International Journal of Social Robotics
https://doi.org/10.1007/s12369-025-01276-9

Artificial Morality: Differences in Responses to Moral Choices by Human and Artificial Agents

Diana Armbruster[1] · **Sarah Mandl**[1] · **Anne Zeiler**[2] · **Anja Strobel**[1]

Accepted: 3 June 2025

Abstract
A consensus on moral "rights" and "wrongs" is essential for ensuring societal functioning. Moral decision-making has been investigated for decades focusing on human agents. More recently, research has started into how humans evaluate artificial moral agents. With increasing presence of artificial intelligence (AI) in society, this question becomes ever more relevant. We investigated responses from a third-party perspective to moral judgments of human and artificial agents in high-stakes and low-stakes dilemmas. High-stakes dilemmas describe life-or-death scenarios while low-stakes dilemmas do not have lethal albeit nevertheless substantial negative consequences. In two online studies, participants responded to the actions resp. inactions of human and artificial agents in four high-stakes scenarios (N_1 = 491) and four low-stakes dilemmas (N_2 = 490). In line with previous research, agents received generally more blame in high-stakes scenarios and actions resulted overall in more blame than inactions. While there was no effect of scenario type on trust, agents were more trusted when they did *not* act. Although humans, on average, were blamed more than artificial agents they were nevertheless also more trusted. The most important predictor for blame and trust was whether participants agreed with the moral choice of an agent and considered the chosen course of action as morally appropriate – regardless of the nature of the agent. Religiosity emerged as further predictor for blaming both human and artificial agents, while trait psychopathy was associated with more blame of and less trust in human agents. Additionally, negative attitudes towards robots predicted blame and trust in artificial agents.

Keywords Moral dilemmas · Moral judgement · Artificial agents · AI · Blame · Trust

1 Introduction

Agreement on what is morally "right" and "wrong" is essential to the functioning of society. Until very recently humans have been deemed to be the only beings capable of moral decisions, although some animals have been, albeit inconsistently, seen as moral creatures (even if not as full moral agents) and the issue of animal morality is currently discussed anew (cf. Monsó et al. 2018). However, the ongoing rapid development of the capabilities of artificial intelligence (AI) might give rise to a new type of moral agent. Thus, for future societies of humans coexisting with various types of artificial and hybrid agents to function, questions need to be addressed regarding their respective tasks and responsibilities (Meyer et al. 2023). Given their current rapid development, these new agents will likely gain in autonomy and their decisions might (to varying degrees) be dissimilar from those that humans might take (Gesmann-Nuissl 2018). Increasingly, artificial agents will also be involved in situations requiring moral judgments (Jentzsch et al. 2019). Although there is no complete agreement among all humans even on key moral issues and considerable individual and cultural differences exist in human moral decision-making (Friesdorf et al. 2015; Graham et al. 2016), disagreements between humans and artificial agents on moral questions represent a new type of challenge. Since such disagreements also touch on the more fundamental issue of roles, rights, and responsibilities of humans and the emerging different artificial societal actors (Meyer et al. 2023), this challenge needs to be timely met. Thus, the question of how humans

✉ Diana Armbruster
diana.armbruster@psychologie.tu-chemnitz.de

[1] Personality Psychology and Assessment, Institute of Psychology, Chemnitz University of Technology, 09120 Chemnitz, Germany

[2] YOUSE GmbH, 13187 Berlin, Germany

Published online: 01 July 2025

judge and respond to the decisions of artificial moral agents warrants further investigation, even if – for now – this concerns decisions in fictional scenarios.

Moral decisions in humans have been investigated for decades using different approaches (review: Ellemers et al. 2019) with moral dilemmas being one well-established method. Moral dilemmas describe a brief (fictional) scenario with usually two mutually exclusive outcomes to choose (first-party perspective). Alternatively, the decisions of others in a dilemma situation might be presented to be judged from a third-party perspective (Christensen and Gomila 2012). Importantly, the application of different moral principles suggests conflicting courses of action in these scenarios. *Deontological* principles are based on universal rules of what is right and wrong and are thus independent of a specific situation and its outcome. Conversely, *utilitarian* principles focus on the outcome of an action (or inaction) and ultimately aim to maximize benefit for as many people as possible even if it means harming one or a few individuals (Gawronski and Beer 2017). Notably, these two principles are not necessarily in conflict with each other in every situation where moral decisions are required. However, moral dilemmas are specifically designed to pit these principles against each other. Furthermore, research into moral scenarios highlighted the importance of various design and content factors that affect responses (Christensen and Gomila 2012). Among those factors are methodological aspects (e.g., word count, question format, expression style) and conceptual variables (e.g., personal force, intention, benefit recipient; Christensen et al. 2014; Christensen and Gomila 2012). It should also be noted that the validity of various scenarios, including the well-known *trolley dilemma*, has been questioned because some of these dilemmas are lacking in realism (Fried 2012; Gold et al. 2014; Kahane 2015). In response, new scenarios based on real-life events have been developed (e.g., Körner and Deutsch 2022) and every-day dilemmas with lower stakes have been proposed (e.g., Singer et al. 2019), though these have been less studied to date.

