants approached our research assistants voluntarily. They were informed they could participate in a psychological experiment which would take about 45 min of their time (data were simultaneously gathered for two other studies). They were given informed consent forms, after which the participants who consented to the study were guided to our laboratory space.

The laboratory had four notebook computers with 15″ screens positioned to guarantee maximum privacy. Participants were instructed to use headphones playing pink noise at a pleasant level to cover up any background noise. The experiment was programmed using Python on top of Pygame version 1.96.

Both the experimenters and participants were blind to the randomization. Participants started the experiment by filling in exploratory individual differences measures and then continued to the actual task and dependent variables. Finally, they were compensated, debriefed, thanked, and dismissed. Each participant was compensated with two 5€ (a total of 10€) vouchers.

3.1.3 | Materials

Vignette

The base of the vignette was similar to the one used in Study 1. We altered the vignette in a way that allowed us to observe to what extent the human-likeness and perceived mindedness of the robot might influence moral judgments. In addition to the human agent, the vignette had two different robot agent conditions: a floater drone and an android.

[3] Income was self-reported on a 10-point ladder scale in relation to average income in the participants' home country. The options were 'I belong to the wealthiest 5%', 'I belong to the wealthiest 10%', 'I belong to the wealthiest 20%', 'I belong to the wealthiest 33%', 'My income is slightly above average', 'My income is average', 'My income is almost average', 'I belong to the poorest 33%', 'I belong to the poorest 20%', and 'I belong to the poorest 10%'. In 2019, when the data were collected, average annual wages in the country were $46,249 (OECD, 2022).

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

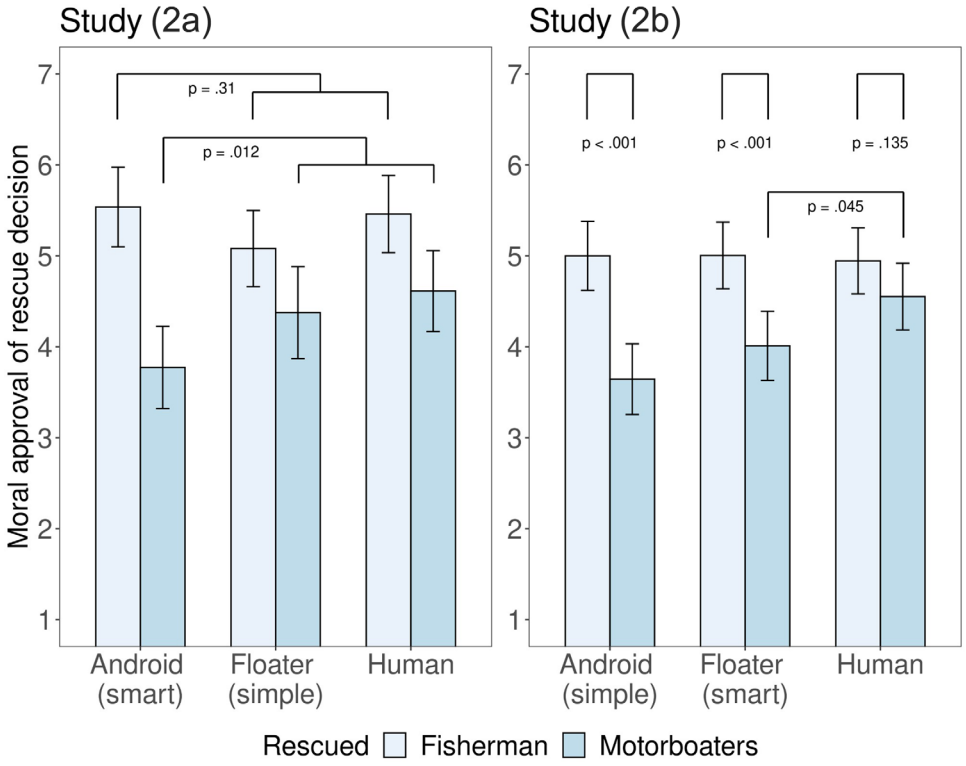

FIGURE 2 Estimated marginal means for Studies 2a and 2b. The android agent was judged significantly more harshly for saving the culpable party (the motorboaters), regardless of whether it was described as having human-like intelligence or not. The floater drone agent was judged more harshly only when it was described as intelligent. Error bars are 95% CIs. The dependent variable is a mean score of five items.

The drone was described as only operating based on its programming, and not having a human-like capacity for independent decision-making (simple drone). The android, on the other hand, was described as having a human-like independent decision-making ability and as functioning based on its own abilities (smart android). See Appendix A in Supplementary Materials for the vignette.

Dependent variable: Moral approval of the decision

The dependent variable was the same five-item measure (Cronbach's $\alpha = 0.89$) used in Study 1.

3.2 | Results and discussion of Study 2a

We ran a full factorial two-way ANOVA on moral approval by entering both experimental factors into the model as predictors (full statistics in Table 2). As in Study 1, there was a significant two-way interaction between agent and rescue decision [$F(1, 180) = 3.21$, $p = .042$, $\eta^2_p = 0.034$]. The pattern of means suggested that the android robot's decision to rescue the culpable party was the least approved decision by our participants (see Figure 2).

Contrast analysis revealed no significant difference between the android and the other two agents when the innocent fisherman was rescued ($B = 0.27$, 95% CI = [−0.26, 0.80], $t(180) = 1.00$, $p = .32$, $d = 0.15$), but rescuing the culpable motorboaters was significantly less approved for the android compared to the other two agents ($B = -0.72$; 95% CI = [−1.28, 0.16]; $t(180) = -2.52$; $p = .012$, $d = -0.37$). The difference between these two contrast comparisons was itself statistically significant ($B = 0.98$; 95% CI = [0.22, 1.75]; $t(180) = 2.52$; $p = .012$, $d = 0.38$), indicating that the android agent was specifically judged more negatively for making the utilitarian choice of rescuing the culpable party.

In sum, the results suggest that the preference for a non-utilitarian decision by the robot agent in Study 1 may have stemmed from participants assuming some level of independence or human-likeness in the robot, as we did not observe this effect for the non-humanoid floater drone. However, it is still unclear why the effect we observe is in this direction. In Studies 1 and 2a, participants judged a utilitarian decision by a robot as less morally acceptable than the alternative, which contradicts, for example, Malle et al. (2015). Thus, the results may be explainable by a preference for robots to abide by a folk sense of justice rather than any preferences about utilitarianism per se. Since our vignette does not contrast the utilitarian action with a deontological *inaction* (as in Malle et al., 2015), but an *action* that prioritizes innocent lives, differences in judgment may stem more from the culpability of the rescued party.

However, we also needed to rule out competing factors relating to the description of the robot, and we designed the next study for this purpose. In short: was the android robot judged differently due to its

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

stated human-like cognitive capacities, or due to its stated human-like appearance (Laakasuo et al., 2021a)?

4 | STUDY 2B

This study was almost identical to Study 2a, except that we switched the cognitive descriptions of the drone robot and the android robot, in order to test whether the unique preferences for the android agent in Study 2a stemmed from its stated cognitive abilities or its stated human-like appearance. Our intention was to see if mind perception was one of the main explanatory factors behind our previous results that implied that the decisions of robots with human-like cognitive skills were specifically judged more negatively.

4.1 | Method

4.1.1 | Participants and design

This study was conducted in a laboratory environment. In total, 266 adult participants proficient in Finnish were non-intrusively recruited from a large public library in the city centre of Helsinki. Participants were automatically randomized into one of six conditions in a 3 (agent: floater drone, android, human) × 2 (rescue decision: fisherman, motorboaters) between-subjects design. The sample size rationale for this study was based on the heuristic that a little over 5% of explained variance, translated to a Cohen's *f* of about 0.235, would be a realistic effect size for interactions. This heuristic was based on our prior experience with similar off-line studies in moral psychology. According to G*Power calculations for a 2 × 3 between-subjects design, approximately 240 participants would be needed for 90% power to detect effects of this size. The sample size was also constrained by practical and economic concerns stemming from running a physical laboratory and coordinating a large number of participants. See Table 1 for the results of a sensitivity power analysis on the main effects and Appendix F in Supplementary Materials for a simulation-based sensitivity power analysis on the interaction effects of Study 2b.

The participants had a mean age of 31.58 (SD = 10.95; range = [18, 72]). Of these, 158 were women, 97 men, and 11 preferred not to state their gender; 73% had a Bachelor's degree or higher; and 72% reported having an income that was less than the average income in their country.[4]

4.1.2 | Procedure

The procedure (including recruitment) was identical with Study 2a. Participants were informed that they could participate in a psychological experiment which would take about 30 min of their time (data were simultaneously gathered for one other study). Each participant was compensated with one voucher worth 5€.

4.1.3 | Materials

Vignette

In this study, the android robot was described as only operating based on its programming, and not having a human-like capacity for independent decision-making (simple android). The floater drone, on the other hand, was described as having a human-like independent decision-making ability and as functioning based on its own abilities (smart drone). Otherwise, the vignette was identical to that of Study 2a. See Appendix A in Supplementary Materials for the vignette.

Dependent variable: Moral approval of the decision

The dependent variable was the same as in previous studies (Cronbach's $\alpha = 0.91$).

4.2 | Results and discussion of Study 2b

We ran a full factorial two-way ANOVA on moral approval by entering both experimental factors into the model (full statistics in Table 2). Again, there was a significant two-way interaction between the experimental factors [$F(2, 260) = 3.30$, $p = .038$, $\eta^2_p = 0.025$]. The mean pattern (see Figure 2) suggested that rescuing the culpable party was approved less when the agent was a robot of any kind than when the agent was a human.

The simple effect of rescue decision was significant for the android agent ($B = 1.36$, 95% CI = [0.81; 1.90], $t(260) = 4.92$, $p < .001$, $d = 0.61$) and the floater drone agent ($B = 0.99$, 95% CI = [0.46; 1.52], $t(260) = 3.70$, $p < .001$, $d = 0.46$), but not for the human agent ($B = -0.39$, 95% CI = [−0.12; 0.91], $t(260) = -1.50$, $p = .135$, $d = -0.19$). This replicated the result of the previous studies, indicating that robot agents' decisions to rescue the culpable were judged more negatively than decisions to rescue the innocent despite a utilitarian motivation, but the same was not true for the decisions of human agents. One further contrast analysis revealed that there was a significant difference in approval for rescuing the motorboaters between the human agent and the drone agent, with the drone agent's decision being judged more negatively ($B = -0.53$, 95% CI = [−1.07; −0.01], $t(260) = -2.03$, $p = .044$, $d = -0.25$). Given that approval for this decision was higher for the drone than the android agent, it follows that the decision to rescue the motorboaters was judged more negatively when the agent was either kind of robot than when the agent was a human.

Based on the results, it seems that the cognitive capacities of the robot agent are not the only thing that affects judgment of their decisions. In Study 2a, the android agent was described as capable of independent thinking, and its utilitarian decision was judged more negatively than the non-utilitarian one, whereas the human agent's

[4] Income was self-reported on an 8-point ladder scale in relation to average income in the participants' home country. The options were 'About the lowest 5%', 'About the lowest 5%–20%', 'Slightly below average', 'Average', 'Slightly above average', 'About the highest 20%–5%', 'About the highest 5%–1%', and 'About the highest 1%'. In 2019, when the data were collected, average annual wages in the country were $46,249 (OECD, 2022).

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

decisions were judged as equally approvable. There was no such effect for the floater drone agent in Study 2a, where the floater drone was described as a mere automaton. However, in Study 2b, when the floater drone was described as capable of independent thinking and the android as a mere automaton, both robot agents' utilitarian decisions were judged more negatively than a human's utilitarian decision. Thus, there may be an effect of both the stated cognitive capabilities of the robot and the physical appearance of the robot (see Bigman et al., 2019; Laakasuo et al., 2021a; Złotowski et al., 2015).

Before moving to examine culpability more closely, we examined one more factor: whether the implicit wealth of the boaters may have played a role. Participants may have seen the fisherman as more deserving to be rescued due to his 'underdog' socio-economic status, compared to the implicitly rich motorboat owners, in the story. In Study 2c, we examined whether a robot rescuing the culpable party was judged more harshly due to the party's perceived wealth specifically.