Third-party perspective. Responses of third parties (i.e., 'observers') to moral decisions of others are of additional interest because third parties are important for maintaining social cooperation and interpersonal trust which depend on consistent reprimanding and punishment of moral transgressions (Boyd et al. 2003). Observers have been suggested to act as everyday judges and to object to perceived violations of moral or social norms even if they are not involved themselves (Weiner 2006). They are of particular importance in this regard in larger and increasingly anonymous societies with frequent one-off interactions (Bendor and Swistak 2001; Boyd et al. 2003). Thus, although most studies on moral decision-making have investigated the first party perspective, there is a strong ongoing research interest in responses of observers to moral choices of others (e.g., Behnke et al. 2020).

Moral choices of artificial agents. While the majority of third-party studies investigated moral decisions by humans, research has started on responses to moral choices by artificial agents (e.g., Awad et al. 2018; Bigman and Gray 2018; Malle et al. 2019; Malle et al. 2015). The studies used similar approaches to the ones on human decision-makers and presented moral dilemmas to assess whether various decisions of (fictional) artificial moral agents were deemed appropriate. In addition, measures of blame and trust have been explored (cf. Malle et al. 2015). Findings suggest that artificial agents are expected to act more utilitarian than human agents in adaptations of the trolley dilemma, i.e., to sacrifice one individual to save four (Malle et al. 2015). Furthermore, human and artificial agents receive different amounts of blame for the same moral choices, although findings are inconsistent on whether artificial agents can be blamed at all, i.e., whether they can be morally responsible. While some studies reported that artificial agents were blamed (although to a different degree) by most participants (Malle et al. 2015, 2019), others found that the majority did not or just to a small extent ascribe morality to robots and consequently did not blame them (Bretschneider et al. 2022; Mandl et al. 2022). Bigman and Gray (2018), who investigated to what extend human decision makers are preferred to artificial agents in various (fictional) medical, legal, or military scenarios, concluded that there is a general preference for human agents. They described an 'aversion' to moral decision-making by artificial agents the degree of which depended on the agents' perceived experience and expertise (Bigman and Gray 2018). However, they also emphasized that they investigated life-or-death scenarios and that results might differ in dilemmas with less at stake. In sum, there is growing interest into how artificial moral agents are evaluated by human observers accompanied by an ongoing debate about whether and to what extent such agents are actually capable of moral decisions and whether they should be allowed to make them (Malle 2016; Meyer et al. 2023; Misselhorn 2018).

Individual differences in evaluating moral choices. In addition to general preferences for human vs. artificial moral agents, individual differences in evaluating moral choices by artificial agents can be expected. Research focusing on human moral agents indicated gender differences with men being more likely to endorse utilitarian options (e.g., Armbruster et al. 2021; Banerjee et al. 2010; Bjorklund 2003; Capraro and Sippel 2017; Fumagalli et al. 2010) although some studies did not find differences between men and women (e.g., Brannon et al. 2019; Seyedsayamdost 2015). Furthermore, there are gender differences in attitudes

towards robots and AI (Funk et al. 2020; Sindermann et al. 2021) as well as towards technology in general (meta-analysis: Cai et al. 2017) with men reporting more positive attitudes. Thus, gender differences might also exist in judging decisions of artificial moral agents.

In addition to gender effects, general population differences in attitudes towards technology, robots, or AI are likely to affect responses towards artificial agents making moral decisions. Previous studies show, for instance, that negative attitudes towards robots influence interactions with them (Babel et al. 2022; Nomura et al. 2008) and evaluations of their behavior styles (Syrdal et al. 2009), while affinity for technology interaction as a more general measure of approaches to technology has been linked to self-reported usage of technical systems (Franke et al. 2019). Therefore, different outlooks on both broader as well as specific technological issues might predict responses to artificial moral agents.

Personality traits have also been found to be associated with moral judgement tendencies including, for instance, *Need for Cognition* (NFC) and trait *Psychopathy*. NFC refers to the tendency to engage in and enjoy effortful cognitive activities (Cacioppo and Petty 1982). Previous studies have linked NFC to utilitarian judgement tendencies (e.g., Conway and Gawronski 2013) although findings are not consistent and associations with deontological preferences have also been found (e.g., Körner et al. 2020; Park et al. 2016). Higher NFC levels had initially been suggested to lead to increased deliberation of costs and benefits and thus been associated with utilitarian decisions. However, recently NFC has also been proposed to result in increased reflections of moral norms to explain correlations with deontological decisions (Körner et al. 2020). Furthermore, NFC predicts self-reported moral behavior, e.g., donating and supporting others in need or considering action consequences for others, suggesting that enjoyment of and engagement in effortful cognitions might represent a 'moral capacity' (Strobel et al. 2017). NFC has also been found to be negatively associated with punitive reactions (Sargent 2004) and, with regard to robots, to be linked to more positive attitudes (Reich-Stiebert and Eyssel 2015; Spatola and Wykowska 2021). Contrariwise, deviant moral behavior has been recognized as a core element of psychopathy as reflected in its early descriptions as 'moral derangement' (Rush 1812) or 'moral insanity' (Prichard 1835). Psychopathy comprises characteristics like antisocial behavior, lack of empathy, remorse or guilt, glibness, shallow affect and impulsivity (Hare and Neumann 2009). Since psychopathy exists on a continuum, the clinical construct was adapted to the sub-clinical domain (cf. Hare 1985). Findings of moral dilemma studies suggest that individuals with higher trait psychopathy are more inclined to make utilitarian choices (meta-analysis: Marshall et al. 2018), which results in less casualties in (fictional) sacrificial dilemmas. This conflicts with real-life behavior of persons with increased levels of psychopathy, who usually appear to be not particularly concerned with enhancing the 'greater good' as they are responsible for a disproportionate amount of physical, financial, social or emotional harm (Kiehl and Hoffman 2011). Further research has shown that a general lack of empathy and reduced compassion for others (Glenn et al. 2009; Seara-Cardosa et al. 2013) together with a reduced dislike for performing harmful actions (Patil 2015) may contribute to these 'utilitarian' choices. Recent studies revealed that individuals scoring higher in trait psychopathy actually show reduced deontological *and* reduced utilitarian inclinations but increased action tendencies (Gawronski et al. 2017; Körner et al. 2020). Trait psychopathy belongs to the group of 'dark' traits which have been originally suggested to comprise a Dark Triad together with narcissism and Machiavellianism (Paulhus and Williams 2002). Later research proposed additional traits (e.g., spitefulness, egoism) as the basis of the dark core of personality (Moshagen et al. 2018). Given the links between these traits and moral attitudes in general, they might also impact reactions towards moral choices by artificial agents.

Religion has been identified as another important factor influencing moral decisions and attitudes (Cohen 2015; Graham et al. 2016) with deontological judgements being positively associated with religiosity (Szekely et al. 2015). Furthermore, religiosity has been linked to negative views on interactions with robots (Giger et al. 2017) and to more fearful attitudes towards them (Katz and Halpern 2014), although different religions have been proposed to exert different effects in this regard (Halpern and Katz 2012; MacDorman et al. 2009; Shaw-Garlock 2009). Nevertheless, religiosity might shape attitudes towards artificial moral agents.

In this study, we aim to investigate the following variables to determine whether and to what degree they affect how humans respond to moral choices made by artificial entities: (a) dilemma type (high vs. low stakes), (b) agent type (human vs. artificial), (c) agent's choice (action vs. inaction), and (d) personal characteristics (e.g., personality traits, attitudes). All variables were selected because previous findings have linked them to differences in moral decision-making. However, to our knowledge, they have not been investigated together and some of them have also only been examined with regard to humans' moral choices but not to decisions of artificial moral agents. Understanding the role of these variables will help to further tailor interactions between humans and robots/AI in moral situations, and to avoid setups that result in frustration with and rejection of artificial agents. Specifically, we evaluate ratings on whether

moral decisions are considered (1) appropriate as well as (2) how much blame an agent deserved for their decision and (3) to what degree the agent can be trusted. All three variables are key parameters in moral interactions, and their joint inclusion allows a more fine-grained assessment of people's responses to the decisions of artificial moral agents (cf. Malle et al. 2015). We also investigate the effect of situation and person variables on response outcomes. Situation-related, we contrast responses to high-stakes with low-stakes dilemmas. In general, low-stakes scenarios have been studied less frequently in moral psychology research (cf. Singer et al. 2019). However, they are more likely to occur in real life. As the presence of artificial entities increases, the probability of low-stakes moral dilemmas involving them also rises. Responses to high-stakes scenarios, on the other hand, are of interest because of their severe and often irreversible consequences. Person-wise, we investigate effects of religiosity, NFC, Dark Triad traits, attitudes towards technology and robots as well as age and gender. These variables have previously been linked to moral judgments and/or preferences resp. behavior towards robots/AI. However, most of them have not been investigated in the context of moral choices of artificial entities.

Based on previous findings on artificial moral agents albeit in different moral dilemmas (e.g., Bigman and Gray 2018; Malle et al. 2019; Malle et al. 2015), we formulated and preregistered the following hypotheses (see https://osf.io/cmqbs/): (1a) Artificial moral agents are blamed less for acting in high as well as low-stakes dilemmas compared to human agents, and are thus rated more positively. (1b) Artificial moral agents are blamed more for inaction in both dilemma types compared to human agents. Furthermore, as Need for Cognition (NFC) has been associated with generally less punitive reactions (Sargent 2004) and more positive attitudes towards robots (Reich-Stiebert and Eyssel 2015; Spatola and Wykowska 2021), NFC was hypothesized to be (2) negatively associated with blaming artificial agents and positively with trusting them, particularly when those agents chose action over inaction. We also preregistered the following research questions: Compared to human agents, how much trust is attributed to artificial moral agents after action resp. inaction? How do attitudes towards robots influence the relationship between (a) NFC and blame judgement and (b) NFC and trust in artificial moral agents? How do Dark Triad personality traits (narcissism, psychopathy, Machiavellianism) influence blame judgement resp. trust in artificial moral agents after a decision for action/inaction in moral dilemmas? To what extent does Affinity for Technology Interaction (ATI) influence the relationship between agent type and the evaluation of the decision to act resp. not to act? Are there differences between evaluation of artificial moral agents and human actors in high- vs. low-stake dilemmas? Answering these questions will contribute to identifying further key regulators of human-robot interactions in moral contexts, and thus help to shape their encounters in beneficial ways.

2 Materials and Methods

We report how we determined our sample size, all data exclusions, all manipulations, and all measures in the study (Simmons et al. 2012). All materials, data, and analyses of this study are available online (https://osf.io/cmqbs/). Both studies were preregistered (https://osf.io/cmqbs/).