5 | STUDY 2C

In Study 2c, we reduced the number of motorboaters to one. Thus, the number of lives was matched across rescue decisions—enabling us to assess the relevance of culpability in isolation. Additionally, we manipulated the motorboater's socio-economic status (SES; high vs. low) to understand whether participants' preference for rescuing the fisherman was driven partly by wealth-related stereotypes. We kept all the other materials constant with respect to Studies 2a–2b. The results suggested a minor effect of socio-economic status, a major effect of culpability, and again, more negative judgments towards the decisions of robots than those of humans.

5.1 | Method

5.1.1 | Participants and design

We recruited 503 adult participants through the online crowdsourcing platform Prolific Academic (www.prolific.co). The questionnaire was administered online, using Qualtrics XM (www.qualtrics.com) in English. Participation was restricted to participants who were nationals of the US, the UK, Ireland, Canada, Australia, or New Zealand, spoke English as their first language and were fluent in it, and reported having no long-term health conditions or disabilities. Participants were randomized into one of twelve conditions in a 3 (agent: human, android, floater drone) × 2 (rescue decision: fisherman, motorboater) × 2 (SES: high, low) between-subjects design. The sample size was determined by funding constraints for the study. However, even after exclusions, the final sample size matched the heuristic of at least 30 participants per cell. Furthermore, the same power calculations were applied as used in Study 2b, but since there was an extra factor with two levels, we doubled the sample size (the power calculation tools we used at the time did not support three-way ANOVAs). See Table 1 for the results of a sensitivity power analysis on the main effects and Appendix F in Supplementary Materials for a simulation-based sensitivity power analysis on the interaction effects of Study 2c.

We excluded 57 participants who failed two or more out of three attention checks, one or more out of two post-test manipulation checks, and/or reported having less than native fluency in English. The final sample size was 446 with a mean age of 38.65 (SD = 12.23; range = 18–74). Of these, 278 respondents were women and 168 men; 54% had at least a Bachelor's degree; and 26% reported having an income that was less than the average income in their country.[5]

5.1.2 | Procedure

Participants read a brief description of the questionnaire on Prolific, provided informed consent, read the vignette, answered some personality and attitude measures (not reported here), and answered questions related to the story, before providing demographic information and reading a debrief. Participants were compensated according to the minimum requirements of the Prolific Academic (£1.70 for a ~20-min experiment).

5.1.3 | Materials

Vignette

The vignette was similar to that of Study 2a with the following changes: we dropped the number of motorboaters to one and we added a description of the motorboater as appearing to be of 'high' or 'low' status, social class, and income. Thus, the difference between the SES conditions was only a single word. See Appendix A in Supplementary Materials for the vignette.

Dependent variable: Moral approval of the decision

The dependent variable was the same as in the previous studies (Cronbach's $\alpha = 0.89$).

5.2 | Results and discussion of Study 2c

We ran a full factorial three-way ANOVA on moral approval by entering all three of our experimental factors into the model as predictors (full statistics in Table 2). The only significant effects we observed were the main effect of the rescue decision [$F(1, 434) = 364.21$, $p < .001$, $\eta^2_p = 0.459$] and an interaction between SES and the rescue

[5] Income was self-reported on a 6-point ladder scale in relation to average income in the participants' area of residence. The options were 'Top 1%', 'Top 20%', 'Above average', 'About average', 'Below average', and 'Lowest 20%'. Since the sample consisted of participants from several countries, it is not possible to provide a general average income for context. A majority of the participants (84%) were from the United Kingdom. In 2019, when the data were collected, average annual wages in the United Kingdom were $47,937 (OECD, 2022).

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

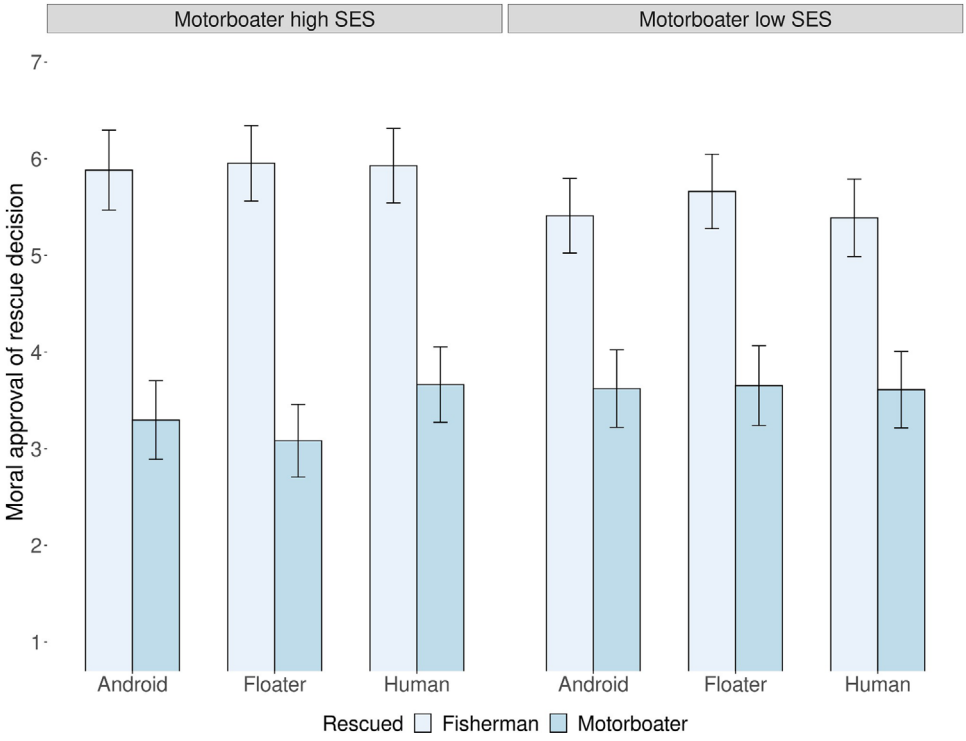

FIGURE 3 Estimated marginal means for the rated approval of decisions in Study 2c. The high SES of the culpable party makes judgments slightly harsher, but the largest effect is that of culpability itself. Error bars are 95% CIs. The dependent variable is a mean score of five items.

decision [$F(1, 434) = 9.48$, $p = .002$, $\eta^2_p = 0.021$]. The three-way interaction between agent, rescue decision and SES was non-significant [$F(2, 434) = 0.25$, $p = .778$, $\eta^2_p = 0.001$]. The pattern of means suggested that the interaction between SES and rescue decision may stem from differences in opposite directions within the two rescue decision conditions.

Contrast analyses showed that when the fisherman was rescued, it was approved more if the motorboater had high SES than when he had low SES ($B = 0.43$, 95% CI = [0.11, 0.75], $t(434) = 2.65$, $p = .008$, $d = 0.26$). When the motorboater was rescued, it was approved less if the motorboater had high SES than when he had low SES, although this effect was non-significant ($B = -0.28$, 95% CI = [−0.60, 0.04], $t(434) = -1.70$, $p = .089$, $d = -0.16$). The difference between these two contrasts was significant ($B = -0.715$, 95% CI = [−1.16, −0.26], $t(434) = -3.07$, $p = .002$, $d = 0.30$).

In sum, it seems that the high SES of the motorboater affected the approval of both rescue decisions, making rescuing the motorboater slightly worse and rescuing the fisherman better in participants' eyes. However, as shown in Figure 3, the size of this effect is small compared to the effect of the motorboater's culpability.

There were no other significant interaction effects. Thus, the results did not replicate Studies 1–2b, where we found an interaction between the rescue decision and agent type. In one further contrast analysis, only in the high SES condition, a human agent's decision to rescue the culpable party was judged more positively than the same decision by robot agents on average ($B = -0.47$, 95% CI = [−0.95, 0.00], $t(434) = -1.94$, $p = .053$, $d = -0.19$), but this effect was not significant by conventional standards. The lack of a stronger agent effect, either as a main effect or as an interaction, may stem from lower power in Study 2c compared to the previous studies, but also from Study 2c having a notably different design. That is, the lack of a utilitarian decision option in the vignette also removed the conflict between two moral concerns. It may be, for example, that a moral decision situation that participants find 'obvious' elicits more uniform judgments of decisions made by different agents (although see Study 5 and the meta-analyses below).

At a broad level, although variation in socio-economic status influenced participants' moral judgment (i.e., rescuing the culpable party was condemned even more when they were of high status), culpability itself seemed to drive the effect on judgment. A preference for rescuing the innocent party emerged whether the culpable party was described as high or low in socio-economic status.

Studies 2a–2c examined plausible alternative explanations of the results of Study 1, implicating the robot's appearance and mind and the culpable party's social standing as factors affecting moral judgments. In the following studies, we turned to evaluate our prospective explanation: the culpability of the accident victims. How much more do participants care about the culpability of the rescued party than utilitarian concerns? We returned to the original setting with two people on a motorboat and one person in a smaller rowboat, but manipulated which (if any) party was responsible for the boat collision.

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 3 Distribution of responses and number of responders per question in Study 3

Response	1 ('not acceptable at all')	2	3	4	5	6	7 ('very acceptable')
Motorboaters drunk							
Rescue fisherman ($N = 19{,}315$)	2.8%	1.4%	1.9%	4.8%	5.6%	11.0%	72.5%
Rescue motorboaters ($N = 19{,}195$)	36.8%	17.4%	12.8%	9.9%	6.2%	4.2%	12.8%
Pure accident							
Rescue fisherman ($N = 18{,}533$)	8.3%	5.2%	8.9%	22.1%	10.4%	8.7%	36.5%
Rescue motorboaters ($N = 18{,}383$)	8.4%	4.0%	5.3	20.1%	11.3%	12.5%	38.6%

Note: The number of participants varies between questions due to the survey being open to the general public and full responses not being forced. Responses to any given question may include responses from participants who did not respond to any or some of the other questions. Numbers based on publicly available response summaries on the survey website (https://yle.fi/aihe/artikkeli/2020/04/28/kumpi-robotin-pitaisi-pelastaa-viaton-kalastaja-vai-juoppokuski-vastaa-ja-katso), accessed 21 January 2021.

6 | STUDY 3

As part of a large-scale public survey, in collaboration with the Yleisradio public service media and broadcasting company in Finland, we included a replication of the robot agent condition in Study 1. We also included a variant of the vignette in Study 1 in which the boat collision is caused by a motor malfunction rather than irresponsible behaviour, thus removing the effect of culpability.

6.1 | Method

6.1.1 | Participants and design

The survey was an online questionnaire open to all visitors to the Yleisradio web page and conducted in Finnish, with the questionnaire opening on 28 April 2020. The survey included other vignettes relating to the theme of morally judging robots or AIs. There was no randomization and no experimental manipulations. Approximately 19,000 people had responded to the survey by 21 January 2021, when we copied the publicly available data for analysis (see Table 3 for more details). There was no specific sample size rationale for this study due to the openness of the survey, and the analyses are based on only the public results available to all visitors on the survey page.

6.1.2 | Procedure

Due to the open nature of the survey, participan could respond to any of the vignettes on the web page in any order and skip responding to any questions. The order of presentation of the vignettes relevant for Study 3 was fixed: the basic form of the vignette (drawn from Study 1) was presented first, followed by an alternative version of the vignette (see below).

6.1.3 | Materials

Vignette

The basic form of the vignette was identical to the robot agent condition of Study 1, excluding the end of the vignette (see Appendix A in Supplementary Materials). Participants were presented with two possible endings (the robot rescues the innocent fisherman or the culpable motorboaters) and gave their responses to the dependent variable (see below). To examine whether the effect was driven by differences in culpability, participants were asked to consider another scenario in which the motorboaters had been acting responsibly. In this alternative scenario, the accident was attributed to a malfunction rather than the boaters' inebriation.