2.1 Sample

We conducted two online studies and originally recruited $N_1 = 500$ and $N_2 = 500$ participants using Prolific Academic (www.prolific.co; Palan and Schitter 2018) who responded to decisions of human and artificial agents in high-stakes (study 1) and low-stakes (study 2) moral dilemma scenarios. Individuals who took part in study 1 could not participate in study 2. To ensure comprehension of the scenarios' text, participants had to be either German native speakers or had to speak the language at a proficient level. In study 1, $n = 247$ female (49.4%), $n = 246$ male (49.2%), and $n = 7$ diverse individuals (1.4%) participated. In study 2, $n = 257$ female (51.4%), $n = 233$ male (46.6%) and $n = 10$ diverse individuals (2.0%) took part. Since gender effects were investigated in all analyses, diverse subjects had to be excluded because of the small subsample size. Furthermore, two individuals indicated to be under 18 years of age and were excluded from all analyses as well. Thus, in the final first sample, $N = 491$ individuals remained (mean age = 30.53, SD = 11.06, range = 18–72 years) while the second sample consisted of $N = 490$ individuals (mean age = 31.46, SD = 10.97, range = 18–72 years).

2.2 Moral Dilemmas

In study 1, four high-stakes dilemmas detailing public health related scenarios were adapted from Gawronski et al. (2017). They were rephrased in the third-party perspective with either a treating physician or a treating care robot as moral agent who decided to follow a suggested course of action or not (cf. Malle et al. 2015). Similarly, in study 2 four low-stakes dilemmas in the third-party perspective were employed. Content-wise, low stake dilemmas included mainly administrative decisions (e.g., on study grants or library bans) with either a human or an artificial moral agent who decided for or against a certain action. Three of the four low-stakes dilemmas were adapted from pre-existing

scenarios. The *library ban* scenario is a modified version of the *smart house* dilemma (Liao et al. 2019) and the *early parole* scenario is based on a dilemma used by Bigman and Gray (2018). The *obese patient* dilemma was adapted from a conflict scenario posted on the Open Roboethics institute's website (Open Roboethics Institute 2015), while the *scholarship* dilemma is inspired by real-life differences between funding bodies in how they take into account different characteristics of an applicant when deciding who gets funding and who does not. While in each high-stakes dilemma the lives of a single individual or multiple persons were on the line, none of the low-stakes scenarios contained a life-or-death situation. Thus, in contrast to the sacrificial high-stakes scenarios, low-stakes dilemmas represent every-day moral conflicts. The exact wording of all dilemmas can be found in the supplement and online (https://osf.io/cmqbs/, https://osf.io/mf2ax/ and https://osf.io/x8d4m/).

After reading the main text of a scenario, participants first indicated whether they judged it morally appropriate or not for the human or artificial agent to take the suggested action. Following, participants were presented with the decision made by the respective agent who either chose to act or not. Agreement (yes vs. no) between participants' own moral judgements and an agent's actual decision was later calculated for each scenario to which participants responded. Participants then rated how much blame the agent deserved for their choice on a visual analog scale ranging from "no blame at all" to "maximum blame"; see also supplement for a depiction of the scale) and to what degree the agent could be trusted. Trust was evaluated on an 8-point Likert scale ranging from 0 = "not at all" (trustworthy) to 7 = "very much". In both studies we used a 2 (agent: human vs. artificial) × 2 (decision: action vs. inaction) design. To obtain a balanced distribution in these four conditions across the four dilemmas, a Latin square was used in both studies resulting in 16 randomization groups in each study with differing condition combinations to which participants were randomly assigned.

2.3 Questionnaires

To assess Need for Cognition, the German version of the *NFC Short Scale* (Bless et al. 1994) was used. Participants responded to the 16 items of the questionnaire on a 7-point Likert scale with higher sum scores indicating higher Need for Cognition levels. Internal consistency of the NFC short scale was $\alpha = 0.89$ (sample 1) and $\alpha = 0.87$ (sample 2), respectively. Religiosity was assessed with three items each from the Duke University Religion Index (DUREL; Koenig and Büssing 2010) and the Centrality of Religiosity Scale (CRS; Huber and Huber 2012). In addition, participants responded to one item assessing self-perceived general religiosity (ranging from 1 - not religious to 10 - religious). Combining these z-standardized items resulted in a scale with an internal consistency of $\alpha = 0.90$ and 0.91, respectively, in the two samples. Thus, a single indicator of religiosity was used with higher mean scores indicating higher trait levels. The German Version of the *Dirty Dozen* (Küfner et al. 2015) was employed to measure the three traits of the Dark Triad: psychopathy, narcissism, and Machiavellianism (assessed with four items each). Participants responded on a 9-point Likert scale with higher means in the respective sub-scales indicating higher trait levels. In the two samples, Cronbach's alpha for the subscales were as follows: psychopathy $\alpha = 0.67$ and 0.62, narcissism $\alpha = 0.74$ and 0.82, and Machiavellianism $\alpha = 0.79$ and 0.82. We used the *Affinity for Technology Interaction* (ATI) scale (Franke et al. 2019) to assess the tendency to actively engage in technology interaction. The 9-item scale was originally developed in German and uses a 6-point Likert scale response format. Internal consistency of the ATI scale in the two samples was $\alpha = 0.91$ and 0.92, respectively. The German version of the *Negative Attitudes towards Robots Scale* (NARS; Nomura et al. 2008) was used to capture attitudes concerning interaction and communication with robots. The scale comprises a total of 14 items on three sub-scales assessing negative attitudes towards (1) situations and interactions with robots (six items; $\alpha = 0.75$ and 0.74), (2) social influence of robots (five items; $\alpha = 0.72$ and 0.70), and (3) emotions in interaction with robots (three items; $\alpha = 0.63$ and 0.67). The NARS uses a 5-point Likert scale response format (Nomura et al. 2006) with higher mean scores indicating more pronounced negative attitudes towards robots.

2.4 Procedure

In both online studies, participants were first informed about the study aims and protocol as well as about data protection policies and were asked to indicate their consent via button press. Afterwards, demographical variables (i.e., age, biological sex, social gender, educational level) were assessed. Following, participants completed questionnaires to measure Need for Cognition (NFC), Religiosity, Dark Triad traits, as well as Affinity for Technology Interaction (ATI) and Negative Attitudes towards Robots (NARS). Afterwards, participants were introduced to the moral dilemma experiment. They then completed the four scenario × condition combinations they had been randomly assigned to. As described above, for each scenario participants first indicated whether a suggested course of action was morally appropriate or not for a respective agent to take. Following, they were informed about the choice of the agent and rated blame for and trust in the agent. After completing the survey, participants were thanked. Participants of both samples

received 5 € each for completing the study. The study design and protocol was approved by the ethics committee of Chemnitz University of Technology (#101499823).

2.5 Statistical Analysis

All analyses were performed with SPSS 29 (IBM Statistics). Using repeated measurements ANOVAs, we first analyzed effects on (a) participants' ratings on whether it was morally appropriate for the respective agent to take the suggested action. Agent type (human vs. artificial) was entered as within-subject factor, while dilemma type (low vs. high-stakes) and gender were entered as between-subject factors. In additional ANOVAs effects of dilemma type, agent type, agent's decision (action/inaction) and gender on (b) blame and (c) trust ratings were investigated. Following, regression analyses (enter method) were conducted to further assess the role of personality traits and attitudes as potential predictors of the amount of blame agents received and to what degree they were trusted. The following variables were entered as predictors for responses to human agents: dilemma type, agreement with agents' moral choices, gender, age, Need for Cognition (NFC), religiosity, and the Dark Triad traits narcissism, psychopathy and Machiavellianism. Agreement ratings were based on whether participants' own moral judgements in a given scenario were in line with the presented moral choice of the agent or not. In addition to the aforementioned predictors, four more variables were included to predict responses to artificial agents: Affinity for Technology Interaction (ATI) and the three NARS subscales negative attitudes towards (1) situations and interactions with robots, (2) social influence of robots, and (3) emotions in interaction with robots. Ancillary correlation analyses were conducted to investigate associations between the various personality traits and attitudes as well as between trust and blame whose results can be found in the supplement.

3 Results

3.1 ANOVA: Effects on Moral Appropriateness Ratings of Actions in Dilemma Situations

Participants rated for both high-stakes and low-stakes dilemmas whether a suggested course of action was appropriate or not for the respective agent. There was no difference between the averaged appropriateness ratings to the two dilemmas types (high-stakes: 49.03%; low-stakes: 50.05%; $F_{1, 976} = 0.51$, $p = .477$). Furthermore, there was no general effect of agent type ($F_{1, 976} = 0.09$, $p = .767$) with endorsement rates of suggested actions being 49.39% for human agents and 49.80% for artificial agents. There was also no interaction between dilemma type and type of agent ($F_{1, 976} = 0.64$, $p = .422$) and men and women did not differ in their overall endorsement of suggested actions ($F_{1, 976} = 0.004$, $p = .948$). There were also no interactions between gender and agent type ($F_{1, 976} = 0.74$, $p = .389$) or gender and dilemma type ($F_{1, 976} < 0.001$, $p = .993$). However, there was a three-way interaction between dilemma type, agent and gender ($F_{1, 976} = 7.61$, $p = .006$, $\eta_p^2 = 0.008$; see Fig. 1). The effect is mainly due to high-stakes dilemmas with follow-up analyses revealing a significant gender × agent interaction ($F_{1, 489} = 7.01$, $p = .008$, $\eta_p^2 = 0.014$) which was not found in the low-stakes scenarios ($F_{1, 487} = 1.69$, $p = .195$). In high-stakes scenarios, women were somewhat less likely than men to endorse the actions of human agents (0.45 vs. 0.51). Contrariwise, more women than men endorsed the actions of artificial agents in high-stakes scenarios (0.53 vs. 0.47).

3.2 ANOVA: Effects on Blame of Moral Agents

Blame ratings differed by dilemma type ($F_{1, 976} = 63.61$, $p < .001$, $\eta_p^2 = 0.061$) with increased blame in high-stakes scenarios. Overall blame ratings were also higher for human compared to artificial agents ($F_{1, 976} = 46.24$, $p < .001$, $\eta_p^2 = 0.045$) and for actions compared to inactions ($F_{1, 976} = 83.56$, $p < .001$, $\eta_p^2 = 0.079$). Furthermore, there were significant interactions between dilemma type and agent ($F_{1, 976} = 7.65$, $p = .006$, $\eta_p^2 = 0.008$), dilemma type and decision type (action/inaction; $F_{1, 976} = 17.01$, $p < .001$, $\eta_p^2 = 0.017$) and agent type and decision type ($F_{1, 976} = 10.01$, $p = .001$, $\eta_p^2 = 0.010$; see Fig. 2). Follow-up analyses revealed that in both high-stakes and low-stakes scenarios there was more overall blame for human agents (all $p \leq .003$) and for actions compared to inactions (all $p \leq .001$). However, there was an agent type × decision type interaction effect on blame in the high-stakes scenarios ($F_{1, 489} = 8.64$, $p = .003$, $\eta_p^2 = 0.017$) but not in the low-stakes dilemmas ($p = .121$). Overall, we could confirm one part of our hypothesis regarding differences in blaming human vs. artificial agents. Artificial agents were indeed blamed less for choosing to act, but they were also blamed less when deciding not to act. Contrary to our expectation, there were thus general higher blame ratings for human agents, not just in the "action" condition. There was no main effect of gender ($F_{1, 976} = 0.028$, $p = .866$), nor were there any interaction effects involving gender on blame ratings (all $p \geq .244$).