Dependent variable

For both versions of the vignette, participants responded to two questions: 'How acceptable would you find it if the robot decided to rescue the fisherman?' and 'How acceptable would you find it if the robot decided to rescue the motorboaters?'. Participants responded to both questions on a 7-point Likert scale, where 1 denoted 'not acceptable at all' and 7 denoted 'very acceptable'.

6.2 | Results and discussion of Study 3

Due to the open nature of the survey, any visitor on the web page was free to respond to a given question and not reply to another (leading to uneven numbers of responses to the questions), and due to privacy concerns, data with identifiers was not available. Only the public results on the survey web page could be accessed: these included the number of responses for each question and the distribution of responses. Due to this, we could not conduct within-subjects analyses. We opted for comparing conditions using Welch's *t*-tests, essentially treating each data point as an individual participant and each question as an experimental group, coupled with simulations of possible within-subjects data structures. That is, we analysed the data as if the samples were independent, and then simulated analyses using the same data randomly organized into equal-sized paired samples, with 100,000 iterations for each paired-samples test. The paired-samples tests converged on similar results as the independent-samples tests; we report the latter below, and the former can be found in Appendix B in Supplementary Materials.

The results of Study 3 were in line with the previous studies, as seen in Figure 4. When the motorboaters were culpable and the fisherman

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

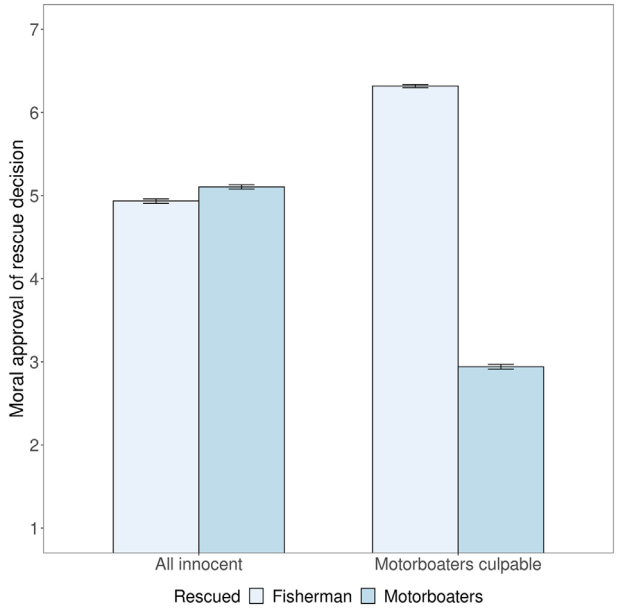

FIGURE 4 Estimated marginal means from Study 3. Error bars represent 95% CIs. Note: sample sizes for each question are different due to the method.

innocent, rescuing the fisherman (M = 6.32, SD = 1.40) was rated much more acceptable than rescuing the motorboaters (M = 2.94, SD = 2.10); Welch's $t(38{,}050) = 197.68$, $p < .001$, $d = 1.89$. When both parties were innocent, there was a noticeably smaller difference in judgments, with the non-utilitarian option of rescuing the fisherman (M = 4.93, SD = 1.98) being rated less acceptable than rescuing the motorboaters (M = 5.10, SD = 1.95); Welch's $t(41{,}864) = -8.85$, $p < .001$, $d = -0.09$. Study 3 provided evidence from a large sample that people strongly prefer a robot agent's decision to rescue a single innocent person over two culpable ones. Additionally, the results of Study 3 showed that even in the absence of differences in culpability, preferences for utilitarian decisions, when comparing rescuing one to rescuing two, are weak, at least when the agent is a robot.

The results of this public online survey inspired the design of Studies 4a–4b and 5, which we pre-registered. The results suggested that our previous results stemmed specifically from people wishing that the lifeguard robot prioritizes folk justice over utilitarian concerns, rather than, for example, our participants disliking the robot making a utilitarian decision per se. In Study 3, in the absence of culpability, the participants had a weak preference for utilitarian decisions. However, the design of Study 3 prevented us from controlling for several factors that we would normally wish to control for in a study, and the fixed order of the questions may have introduced an order effect that could have attenuated utilitarian preferences. Additionally, given the theme of the public survey being specifically moral thinking about robots and AI, we were unable to add questions about human agents. We wished to examine the effects of culpability and utilitarianism in a more controlled survey where participants could be randomized into experimental conditions.

7 | STUDY 4A

We designed this study in order to investigate more nuanced differences between a utilitarian decision (rescuing two versus rescuing one) and a culpability-based decision (rescuing the innocent vs. rescuing the guilty). We manipulated the culpability of the parties as a between-subjects factor so that the culpable party consisted of either one or two people. This way, we could start to dissociate the effect of utilitarian concerns from that of culpability: in one of the conditions, the two moral concerns would conflict, and in the other, they would be aligned. We also changed our dependent variables from multi-item batteries to single-item measures, based on suggestions by Malle et al. (2019). This study was pre-registered online: https://osf.io/5bpxq/.

7.1 | Method

7.1.1 | Participants and design

One thousand (N = 1000) participants were recruited through Prolific Academic. The study was conducted on Qualtrics XM in English, took approximately 13 min, and participants were compensated £1.09 for their time. Participants were randomly assigned to one of four between-subjects conditions in a 2 between (agent: human, android) × 2 between (culpable: fisherman, motorboaters) × 2 within (rescue decision: fisherman, motorboaters) mixed design. The within-subject factor was counterbalanced, such that half of the participants first considered the decision to rescue the fisherman, and then the decision to rescue the motorboaters, while the other half considered these decisions in the opposite order.

As per our pre-registration, we only included participants with complete responses, who indicated having no long-term health problems, and who indicated speaking English at a native-level fluency. Participation was restricted to participants who were resident in and nationals of the US, the UK, Ireland, Canada, Australia, or New Zealand. One participant who indicated being resident in a country from outside of this list was excluded from all analyses. We further excluded participants who failed at least two attention checks, at least one post-test manipulation check, or an inattentive responding test at the end of the experiment.[6] The final sample size after exclusions was 912 participants, close to our target (see power analysis below), with a mean age of 41.22 (SD = 13.12, range = [18, 87]). Of the participants, 481 were women, 430 were men, and one wished to not state their gender. A total of 66.0% of participants had at least some level of university education, and 28.5% reported having an income that was lower than the average income in their country.[7]

[6] The participants were asked two questions: whether they make their own shoes and whether they make their own clothes. Participants who answered 'yes' to both were excluded due to implausibility. We acknowledge that it is not impossible for someone to be both a tailor and a shoemaker, but we posit that in an online questionnaire, a person claiming to be both is most likely a person 'straight-lining' their responses.

[7] Income was self-reported on a 6-point ladder scale in relation to average income in the participants' area of residence. The options were 'Top 1%', 'Top 20%', 'Above average', 'About average', 'Below average', and 'Lowest 20%'. Since the sample consisted of participants from several

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

7.1.2 | Power analysis

To calculate our required sample size, we first ran simulations using several different mean patterns of potential results (based partially on prior results, and partially on speculation, as we had no clear a priori hypotheses about some of the interactions). Based on the simulations, we found that detecting a three-way interaction effect would have either required an unfeasibly large sample size or the assumption of an effect size much larger than what we observed in the prior studies for interactions. Since we were interested in testing hypotheses about the three-way interaction, we decided to compensate for this by decomposing the interaction to specific contrast analyses (see pre-registration). We then calculated our sample size requirements so that we had enough power for these contrast analyses. Specifically, we used the pwr package (Champely, 2020) in R to calculate sample size requirements for detecting medium-sized effects ($d = 0.30$) with 90% power for both paired and independent samples *t*-tests of two conditions (as the design was mixed, with our hypotheses concerning both within-subjects and between-subjects effects). We chose $d = 0.30$ as the benchmark for contrast analyses based on trends in the previous studies. The estimated requirements were, rounding up, 235 participants per group for between-subjects effects, and 119 pairs for within-subjects effects. Thus, we opted to collect 235 participants per group, plus an additional 60 participants to compensate for potential exclusions, leading to a total of 1000 participants. See Table 1 for the results of a sensitivity power analysis on the main effects and Appendix F in Supplementary Materials for a simulation-based sensitivity power analysis on the interaction effects of Study 4a.

7.1.3 | Procedure

Participants provided informed consent, after which they read the instructions for the experiment. They then proceeded to a vignette screen, where they had to stay for 90 s minimum to read the vignette carefully. After the vignette, participants answered three 'gate-keeper' questions about the vignette: whether the lifeguard was a human or an android, which (if any) party was drunk, and whether the lifeguard could rescue both parties. If they failed any of these questions, they were instructed to re-read the vignette and try again.

Having passed the gate-keeper questions, participants first read one hypothetical ending and answered seven single item DV questions about the lifeguard's decision, then repeated this for the second hypothetical ending. After this, they responded to seven exploratory DVs (not reported here; see Appendix D in Supplementary Materials). All Likert questions were presented as 7-pointed Likerts, and participants were also asked to respond to these questions on a 100-point slider scale. In a post-test manipulation check, participants again answered three questions about the vignettes they had read, provided demographic information, read a debriefing section, and clicked a confirmation link to get their compensation via Prolific.

7.1.4 | Materials

Vignette

We adapted the vignette from Study 1 and manipulated the description of how the two parties behaved: either the fisherman or the motorboaters were described as being drunk and steering in an irresponsible manner, leading to a crash. The ending of the vignette was removed, as each participant would read two different hypothetical endings to the vignette after first familiarizing themselves with the scenario. We kept the description of the behaviour of the drunk party as similar as possible between conditions (see pre-registration for the vignette). The floater drone condition was removed, as the most consistent differences in our previous studies were observed between the android and the human condition.

Dependent variables

For each ending to the vignette, participants responded to seven moral judgment questions: how (1) morally wrong, (2) acceptable, (3) permissible, (4) blameworthy, (5) deserving of punishment, (6) deserving of praise, or (7) harmful they found the agent's decision. Responses to each item were made on a 7-point Likert scale. We will only report 1 (moral wrongness) and 2 (acceptability) in the results section, as they most closely match the composite measure used in the previous studies, and most of our hypotheses concerned these two variables (see Appendix B in Supplementary Materials for analyses on the remaining five variables, and Appendices C and D for exact wordings of all reported and unreported variables). The moral wrongness item was 'How morally wrong would you find the [agent's] decision to save [the fisherman/the motorboaters]?' (with 1 denoting 'not morally wrong at all' and 7 denoting 'very much morally wrong'), and the acceptability item was 'How acceptable would you find the [agent's] decision to save [the fisherman/the motorboaters]?' (with 1 denoting 'not acceptable at all' and 7 denoting 'very acceptable').

7.2 | Results and discussion of Study 4a

We ran a full factorial three-way ANOVA for both DVs (acceptability and wrongness), with the agent and the culpability conditions as between-subjects factors, and the rescue decision as a within-subjects (repeated measures) factor (full statistics in Table 4).[8] The pattern of means for acceptability suggested that the least acceptable decision

countries, it is not possible to provide a general average income for context. A majority of the participants (67%) were from the United Kingdom. In 2020, when the data were collected, average annual wages in the United Kingdom were $47,147 (OECD, 2022).

[8] We also conducted an exploratory analysis using the mean of all seven single-item moral judgment measures as the dependent variable. The wrongness, blame, punishment, and harm items were reverse-coded. Cronbach's α was 0.87 for judgments of rescuing the motorboaters and 0.90 for judgments of rescuing the fisherman, both acceptable. The results of the analysis were essentially the same as the results of the analyses reported here.

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 4 Full statistics of three-way ANOVAs for Study 4a

DV	Factor name	*F*	*p*	η^2_p
Acceptability	Agent	11.91	<.001***	0.007
	Decision	29.79	<.001***	0.016
	Culpability	1.89	.168	0.001
	Agent × Decision	7.47	.006**	0.004
	Agent × Culpability	4.59	.032*	0.003
	Decision × Culpability	1289.41	<.001***	0.415
	Agent × Decision × Culpability	5.23	.022*	0.003
Wrongness	Agent	8.63	.003**	0.005
	Decision	14.04	<.001***	0.008
	Culpability	0.13	.719	0.000
	Agent × Decision	6.47	.011*	0.004
	Agent × Culpability	4.27	.039*	0.002
	Decision × Culpability	916.32	<.001***	0.340
	Agent × Decision × Culpability	2.56	.110	0.001

Note: Agent: human vs. android; Decision: rescue fisherman vs. rescue motorboaters; Culpability: fisherman culpable vs. motorboaters culpable. The Decision factor is within-subjects, other factors between-subjects. Satterthwaite's method was used to correct degrees of freedom. * = significant at $p < .05$; ** = significant at $p < .01$; *** = significant at $p < .001$.