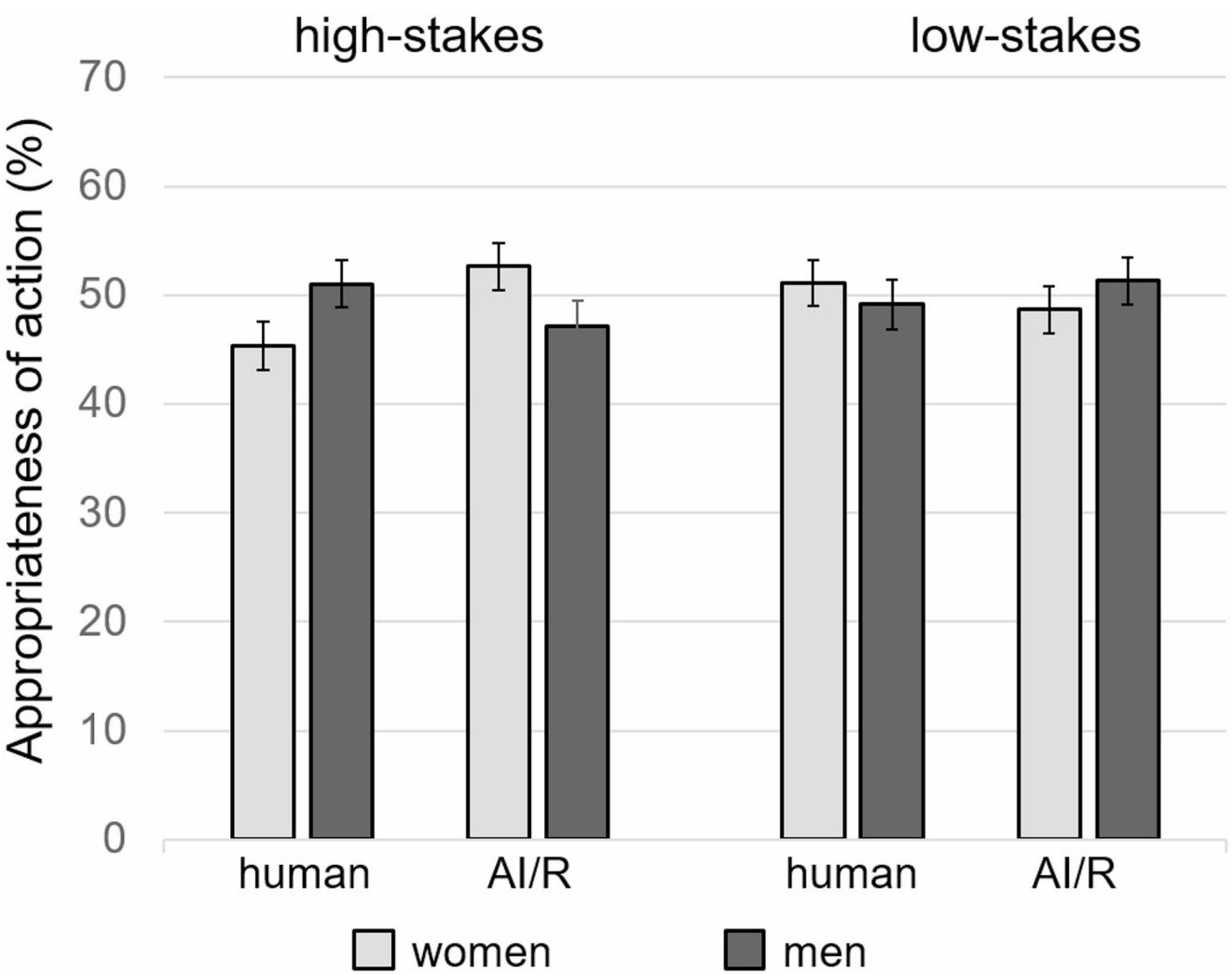

Fig. 1 Mean percentage of agreement with suggested actions in high-stakes and low-stakes dilemmas with human and artificial agents (AI/R)

3.3 ANOVA: Effects on Trust of Moral Agents

Overall, trust in the respective agents did not differ in high- compared to low-stakes scenarios ($F_{1, 826}$ = 0.21, p=.644). However, there was a significant effect of agent type ($F_{1, 826}$ = 6.35, p=.012, η_p^2 = 0.008) with higher trust ratings for human agents as well as a decision type main effect ($F_{1, 826}$ = 15.68, p <.001, η_p^2 = 0.019) with higher trust ratings for agents who chose not to take action. There was also an interaction between dilemma type and decision type ($F_{1, 826}$ = 7.79, p =.005, η_p^2 = 0.009) and between agent and decision type ($F_{1, 826}$ = 5.19, p=.023, η_p^2 = 0.006; see Fig. 3). Follow-up analyses revealed an interaction between agent and decision type in high-stakes scenarios ($F_{1, 489}$ = 4.07, p =.044, η_p^2 = 0.008) that was not present in the low-stakes dilemmas (p =.193). Trust in human agents in high-stakes scenarios is only higher when agents do not take action while there is no difference between human and artificial agents when they do decide to act. Similar to blame ratings, gender did not affect trust ratings ($F_{1, 826}$ = 0.004, p =.948) nor where there any significant interactions involving gender (all $p \geq .300$).