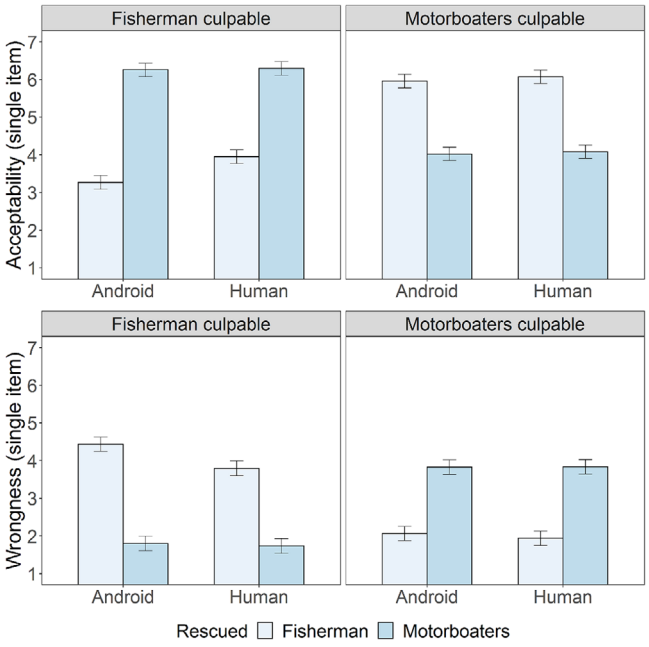

FIGURE 5 Estimated marginal means for the rated wrongness and acceptability of decisions in Study 4a. Utilitarian concerns did not matter for judgments of the human agent's decisions but did matter when judging the robot agent. Error bars are 95% CIs. The dependent variables are single items.

was when the android rescued a culpable fisherman, leaving two innocent motorboaters to die (see Figure 5). This decision makes sense as the 'worst' option, as it conflicts with both utilitarianism and a folk sense of justice. The results were similar for wrongness (see Figure 5). Each two-way interaction was significant for both wrongness and acceptability judgments (see Table 4); however, the three-way interaction between agent, rescue decision, and culpability was significant only for acceptability [$F(1, 1816) = 5.23$, $p = .022$, $\eta^2_p = 0.003$] but not for wrongness [$F(1, 906) = 2.56$, $p = .110$, $\eta^2_p = 0.001$]. These findings are in alignment with our previous studies, where a robot's 'bad' decision was judged more negatively than a 'bad' decision by a human. The results reported below are the results of analyses included in the pre-registration to test our hypotheses.

Contrast comparisons confirmed that rescuing the culpable was seen as worse than rescuing the innocent, regardless of agent: ($B = -2.32$, 95% CI = [−2.44, −2.19], $t(908) = -35.91$, $p < .001$, $d = -2.32$) for acceptability and ($B = 2.09$, 95% CI = [1.95, 2.22], $t(908) = 30.27$, $p < .001$, $d = 2.01$) for wrongness. Further planned contrasts confirmed that rescuing the culpable fisherman was judged as less acceptable for the robot than for the human ($B = -0.69$, 95% CI = [−0.94, −0.43], $t(1816) = -5.29$, $p < .001$, $d = -0.25$). When the decision was to rescue the culpable motorboaters, there was no significant difference ($B = -0.06$, 95% CI = [−0.31, 0.20], $t(1816) = -0.43$, $p = .666$, $d = -0.02$). This pattern repeated for wrongness judgments: that is, relative to the human lifeguard, it was worse for the robot to rescue the culpable fisherman ($B = 0.64$, 95% CI = [0.6, 0.91], $t(1815) = 4.56$, $p < .001$, $d = 0.21$), but no worse to rescue the culpable motorboaters ($B = 0.00$, 95% CI = [−0.28, 0.27], $t(1874) = -0.03$, $p = .978$, $d = 0.00$). Thus, the significant interactions we observed (see Table 4) were driven by the condition where the robot acted against *both* the utilitarian standard of saving two *and* the folk justice standard of saving the innocent.

In sum, our participants once again selectively judged specific decisions made by robot agents more negatively than the same decisions made by human agents. The hypothesis that participants would judge

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

the android's decision of rescuing *any* culpable party more negatively than the same decision by a human was only partially supported. Rather, participants selectively judged the android's decision more negatively than the human's when the decision violated both utilitarianism and a folk sense of justice. When judging human lifeguards' decisions, participants preferred it when the lifeguard saved innocent lives, regardless of their *number*. When evaluating a robot's decision, participants appeared to consider both the number of lives as well as their innocence or culpability. Still, in general, the effect of culpability was substantially larger than that of utilitarianism, in line with Study 3.

8 | STUDY 4B

This study was a replication of Study 4a conducted in the Czech Republic and in the Czech language.

8.1 | Method

8.1.1 | Participants and design

We recruited 958 participants through the Pokusní králíci volunteer participant pool website (https://pokusnikralici.cz/) and its associated Facebook page (https://www.facebook.com/pokusnikralici/). Participants were directed to a Czech language translation of Study 4a on Qualtrics through the website. The procedure was identical to Study 4a except for participation reward, as the Pokusní králíci participant pool is completely volunteer-based and the website does not offer monetary rewards to participants.

The final sample size after exclusions (using the same criteria as in Study 4a, except for the country of residence and English fluency) was 811, which was below our target (see power analysis for Study 4a), but given high effect sizes in Study 4a, we proceeded with the analysis. See Table 1 for the results of a sensitivity power analysis on the main effects of Study 4b. Of the participants, 496 were women, 305 men, 1 non-binary, 5 agender, 3 preferred not to state their gender, and 1 stated they were not sure. The mean age of the participants was 38.20 (SD = 11.80, range = [18, 80]). Altogether 64% had at least some level of university education, and 22.6% reported having an income that was lower than the average income in their country.[9]

8.1.2 | Power analysis

The sample size rationale was identical to that of Study 4a. See Table 1 for the results of a sensitivity power analysis on the main effects and Appendix F in Supplementary Materials for a simulation-based sensitivity power analysis on the interaction effects of Study 4b.

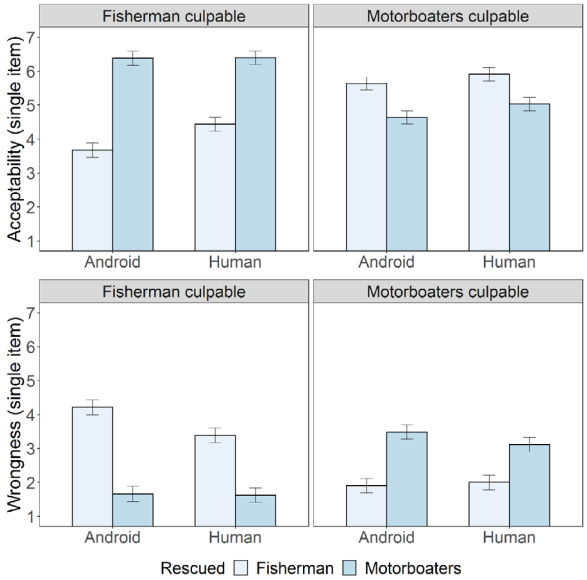

FIGURE 6 Estimated marginal means for the rated wrongness and acceptability of decisions in Study 4b. Utilitarian concerns did not matter for judgments of the human agent's decisions but did matter when judging the robot agent. Error bars are 95% CIs. The dependent variables are single items.

8.1.3 | Procedure

The procedure was identical to that of Study 4a.

8.1.4 | Materials

The materials were identical to those of Study 4a, except for a minor change in the vignette setting (describing the event happening on a lake rather than at sea—this change was made due to the Czech Republic being a landlocked country, making a scenario at sea arguably less familiar or believable).

8.2 | Results and discussion of Study 4b

We used the same analytical procedure and tested the same specific contrasts as in Study 4a, as Study 4b was a direct replication.[10] We replicated the general pattern of the results; see Figure 6 and Table 5 for results of the three-way ANOVAs. Here, the two-way interaction between agent and culpability, which was significant for both

[9] Income was self-reported on a 6-point ladder scale in relation to average income in the participants' area of residence. The options were 'Top 1%', 'Top 20%', 'Above average', 'About average', 'Below average', and 'Lowest 20%'. The data were collected in 2021; the latest available OECD statistics are from 2020. Average annual wages in the country in 2020 were $29,885 (OECD, 2022).

[10] We also conducted an exploratory analysis using the mean of all 7 single-item moral judgment measures as the dependent variable. The wrongness, blame, punishment, and harm items were reverse-coded. Cronbach's α was 0.77 for judgments of rescuing the motorboaters and 0.84 for judgments of rescuing the fisherman, both acceptable. The results of the analysis were essentially the same as the results of the analyses reported here.

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 5 Full statistics of three-way ANOVAs for Study 4b

DV	Factor name	*F*	*p*	η^2_p
Acceptability	Agent	24.83	<.001***	.015
	Decision	92.71	<.001***	.054
	Culpability	1.31	.253	.001
	Agent × Decision	4.69	.030*	.003
	Agent × Culpability	0.12	.729	.000
	Decision × Culpability	510.36	<.001***	.240
	Agent × Decision × Culpability	9.13	.003**	.006
Wrongness	Agent	13.28	<.001***	.008
	Decision	27.23	<.001***	.017
	Culpability	1.58	.208	.001
	Agent × Decision	1.11	.291	.001
	Agent × Culpability	3.60	.058	.002
	Decision × Culpability	507.45	<.001***	.239
	Agent × Decision × Culpability	16.47	<.001***	.010

Note: Agent: human vs. android; Decision: rescue fisherman vs. rescue motorboaters; Culpability: fisherman culpable vs. motorboaters culpable. The Decision factor is within-subjects, other factors between-subjects. Satterthwaite's method was used to correct degrees of freedom. * = significant at $p < .05$; ** = significant at $p < .01$; *** = significant at $p < .001$.

acceptability and wrongness judgments in Study 4a, was non-significant (acceptability: [$F(1, 1614) = 0.12$, $p = .729$, $\eta^2_p = 0.000$]; wrongness: [$F(1, 1614) = 3.60$, $p = .058$, $\eta^2_p = 0.002$]). Additionally, the two-way interaction between agent and decision, which was significant for both DVs in Study 4a, was significant for acceptability [$F(1, 1614) = 4.69$, $p = .030$, $\eta^2_p = 0.003$] but not for wrongness [$F(1, 1614) = 1.11$, $p = .291$, $\eta^2_p = 0.001$]. All other two-way interactions were significant, and in slight contrast to Study 4a, the three-way interaction was significant for both acceptability [$F(1, 1614) = 9.13$, $p = .003$, $\eta^2_p = 0.006$] and wrongness judgments [$F(1, 1614) = 16.47$, $p < .001$, $\eta^2_p = 0.010$].

We replicated the interaction between culpability and the rescue decision. When averaging over agents and the number of people rescued, rescuing the culpable party was judged as less acceptable ($B = -1.63$, 95% CI = [−1.78, −1.49], $t(807) = -22.59$, $p < .001$, $d = -1.59$) and more wrong ($B = 1.75$, 95% CI = [1.60, 1.91], $t(807) = 22.52$, $p < .001$, $d = 1.59$) than rescuing the innocent party.

In Study 4a, we had observed that participants judged the android's decision to rescue the culpable fisherman more negatively than the same decision by the human, but there was no difference in judgments between the agents when they rescued the culpable motorboaters. In Study 4b, participants judged the android agent's decision to rescue the culpable fisherman as less acceptable ($B = -0.76$, 95% CI = [−1.05, −0.47], $t(1614) = -5.19$, $p < .001$, $d = -0.26$) and more wrong ($B = 0.83$, 95% CI = [0.52, 1.14], $t(1614) = 5.25$, $p < .001$, $d = 0.26$) than the same decision by the human. However, they also judged the android's decision to rescue the culpable motorboaters as less acceptable ($B = -0.40$, 95% CI = [−0.68, −0.12], $t(1614) = -2.79$, $p = .005$, $d = -0.14$) and more wrong ($B = 0.37$, 95% CI = [0.07, 0.67], $t(1614) = 2.41$, $p = .016$, $d = 0.12$) than the same decision by the human. This is a deviation from the results of Study 4a. However, note that the effect size is roughly half for the decision to rescue the culpable motorboaters compared to the decision to rescue the culpable fisherman. Thus, the results are in broad terms in alignment with those of Study 4a.

As another deviation from Study 4a, we observed a slightly stronger effect of utilitarian concerns. Namely, the difference in judgments between rescuing the motorboaters or the fisherman was smaller when the motorboaters were culpable than when they were innocent. The android agent rescuing the culpable fisherman was judged as the most wrong and least acceptable scenario. Thus, the android agent's decision was again judged more negatively than the human agent's specifically when the decision conflicted with both utilitarianism and a folk sense of justice.

9 | STUDY 5

In Study 5, we examined whether differences in moral judgments arise when conditioning on culpability: that is, when either both parties involved were responsible for the accident, or when neither party was. This allowed us to evaluate whether the previously documented tendency to hold robots to heightened utilitarian standards arises when the parties are matched in culpability. The study was otherwise identical to Study 4a. The study was pre-registered at: https://osf.io/x5ut9/.

9.1 | Method

9.1.1 | Participants and design

We recruited 1001 participants through Prolific Academic. The study took 13 min (compensation £1.09). Participants provided their responses in English to an online Qualtrics XM questionnaire.

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Participants were randomized into one of four conditions in a 2 between (agent: android, human) × 2 between (culpable: both, neither) × 2 (rescue decision: fisherman, motorboaters) mixed design. The within-subject factor was counterbalanced.

The inclusion and exclusion criteria were identical to those of Study 4a. The final sample size after exclusions was 915, close to our target, with a mean age of 41.54 (SD = 13.88; range = [18, 83]).[11] A total of 499 respondents were women, 414 were men, and 2 were non-binary. Altogether 62.6% of participants had at least some level of university education, and 23.9% reported having an income that was lower than the average income in their country.[12]

9.1.2 | Power analysis

The sample size rationale was similar to that of Study 4a, with approximately 1000 people (a practical, financial limit to data collection) being enough (after estimated 5% exclusion rate) to detect a medium effect size at 90% power for main effects and pairwise comparisons. See Table 1 for the results of a sensitivity power analysis on the main effects and Appendix F in Supplementary Materials for a simulation-based sensitivity power analysis on the interaction effects of Study 5.

9.1.3 | Procedure

The procedure was identical to that of Study 4a, except that the 'gatekeeper' questions and the post-test manipulation checks were changed to match the content of the vignette in Study 5.

9.1.4 | Materials

Vignette

The vignettes were modified from the ones used in Study 4a. The 'all culpable' version of the vignette described the agent noticing that the motorboaters and the fisherman are a drunk and steering in an uncontrolled manner. In this version of the vignette, the boats are about to pass one another, but both parties lose control of their steering, and they crash into one another. The 'all innocent' version of the vignette, in turn, described the agent noticing that the motorboaters and the fisherman are both steering in a responsible manner, and passing each other from a safe distance. However, sudden strong waves make it harder for the boaters to control their boats, and they eventually crash (see pre-registration for the vignette).

Dependent variables

The variables were identical to Study 4a.

9.2 | Results and discussion of Study 5

The analysis procedure was identical to that of Study 4a. We ran a full factorial three-way ANOVA for both of our DVs (acceptability and wrongness), with the agent and the culpability conditions as between-subjects factors and the rescue decision as the within-subjects (repeated measures) factor (full statistics in Table 6).[13] For ANOVA tables and both exploratory and pre-registered hypothesis tests for the other DVs, see Appendix B in Supplementary Materials. The two-way interactions between agent and decision and between decision and culpability were significant for both acceptability and wrongness judgments, whereas the interaction between agent and culpability was not (acceptability: [$F(1, 911) = 1.59$, $p = .207$, $\eta^2_p = 0.001$]; wrongness: [$F(1, 911) = 2.52$, $p = .113$, $\eta^2_p = 0.002$]). Similarly, the three-way interaction was non-significant for both DVs (acceptability: [$F(1, 911) = 0.47$, $p = .492$, $\eta^2_p = 0.000$]; wrongness: [$F(1, 911) = 0.44$, $p = .507$, $\eta^2_p = 0.002$]). Unless otherwise stated, the analyses below are the results of analyses included in the pre-registration to test our hypotheses.

The results supported our prediction that utilitarianism would matter more with no conflicting moral concern related to innocent lives. The results also supported the overall trend from previous studies that robot agents' decisions are selectively judged more negatively than similar decisions by humans when those decisions conflict with either the most salient moral concern or both moral concerns. That is, the 'worst' decision by a robot is worse than the 'worst' decision by a human. However, what turned out to be the 'worst' decision in Study 5 was somewhat surprising. The pattern of means (see Figure 7) suggests that participants judged the situation where the android saved the innocent fisherman as the least acceptable and most wrong decision, contrary to our predictions.

Contrary to our predictions, we did not find evidence for the culpability condition in itself affecting judgments. Thus, the culpability effect observed in prior studies probably depended on the contrast between a culpable and an innocent party depicted in the same scenario. In accordance with our predictions, rescuing the fisherman was judged as significantly more wrong and less acceptable, regardless of culpability or who the agent was. Thus, in the absence of conflicting moral concerns, utilitarian decisions were seen as better overall in Study 5.

[11] The age calculations were conducted with one additional exclusion: a single participant who listed their age as being over 200 years. This participant was not excluded from any other demographic information or analyses reported here, as improbable responses to this demographic question were not a part of our exclusion criteria, and this could plausibly be a simple typing error.

[12] Income was self-reported on a 6-point ladder scale in relation to average income in the participants' area of residence. The options were 'Top 1%', 'Top 20%', 'Above average', 'About average', 'Below average', and 'Lowest 20%'. Since the sample consisted of participants from several countries, it is not possible to provide a general average income for context. A majority of the participants (72%) were from the United Kingdom. In 2020, when the data were collected, average annual wages in the United Kingdom were $47,147 (OECD, 2022).

[13] We also conducted an exploratory analysis using the mean of all 7 single-item moral judgment measures as the dependent variable. The wrongness, blame, punishment, and harm items were reverse-coded. Cronbach's α was 0.79 for judgments of rescuing the motorboaters and 0.85 for judgments of rescuing the fisherman, both acceptable. The results of the analysis were essentially the same as the results of the analyses reported here.

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 6 Full statistics of three-way ANOVAs for Study 5

DV	Factor name	*F*	*p*	η^2_p
Acceptability	Agent	15.81	<.001***	.010
	Decision	311.17	<.001***	.162
	Culpability	0.04	.827	.000
	Agent × Decision	13.43	<.001***	.008
	Agent × Culpability	1.59	.207	.001
	Decision × Culpability	5.53	.019*	.003
	Agent × Decision × Culpability	0.47	.492	.000
Wrongness	Agent	15.30	<.001***	.011
	Decision	260.14	<.001***	.154
	Culpability	0.88	.348	.001
	Agent × Decision	14.99	<.001***	.010
	Agent × Culpability	2.52	.113	.002
	Decision × Culpability	5.39	.020*	.004
	Agent × Decision × Culpability	0.44	.507	.000

Note: Agent: human vs. android; Decision: rescue fisherman vs. rescue motorboaters; Culpability: both parties culpable vs. both parties innocent. The Decision factor is within-subjects, other factors between-subjects. * = significant at $p < .05$; ** = significant at $p < .01$; *** = significant at $p < .001$.

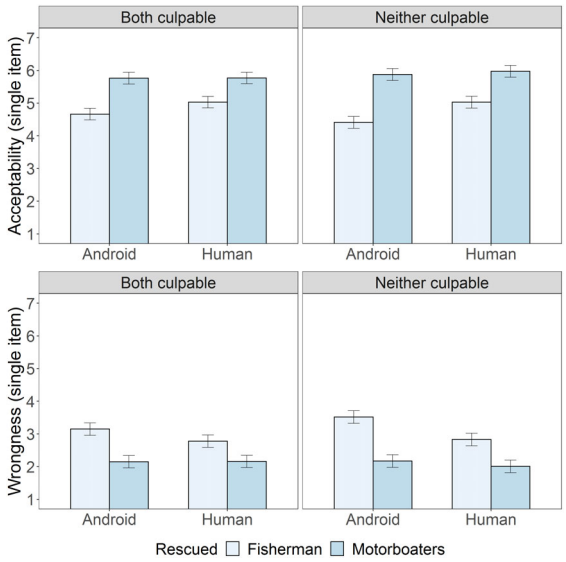

FIGURE 7 Estimated marginal means for rated wrongness and acceptability of decisions in Study 5. The android agent was selectively judged more harshly for the non-utilitarian choice of rescuing the fisherman when both parties were innocent. Error bars represent 95% CIs. The dependent variables are single items.

This result is in contrast to the previous studies, where the conflicting moral concern of culpability was present and seems to have reduced the effect of utilitarian concerns.

The android's decisions were on average judged as more wrong and less acceptable than the human's decisions, again supporting our predictions. Planned contrasts revealed that this effect was driven by the android's *non-utilitarian* decisions. The android's decision to rescue the fisherman was judged as less acceptable ($B = -0.49$, 95% CI = [−0.67, −0.31], $t(1791) = -5.40$, $p < .001$, $d = 0.26$) and more wrong ($B = 0.53$, 95% CI = [0.34, 0.72], $t(1698) = 5.45$, $p < .001$, $d = 0.26$) than the same decision by the human, but the android's decision to rescue the motorboaters was no less acceptable ($B = -0.05$, 95% CI = [−0.23, 0.12], $t(1791) = -0.57$, $p = .566$, $d = -0.03$) nor wrong ($B = 0.07$, 95% CI = [−0.11, 0.26], $t(1698) = 0.78$, $p = .436$, $d = 0.03$).

Exploratory contrast analyses revealed that the android's decision to rescue the innocent fisherman was indeed judged as the most wrong and least acceptable scenario in Study 5. The android's decision to rescue the culpable fisherman was judged more positively than the android's decision to rescue the innocent fisherman: ($B = -0.37$, 95% CI = [−0.64, −0.10], $t(1698) = -2.69$, $p = .007$, $d = -0.13$) for wrongness and ($B = 0.25$, 95% CI = [0.00, 0.50], $t(1791) = 1.97$, $p = .049$, $d = 0.09$) for acceptability. This exploratory result conflicts with our prediction that rescuing the fisherman in the scenario where both parties are culpable would be judged the most negatively (based on the findings of Studies 3–4b). The results still support the general trend that robots' decisions to commit a morally 'bad' action are judged more negatively than similar decisions by humans, but our prediction about what our participants would deem 'bad' was mistaken.