3.4 Regression Analyses: Prediction of Blame and Trust in Human Agents

Regression analyses revealed the following predictors for *blame* of human agents: agreement with agents' moral choices (β = −0.384, p<.001), dilemma type (β = −0.238, p<.001), trait psychopathy (β=0.077, p=.028) and religiosity (β=0.060, p=.037). NFC, narcissism, Machiavellianism, gender, and age did not predict blame for human agents (all $p \geq .063$; see Table 1). *Trust* in human agents was predicted by agreement with agents' choices (β=0.258, p<.001) and trait psychopathy (β = −0.095, p=.015). No other predictor reached significance (all $p \geq .132$; see Table 2).

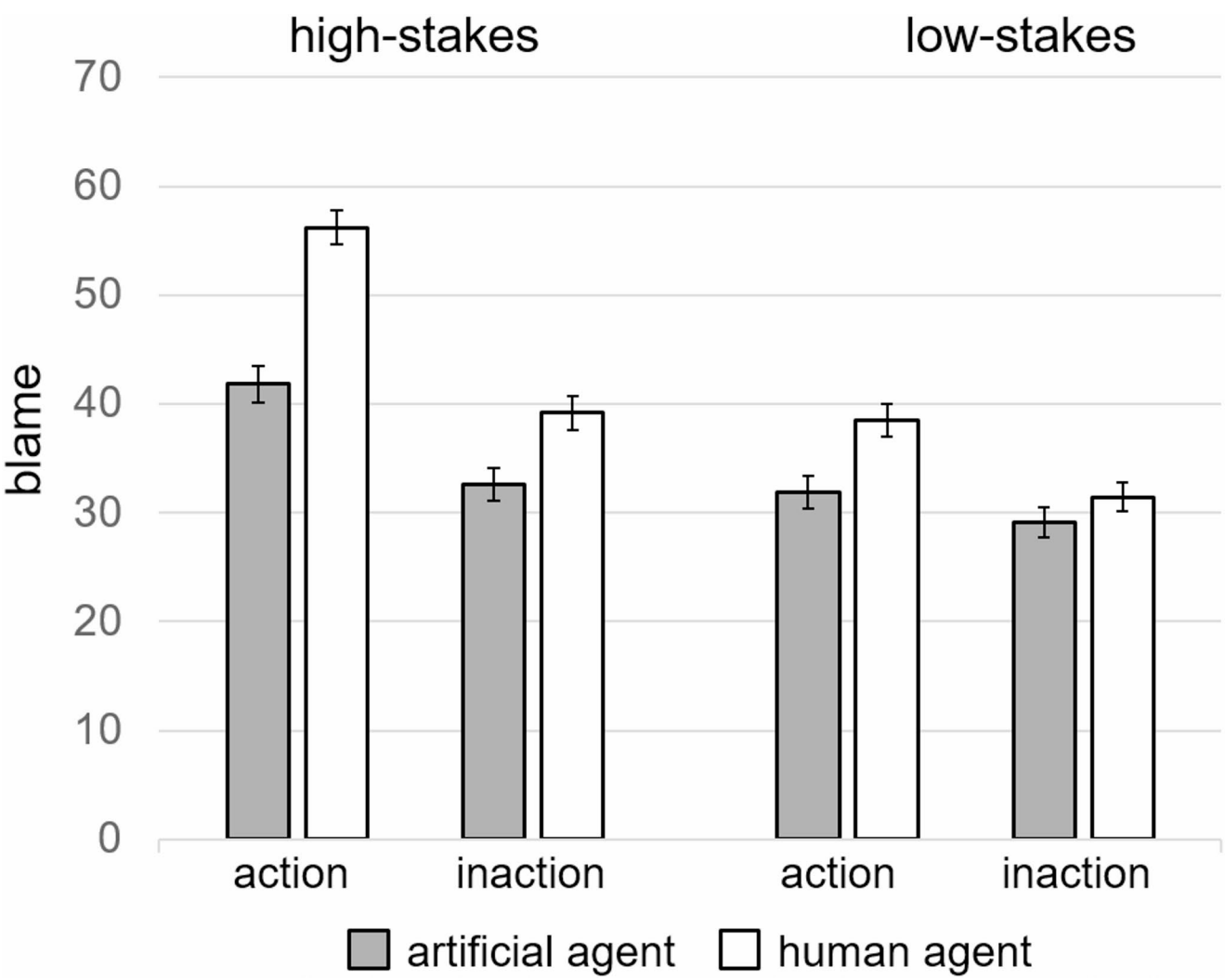

Fig. 2 Mean blame ratings (±SEM; range: 0-100) for actions resp. inactions of human and artificial agents in high-stakes and low-stakes dilemmas

3.5 Regression Analyses: Prediction of Blame and Trust in Artificial Agents

Blame of artificial agents was predicted by agreement with agents' choices (β = −0.323, p<.001), dilemma type (β = −0.106, p<.001), negative attitudes towards situations and interactions with robots (NARS subscale 1, β=0.109, p=.005), narcissism (β=0.096, p=.007), religiosity (β=0.063, p=.040), and age (β = −0.077, p=.014). Affinity for technology interaction, NARS subscales 2 and 3 as well as Machiavellianism, psychopathy, gender, and, contrary to our initial hypothesis, NFC did not predict blame for artificial agents (all p≥.070; see Table 3). *Trust* in artificial agents was predicted by agreement with agents' choices (β=0.218, p<.001) and negative attitudes towards emotions in interaction with robots (NARS subscale 3, β = −0.085, p=.025). All other predictors did not reach significance, including, again contrary to our hypothesis, NFC (all p≥.146; see Table 4).