In sum, the results of Study 5 replicate the previously reported finding that robots' non-utilitarian decisions are judged more negatively when a utilitarian decision is possible. When conditioning on culpability, such that either both parties were culpable or neither party was, we observed a clear tendency for robot agents' non-utilitarian decisions to be selectively judged more negatively than similar decisions by humans. This pattern held whether the robot rescued two innocent people (instead of one innocent person) or two culpable people (instead of one culpable person). Together with the results of Studies 3 and 4a–b, the results also indicate that the effect of culpability was greater than that of utilitarian concerns. That is, the difference between judgments

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

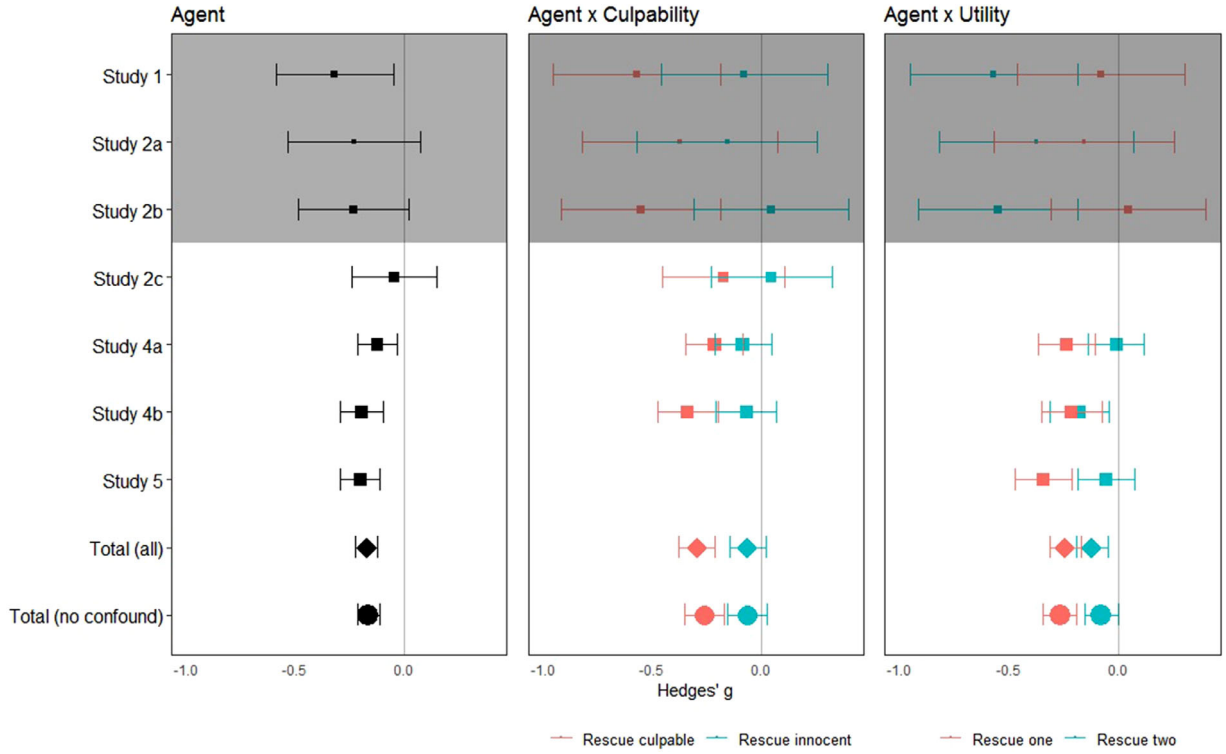

FIGURE 8 Individual and meta-analytic effect sizes. Highlighted area indicates studies with a confound between the number and culpability of the rescued party. Squares and their 95% CI:s represent standardized mean differences (Hedges' g) from individual studies. Diamonds and their 95% CIs represent effect sizes from fixed-effects meta-analyses including all available studies. Circles and their 95% CIs represent effect sizes from fixed-effects meta-analyses excluding Studies 1–2b. Each square, diamond and circles indicates the effect size of a comparison between decisions by human agents and decisions by robot agents. The size of the squares, diamonds and circles indicates group size.

of utilitarian and non-utilitarian decisions is generally smaller than the difference between judgments of decisions to rescue the innocent and decisions to rescue the culpable. Finally, in Study 5, we replicated the finding that a specific decision by a robot was considered the worst option among all possible decisions, namely, rescuing the fisherman when all parties were innocent.

10 | META-ANALYSES

We conducted a series of fixed-effects meta-analyses and a fixed-effects meta-regression with R package metafor (Viechtbauer, 2010), on the combined data of the experimental studies reported here, with the aim of estimating the true magnitude of the observed effect sizes throughout our studies. We chose to conduct fixed-effects analyses (rather than random effects) as the aim of these analyses was primarily to estimate the true effect observed in our studies rather than generalize to the wider population, and the number of studies or individual effects we could include in each meta-analysis was low. We excluded Study 2c from all analyses concerning effects related to the number of people rescued (as Study 2c had no utilitarian decision option), and Study 5 from all analyses concerning effects related to culpability. While Study 5 did contain a culpability manipulation, it deviated from all other studies in that the culpability of both parties was equal in all conditions. For studies with more than one type of robot agent, these were collapsed into a single condition. Analyses were replicated in STATA 15.2 for verification, with similar results to the ones reported here.

First, we found that there was an overall meta-analytic effect of agent ($g = -0.17$, 95% CI = [−0.22; −0.12], $Z = -6.87$, $p < .001$; see Figure 8, left panel, diamonds), indicating that the robot agents' decisions were approved less than the human agents' decisions. Second, we observed a significant difference between the robot and the human agents' decision to rescue the culpable party ($g = -0.29$, 95% CI = [−0.37; −0.21], $Z = -6.94$, $p < .001$; see Figure 8, middle panel, red diamonds), but not the innocent party ($g = -0.06$, 95% CI = [−0.14; 0.02], $Z = -1.47$, $p = .140$; see Figure 8, middle panel, blue diamonds). That is, the robot's decision to rescue the culpable party was judged more negatively than the same decision by a human, but there was no effect when the decision was to rescue the innocent party. Based on a fixed-effects meta-regression with the rescue decision (innocent vs. culpable) entered as a moderator, the difference between these two effect sizes was also significant ($g = 0.23$, 95% CI = [0.11; 0.35], $Z = 3.89$, $p < .001$). Third, there was also a significant meta-analytic effect of agent for both rescuing two people ($g = -0.12$, 95% CI = [−0.19; −0.04], $Z = -3.26$, $p = .001$; see Figure 8, right panel, red diamonds) and rescuing one person ($g = -0.24$, 95% CI = [−0.31; −0.17], $Z = -6.59$, $p < .001$; see Figure 8, right panel, blue diamonds). That

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

TABLE 7 Results of fixed-effects meta-analysis with moderators

	G	SE	*Z*	*P*	CI lower	CI upper
Intercept	0.03	0.03	1.21	.223	−0.02	0.09
Utilitarianism	0.94	0.08	10.55	<.001***	0.76	1.11
Folk justice	−1.22	0.08	−14.17	<.001***	−1.39	−1.05
Agent	−0.15	0.06	−2.55	.010*	−0.28	−0.03
Utilitarianism × Agent	−0.14	0.17	−0.83	.402	−0.49	0.20
Folk justice × Agent	−0.63	0.17	−3.67	<.001***	−0.97	−0.29

Note: Utilitarianism codes for whether dependent variable is an effect size from a contrast comparing rescuing two to rescuing one or not; Folk justice codes for whether it is an effect size from a contrast comparing rescuing the culpable to rescuing the innocent or not; Agent codes for whether effect size is from a contrast comparing a human's or a robot's decisions. The dependent variable is an effect size measure that has been standardized. Predictors are binary and have been centred. Coefficient estimates should be interpreted as deviations from the mean of the overall average effect size of contrasts that were included. * = significant at $p < .05$; *** = significant at $p < .001$.

is, the robot agent was judged more negatively than the human agent for both utilitarian and non-utilitarian decisions, but the difference was larger for non-utilitarian decisions. These two effect sizes again differed significantly in a fixed-effects meta-regression ($g = 0.12$, 95% CI = [0.02; 0.22], $Z = 2.34$, $p = .019$). The results of the above-mentioned analyses were similar whether or not we included Studies 1–2b, where the number of people rescued and their culpability were confounded (Figure 8, compare diamonds and circles). The results were also similar when using random-effects models instead of fixed-effects; see Appendix E in Supplementary Materials.

To estimate the overall effects of utilitarian and folk justice concerns, we conducted a mixed-effects meta-analysis with moderators, in which we controlled for the confound in Studies 1–2b. The results of this analysis are presented in Table 7. We entered effect sizes from utilitarianism effects (contrast between rescuing two and rescuing one) and folk justice effects (contrast between rescuing the culpable and rescuing the innocent) into the model, separately for the two agent types, and from each study. As moderators, we entered a dummy variable coding for agent (0 for human, 1 for robot), and two dummy variables that coded for the type of each effect size (utilitarianism or folk justice). Unconfounded effect sizes from contrasts comparing rescuing two to rescuing one were coded 1 on the utilitarianism dummy and 0 on the folk justice dummy, and vice versa for unconfounded effect sizes from contrasts comparing rescuing the culpable to rescuing the innocent. Confounded effect sizes from Studies 1–2b were coded as 1 for both dummy variables. Finally, we centred the moderator variables and standardized the effect size variables, so that the mean of each was 0.

The absolute effect size of the utilitarianism moderator was smaller than that of the folk justice moderator. There was also a significant interaction between agent and folk justice, so we conducted separate mixed-effects moderator analyses on the two agent groups. The effect of folk justice was larger in the robot group ($g = -1.46$, 95% CI = [−1.62; −1.30], $Z = -17.86$, $p < .001$) than in the human group ($g = -1.15$, 95% CI = [−1.32; −0.99], $Z = -13.69$, $p < .001$), whereas the effect of utilitarianism did not differ between the robot ($g = 0.93$, 95% CI = [0.75; 1.12], $Z = 9.90$, $p < .001$) and human ($g = 0.78$, 95% CI = [0.58; 0.97], $Z = 7.93$, $p < .001$) groups. When using a mixed-effects rather than fixed-effects moderator analysis, the results were similar, though without a significant interaction effect: see Appendix E in Supplementary Materials.

To summarize, decisions made by robot agents were judged overall more negatively than decisions made by human agents (Figure 8, left panel). However, this effect was driven by judgments of specific decisions: norm-abiding decisions by robots and humans were judged more similarly than norm-violating ones (Figure 8, middle & right panels). The decisions of both humans and robots were judged more negatively if they violated rather than abided by a norm: the general direction of moral judgment was similar for the agents. Differences in moral judgments were overall larger when comparing rescuing the culpable to rescuing the innocent than when comparing rescuing two to rescuing one: folk justice concerns had a larger effect than utilitarian concerns. This effect was stronger for the robot agents than for the human agents.

11 | GENERAL DISCUSSION

Our studies have two overall novel findings. First, decisions made by robots were not always judged as worse than the same decisions made by humans. Rather, those decisions were consistently judged as worse when they failed to uphold the most salient moral standard. On the other hand, when the decision aligned with the salient moral standard, differences in judgments between a robot's decision and a human's decision were smaller or non-existent. In other words, a robot's 'bad' decision was worse than a human's 'bad' decision, but a robot's 'good' decision was equal or only slightly worse than a human's 'good' decision. Whether an agent was a human or a robot, moral judgments about their decisions followed a similar trend: rescuing the culpable was worse than rescuing the innocent, and rescuing two was better than rescuing one, all else being equal. Second, the culpability of the rescued party was a stronger indicator of moral judgments about rescue decisions than utilitarian concerns. In general, our participants judged decisions to rescue innocent people more positively than decisions to rescue the culpable, whether this decision conflicted with utilitarianism or not.

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

In the majority of our studies, only specific decisions by robots were judged as worse than similar decisions by humans. A robot's decision to rescue two culpable people rather than one innocent person was judged more negatively than the identical human decision, but there were no differences in judgments when the decision was to rescue the innocent person (Studies 1–2b). A decision violating both utilitarian and folk justice standards was considered worse if the decision was made by a robot rather than a human (Studies 4a–4b). When culpability was held constant between the parties that could be rescued, participants judged non-utilitarian decisions by robots as the most immoral alternative (Study 5). In our meta-analyses, we also found that when comparing judgments of the same decision made by a human and a robot, differences were larger for decisions that violated a norm than decisions that abided by a norm.

The totality of our results suggests that folk justice was a stronger driver of judgments than utilitarianism. In our moderator meta-analysis, where we analysed differences in judgments between two different decisions that either abided by or violated a moral concern, folk justice concerns had a larger effect on the difference than utilitarian concerns. Further, the effect of folk justice concerns was larger for robot than for human agents, whereas the effect of utilitarianism was similar for both agents (though this interaction was not statistically significant when using a mixed-effects model; see Appendix E in Supplementary Materials). In sum, our participants seem to have consistently put more weight on whether an agent rescued innocent or culpable people[14] than on whether an agent rescued as many people as possible, and this was emphasized when the agent was a robot. However, again, the agents were judged in a similar direction: keeping other variables constant, it was better for both robots and humans to rescue the innocent over the culpable and to rescue two over one.

In individual studies, we only observed preferences for utilitarian decisions (by robots or humans) when the utilitarian decision did not conflict with folk justice concerns. In prior literature (e.g., Awad et al., 2018; Hidalgo et al., 2021; Malle et al., 2015, 2016), many findings have indicated a preference for robots to make utilitarian decisions that violated the competing deontological concern of 'do not kill'. In some studies (e.g., Malle et al., 2015; 2016), this preference was stronger for robots than for humans. Why did we not observe a similar effect of utilitarianism over competing moral concerns when robots' decisions were judged? The present results are inconclusive in this regard due to several differences in our vignettes compared to typical utilitarian dilemmas: a lower possible maximum number of lives saved (i.e., two instead of five), the lack of an omission option (i.e., non-utilitarian decisions do not equate to doing nothing), and the added effect of culpability. Thus, it may be that the deontological motivation in typical trolley dilemmas is not strong enough, or that the utilitarian motivation for saving five people over one is stronger than the motivation in the present studies, or that in dilemmas with an omission option, people prefer robots that act rather than stand idly. Our studies highlight that people's judgments of different agents' decisions in moral dilemmas are contextual. Nonetheless, our participants consistently judged robot agent's decisions more negatively for failing to adhere to the most salient moral norm(s).

Our results are broadly aligned with previous studies suggesting that the moral evaluation of a robot decision-maker is moderated by the uncanny valley effect (i.e., the human-likeness of the robot; Grey & Wegner, 2012; Laakasuo et al., 2021a). In other words, decisions made by more human-like robots are judged more negatively than decisions made by machine-like robots. We only used a description of a robot with human-like appearance and intelligence in Studies 4a–5, as the results of Studies 2a–2b indicated that the decisions of humans were more likely to be judged differently from the decisions of robots if the robots were human-like. However, we did not replicate the observation that any decision would be judged more negatively when the agent was human-like (Laakasuo, Palomäki et al., 2021). Again, this may be due to a different moral decision scenario than what is typical for psychological studies on utilitarian judgment.

Robot and human agents or their decisions being judged differently is a common trend in moral psychological studies (Bigman & Gray, 2018; Laakasuo et al., 2021a; Laakasuo et al., 2022; Malle et al., 2015, 2019). Our studies continue in this tradition with a focus on judgments of decisions by different agents (rather than judgments about agents themselves), while ruling out certain alternative explanations. Because our vignettes did not have an omission option, participants' moral judgments of the agents' decisions were not confounded with preferences for or against inactivity (Baron & Ritov, 2004; Gawronski & Beer, 2016; Gawronski et al., 2017). Furthermore, because the scenarios described a lifeguard that acts in response to an event rather than being ordered to do so, the results are not affected by a perceived chain of command (see Laakasuo et al., 2022; Malle et al., 2019).

Our results suggest that the moral decisions of human-like robots are scrutinized more than the moral decisions of humans: 'good' decisions are judged roughly equally, but 'bad' decisions by robots are seen as worse than 'bad' decisions by humans. Our studies cannot give a direct answer as to the reason behind this result, but we speculate that it may stem from differing expectations concerning the actions of robots and humans, based on, for instance, assumptions about differences in cognitive and emotional capacities or social roles. Perhaps robots, especially if they are human-like, are expected to be more 'rational' than humans, and thus when they do something unwanted, the judgment is harsh (as also argued by Komatsu et al., 2021). While robots may be expected to be intelligent, they are generally viewed as less capable of emotion than humans (de Graaf & Malle, 2019; Gamez et al., 2020; Gray et al., 2007; Weisman et al., 2017). Thus, the robot should be able to act intelligently without being hindered by emotion: a life-or-death decision is emotionally taxing for a human, but presumably not for a robot. It could be that our participants could better empathize with a human forced to make a difficult decision than with a

[14] We argue that our results tentatively support the conclusion that our participants' culpability-based judgments were driven more by a wish to protect the innocent, rather than 'punish' the wrongdoers. As stated earlier, both of these motives fit what we termed a folk sense of justice, and we did not specifically ask our participants about their motivations. However, the results of Study 5 work as an indirect test. In Study 5, when both parties were culpable, participants gave more positive judgments for decisions to rescue the party with more people. In other words, it seems that the participants still preferred it when the agent rescued more people, even when the agent could have 'punished' a higher number of culpable people by rescuing the party with just one person.

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

robot in the same situation, and subsequently gave more positive judgments of the human agent's norm-violating decisions. It may also be that judgments about the human agent's rescue decisions are affected by the perceived risk of rescuing a drowning person, while a robot is not seen as risking anything. A robot making an unwanted decision in matters of life and death may also make people question the point of having a robot make the decision in the first place: the justification for using robots is often a reduction in unwanted outcomes. In the context of our vignettes, people may view the function of a robotic rescuer to be to make 'optimal' rescue decisions, and it failing in this task is worse than a human failing in it, as the use of a robot over a human needs additional justification. However, since there is no theory that could properly address these differences between humans and robots, further work is needed (Laakasuo et al., 2021b,c).

In sum, when judging a human-like robot's decisions in a life-or-death situation, people may view it as judging the decisions of an agent that should not have human faults, that does not evoke empathy, and/or that needs to prove its worth as a decision-maker more than a human does. Due to these factors, unwanted decisions by the robot may evoke more negative reactions than unwanted decisions by a human would evoke in the same situation. However, the moral norms that people seem to expect an agent to follow are the same for the robot and for the human, leading to broadly similar judgments of the different agents' decisions.

11.1 | Limitations

Our studies show that culpability had a stronger effect than utilitarian concerns in a 'two versus one' scenario, specifically in the context of lifeguarding. Our results concerning the relative weights of different moral concerns when judging non-human agents may not generalize to other kinds of moral decisions or contexts. More research is needed to investigate whether this effect can be found in scenarios with stronger utilitarian pressures (e.g., 'five vs. one'; Greene et al., 2008), different kinds of culpability, or outside rescue contexts.

In general, our studies also face the usual limitations of experimental studies with self-report measures. First, the sample population is not usually representative in terms of age and other demographics or personality (whether samples are collected online or otherwise). However, we ran studies with different populations and cultures (students, native Finnish speakers, native English speakers, native Czech speakers), which adds robustness to our findings. Second, using self-report measures is associated with participants' tendency to acquiesce in, or obstruct, the experimenter's perceived objectives. We addressed this issue with fully randomized double-blind designs. Moreover, vignette- and questionnaire-based research has been shown to be reliable in studying moral judgments about the decisions of artificial agents (Bigman & Gray, 2018; Laakasuo et al., 2021a; Malle et al., 2015).

Finally, many of our early studies—especially studies conducted in laboratory settings—had relatively small sample sizes for detecting minor interaction effects. However, the results of the later higher-powered pre-registered studies match those of the earlier ones, even with slightly different measures, and our findings are qualified by the meta-analyses.

11.2 | Future studies

Future studies should more generally examine different moral concerns. Although trolley-type dilemmas are common in moral psychology, moral judgment should not be reduced to only utilitarian versus deontological judgment. While there is also little knowledge of how people would weigh different concerns in 'utilitarian calculus', moral psychology should diversify to fill this gap (Chapman, 2018; Voiklis & Malle, 2017). Our studies highlight a potentially under-researched issue: the character perception of moral *patients* rather than moral *agents*. As shown in our studies, the way people perceive the rescued party (i.e., their culpability and their social status) influences moral judgments. Regarding culpability, more research should examine people's motivations for deciding who should be rescued. Additionally, our Studies 2a and 2b, in conjunction with those of Laakasuo et al. (2021a), imply that the stated physical and/or cognitive human-likeness of a robot affects moral judgments of their decisions, which warrants further examination.

Finally, future research is needed to determine which exact subcomponents of moral judgment are of main interest (blame, harm, 'moral acceptance', or causality inferences, to name a few). Our results using a scale measuring general moral approval (Studies 1–2c) and single-item measures of acceptability and wrongness (Studies 3–5) were generally consistent with one another. However, there were some deviations to the patterns we observed when using single-item measures of, for example, perceived harm instead of acceptability or wrongness (see Appendix B in Supplementary Materials). While many different variables we used were interchangeable with one another, there are nuances to different moral judgments, and focusing solely on ratings of wrongness or acceptability may hide interesting patterns in other types of judgment.

12 | CONCLUSION

In sum, our studies—using a novel moral dilemma scenario of a rescue situation—show that human and robot agents' decisions are judged differently primarily when those decisions violate the most salient moral norm in a situation. These differences in judgments are sensitive to both utilitarian concerns and the culpability of the rescued party, something previously underexplored in moral psychology. Moreover, culpability seemed to matter more than strictly utilitarian concerns. People consistently showed a preference for rescuing the innocent over saving a larger number of lives when these options were mutually exclusive. Our studies also addressed the issue of omission or inactivity bias (Gawronski et al., 2017) by having each potential option in the dilemma consist of an active choice (i.e., rescuing one or more people).

We further highlight the need for a focus on different core factors of moral dilemma situations, and for examining moral concerns that

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

are intuitively relevant to many people but that have mostly stayed out of moral psychology due to established methods of asking questions. Especially in the context of judgments about artificial agents making moral choices, it is important to shed light on the various factors that people care about.

All in all, in our studies, we showed that judgments of robots as decision-makers can offer new insights into the questions of human morality and can help elucidate newer research traditions as well. We suggest that moral psychological vignettes incorporate new moral principles and agents to expand and refine the growing understanding of moral cognition.

ACKNOWLEDGMENTS

This research was funded by Jane and Aatos Erkko Foundation (grant number 170112), Academy of Finland (grant number 323207) and Weisell Foundation grants awarded to Michael Laakasuo, who was the principal investigator and conceptualized the research. Marianna Drosinou was additionally funded by Tiina and Antti Herlin Foundation. This research is part of NetResilience consortium funded by the Strategic Research Council within the Academy of Finland (grant number 345186 and 345183). We thank Anke Haas and Noora Lehtonen for help with the original Study 1 design and investigation, and Markus Jokela for help with the meta-analyses.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

DATA AVAILABLITY STATEMENT

The datasets, data codebooks and analysis scripts used are available at https://osf.io/hm5yf/.

Note the following details about the data made available:

Studies 1, 2a, 2b, 2c: Demographic variables such as age are aggregated to make data more anonymous. Covariate measures that may be reported elsewhere are not included. The report does not include data on variables not covered in manuscript.

Study 3: Data with identifiers is not available. Analyses are based on publicly available aggregate numbers, as highlighted in the manuscript.

Studies 4a, 4b, 5: Some demographic variables such as age are aggregated. Some demographic variables collected such as employment are not included in the shared data set due to being irrelevant to analyses and potential identifiers.

ETHICS STATEMENT

The protocols and materials were approved by the University of Helsinki Ethical Review Board in the Humanities and Social Behavioural Sciences. The research followed national ethics guidelines. The results are reported honestly. This is an original article. Authorship reflects individuals' contributions.

ORCID

Jukka Sundvall https://orcid.org/0000-0003-4310-1162
Marianna Drosinou https://orcid.org/0000-0001-9696-6827
Ivar Hannikainen https://orcid.org/0000-0003-0623-357X
Volo Herzon https://orcid.org/0000-0001-7781-1651
Robin Kopecký https://orcid.org/0000-0002-7140-4649
Michaela Jirout Košová https://orcid.org/0000-0002-0441-8426
Mika Koverola https://orcid.org/0000-0001-8227-6120
Anton Kunnari https://orcid.org/0000-0002-2951-6399
Silva Perander https://orcid.org/0000-0001-6711-8079
Teemu Saikkonen https://orcid.org/0000-0001-9619-3270
Jussi Palomäki https://orcid.org/0000-0001-6063-0926
Michael Laakasuo https://orcid.org/0000-0003-2826-6073

REFERENCES

Addison, A., Bartneck, C., & Yogeeswaran, K. (2019). Robots can be more than black and white: Examining racial bias towards robots. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery. (pp. 493–498).

Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). The moral machine experiment. *Nature*, *563*(7729), 59–64. https://doi.org/10.1038/s41586-018-0637-6

Baron, J., & Ritov, I. (2004). Omission bias, individual differences, and normality. *Organizational Behavior and Human Decision Processes*, *94*(2), 74–85. https://doi.org/10.1016/j.obhdp.2004.03.003

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Bigman, Y. E., Waytz, A., Alterovitz, R., & Gray, K. (2019). Holding robots responsible: The elements of machine morality. *Trends in Cognitive Sciences*, *23*(5), 365–368. https://doi.org/10.1016/j.tics.2019.02.008

Campaign to Stop Killer Robots. (2019). *Key elements of a treaty on fully autonomous weapons*.

Carlsmith, K. M., & Darley, J. M. (2008). Psychological aspects of retributive justice. In *Advances in experimental social psychology* (Vol. 40, pp. 193–236). Academic Press. https://doi.org/10.1016/S0065-2601(07)00004-4

Champely, S. (2020). pwr: Basic Functions for Power Analysis. https://CRAN.R-project.org/package=pwr

Chapman, H. A. (2018). A component process model of disgust, anger, and moral judgment. *Atlas of Moral Psychology*, *70*, 70–80.

Christensen, J. F., & Gomila, A. (2012). Moral dilemmas in cognitive neuroscience of moral decision-making: A principled review. *Neuroscience & Biobehavioral Reviews*, *36*(4), 1249–1264. https://doi.org/10.1016/j.neubiorev.2012.02.008

Curry, O. S., Jones Chesters, M., & Van Lissa, C. J. (2019). Mapping morality with a compass: Testing the theory of 'morality-as-cooperation' with a new questionnaire. *Journal of Research in Personality*, *78*, 106–124. https://doi.org/10.1016/j.jrp.2018.10.008

Davison, N. (2017). A legal perspective: Autonomous weapon systems under international humanitarian law. *UNODA Occasional Papers*, *30*, 5–18.

de Graaf, M. M. A., & Malle, B. F. (2019). People's explanations of robot behavior subtly reveal mental state inferences. *2019 14th ACM/IEEE International Conference on Human–Robot Interaction (HRI)* (pp. 239–248). IEEE, Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/HRI.2019.8673308

Gamez, P., Shank, D. B., Arnold, C., & North, M. (2020). Artificial virtue: The machine question and perceptions of moral character in artificial moral agents. *AI & SOCIETY*, *35*(4), 795–809. https://doi.org/10.1007/s00146-020-00977-1

Gawronski, B., Armstrong, J., Conway, P., Friesdorf, R., & Hütter, M. (2017). Consequences, norms, and generalized inaction in moral dilemmas: The CNI model of moral decision-making. *Journal of Personality and Social Psychology*, *113*(3), 343–376. https://doi.org/10.1037/pspa0000086

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Gawronski, B., & Beer, J. S. (2016). What makes moral dilemma judgments "utilitarian" or "deontological"? *Social Neuroscience*, *12*(6), 626–632. https://doi.org/10.1080/17470919.2016.1248787

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619. https://doi.org/10.1126/science.1134475

Gray, K., & Wegner, D. M. (2012). Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition*, *125*(1), 125–130. https://doi.org/10.1016/j.cognition.2012.06.007

Greene, J. D. (2007). Why are VMPFC patients more utilitarian? A dual-process theory of moral judgment explains. *Trends in Cognitive Sciences*, *11*(8), 322–323. https://doi.org/10.1016/j.tics.2007.06.004

Greene, J. D., Morelli, S. A., Lowenberg, K., Nystrom, L. E., & Cohen, J. D. (2008). Cognitive load selectively interferes with utilitarian moral judgment. *Cognition*, *107*(3), 1144–1154. https://doi.org/10.1016/j.cognition.2007.11.004

Haidt, J., & Graham, J. (2007). When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, *20*(1), 98–116. https://doi.org/10.1007/s11211-007-0034-z

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human–robot interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Hidalgo, C. A., Orghian, D., Albo-Canals, J., de Almeida, F., & Martin, N. (2021). *How humans judge machines*. MIT Press.

Kahane, G., Everett, J. A. C., 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. https://doi.org/10.1016/j.cognition.2014.10.005

Komatsu, T., Malle, B. F., & Scheutz, M. (2021). Blaming the reluctant robot: Parallel blame judgments for robots in moral dilemmas across US and Japan. In *Proceedings of the 2021 ACM/IEEE International Conference on Human–Robot Interaction* (pp. 63–72). Association for Computing Machinery. https://doi.org/10.1145/3434073.3444672

Kwon, M., Jung, M. F., & Knepper, R. A. (2016). Human expectations of social robots. In *2016 11th ACM/IEEE International Conference on Human–Robot Interaction (HRI)* (pp. 463–464). IEEE, Institute of Electrical and Electronics Engineers. https://doi.org/10.1109/HRI.2016.7451807

Laakasuo, M., Palomäki, J., & Köbis, N. (2021a). Moral uncanny valley: A robot's appearance moderates how its decisions are judged. *International Journal of Social Robotics*, *13*, 1679–1688. https://doi.org/10.1007/s12369-020-00738-6

Laakasuo, M., Sundvall, J. R., Berg, A., Drosinou, M., Herzon, V., Kunnari, A., Koverola, M., Repo, M., Saikkonen, T., & Palomäki, J. (2021b). Moral psychology and artificial agents (Part one): Ontologically categorizing bio-cultural humans. In S. Thompson (Ed.), *Machine law, ethics, and morality in the age of artificial intelligence* (pp. 166–188). IGI Global. https://doi.org/10.4018/978-1-7998-4894-3.ch010

Laakasuo, M., Sundvall, J. R., Berg, A., Drosinou, M., Herzon, V., Kunnari, A., Koverola, M., Repo, M., Saikkonen, T., & Palomäki, J. (2021c). Moral psychology and artificial agents (Part two): The transhuman connection. In S. Thompson (Ed.), *Machine law, ethics, and morality in the age of artificial intelligence* (pp. 189–204). IGI Global. https://doi.org/10.4018/978-1-7998-4894-3.ch011

Laakasuo, M., Palomäki, J., Kunnari, A., Rauhala, S., Drosinou, M., Halonen, J., Lehtonen, N., Koverola, M., Repo, M., Sundvall, J., Visala, A., & Francis, K. B. (2022). Moral psychology of nursing robots: Exploring the role of robots in dilemmas of patient autonomy. *European Journal of Social Psychology*, *53*, 1–21. https://doi.org/10.1002/ejsp.2890

Malle, B. F. (2021). Moral judgments. *Annual Review of Psychology*, *72*, 293–318. https://doi.org/10.1146/annurev-psych-072220-104358

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In M. I. Aldinhas Ferreira, J. Silva Sequeira, G. Singh Virk, M. O. Tokhi, & E. E. Kadar (Eds.), *Robotics and well-being* (Vol. 95, pp. 111–133). Springer International Publishing. https://doi.org/10.1007/978-3-030-12524-0_11

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human–Robot Interaction*, 117–124. https://doi.org/10.1145/2696454.2696458

Malle, B. F., Scheutz, M., Forlizzi, J., & Voiklis, J. (2016). Which robot am I thinking about? The impact of action and appearance on people's evaluations of a moral robot. *2016 11th ACM/IEEE International Conference on Human–Robot Interaction (HRI)*, 125–132. https://doi.org/10.1109/HRI.2016.7451743

Marcus, G., & Davis, E. (2019). *Rebooting AI: Building artificial intelligence we can trust*. Vintage.

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans*. Penguin.

OECD (2022). *Average wages (indicator)*. (Accessed on 5 June 2022). https://doi.org/10.1787/cc3e1387-en

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

Palomäki, J., Kunnari, A., Drosinou, M., Koverola, M., Lehtonen, N., Halonen, J., Repo, M., & Laakasuo, M. (2018). Evaluating the replicability of the uncanny valley effect. *Heliyon*, *4*(11), e00939. https://doi.org/10.1016/j.heliyon.2018.e00939

Phillips, E., Ullman, D., de Graaf, M. M. A., & Malle, B. F. (2017). What does a robot look like?: A multi-site examination of user expectations about robot appearance. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 1215–1219. https://doi.org/10.1177/1541931213601786

Schein, C., & Gray, K. (2018). The theory of dyadic morality: Reinventing moral judgment by redefining harm. *Personality and Social Psychology Review*, *22*(1), 32–70. https://doi.org/10.1177/1088868317698288

Starr, N. D., Malle, B., & Williams, T. (2021). *I need your advice… human perceptions of robot moral advising behaviors*. arXiv preprint arXiv:2104.06963.

Syrdal, D. S., Dautenhahn, K., Walters, M. L., & Koay, K. L. (2008). Sharing spaces with robots in a home scenario—anthropomorphic attributions and their effect on proxemic expectations and evaluations in a live HRI trial. In *Proceedings of the 2008 AAAI Fall Symposium* (pp. 116–123). AAAI, Association for the Advancement of Artificial Intelligence.

Trovato, G., Lucho, C., & Paredes, R. (2018). She's electric—the influence of body proportions on perceived gender of robots across cultures. *Robotics*, *7*(3), 50. https://doi.org/10.3390/robotics7030050

VanVoorhis, C. R. W., & Morgan, B. L. (2007). Understanding power and rules of thumb for determining sample sizes. *Tutorials in Quantitative Methods for Psychology*, *3*(2), 43–50. https://doi.org/10.20982/tqmp.03.2.p043

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Voiklis, J., Kim, B., Cusimano, C., & Malle, B. F. (2016). Moral judgments of human vs. Robot agents. *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 775–780. https://doi.org/10.1109/ROMAN.2016.7745207

Voiklis, J., & Malle, B. F. (2017). Moral cognition and its basis in social cognition and social regulation. In K. Gray, & J. Graham (Eds.), *Atlas of moral psychology* (pp. 108–120). Guilford Press.

Walker, J. (2019). *Search and rescue robots: Current applications on land, sea, and air*. EMERJ - The AI Research and Advisory Company. https://emerj.com/ai-sector-overviews/search-and-rescue-robots-current-applications/

Wallach, W., & Allen, C., (2008). *Moral machines: teaching robots right from wrong*. Oxford University Press.

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of*

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

Experimental Social Psychology, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

Weisman, K., Dweck, C. S., & Markman, E. M. (2017). Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences*, *114*(43), 11374–11379. https://doi.org/10.1073/pnas.1704347114

Yogeeswaran, K., Złotowski, J., Livingstone, M., Bartneck, C., Sumioka, H., & Ishiguro, H. (2016). The interactive effects of robot anthropomorphism and robot ability on perceived threat and support for robotics research. *Journal of Human–Robot Interaction*, *5*(2), 29–47. https://doi.org/10.5898/JHRI.5.2.Yogeeswaran

Złotowski, J., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and challenges in human–robot interaction. *International Journal of Social Robotics*, *7*(3), 347–360. https://doi.org/10.1007/s12369-014-0267-6

Złotowski, J., Yogeeswaran, K., & Bartneck, C. (2017). Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human-Computer Studies*, *100*, 48–54. https://doi.org/10.1016/j.ijhcs.2016.12.008

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sundvall, J., Drosinou, M., Hannikainen, I., Elovaara, K., Halonen, J., Herzon, V., Kopecký, R., Jirout Košová, M., Koverola, M., Kunnari, A., Perander, S., Saikkonen, T., Palomäki, J., & Laakasuo, M. (2023). Innocence over utilitarianism: Heightened moral standards for robots in rescue dilemmas. *European Journal of Social Psychology*, *53*, 779–804. https://doi.org/10.1002/ejsp.2936

10990992, 2023, 4, Downloaded from https://onlinelibrary.wiley.com/doi/10.1002/ejsp.2936 by University Of Helsinki, Wiley Online Library on [05/03/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License