4 Discussion

4.1 Perceived Appropriateness of Moral Choices by Human and Artificial Agents

We investigated responses to human and artificial agents in high- and low-stakes moral dilemmas, respectively. Overall, there were no differences in perceived appropriateness of suggested actions in the two dilemma types and decision approval did not differ between human and artificial agents. Intriguingly, while there were also no overall gender differences in moral appropriateness ratings, an interaction occurred suggesting differences between men's and women's moral judgements in high-stakes scenarios

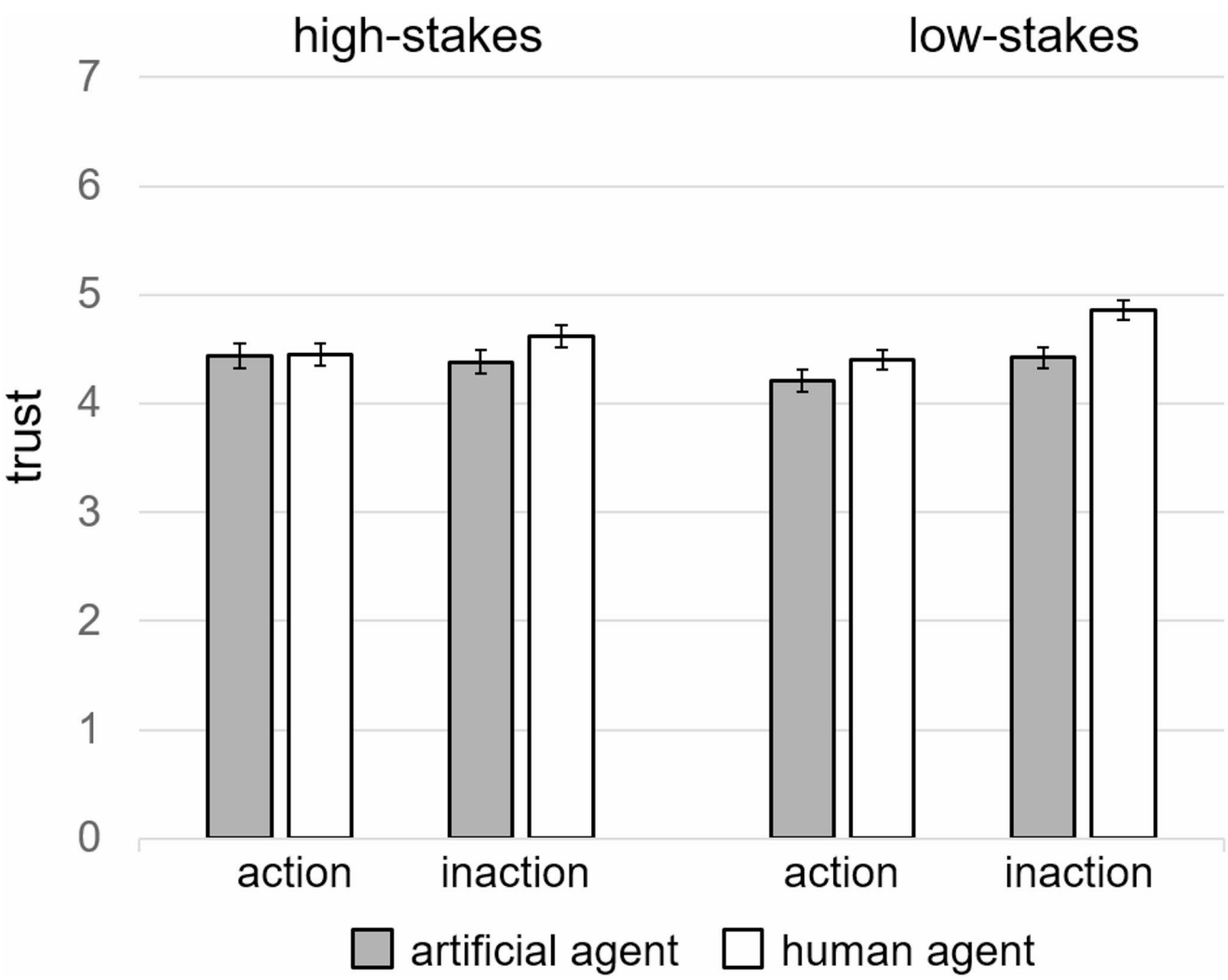

Fig. 3 Mean trust ratings (± SEM; range: 0–7) of acting resp. not acting human and artificial agents in high-stakes and low-stakes dilemma

Table 1 Results of regression analysis on predicting *blame* of human agents

	B	Std.-Error	β	*p*	95% CI lower bound	95% CI upper bound
(Constant)	80.138	5.954		<0.001	68.454	91.823
Dilemma Type	-11.968	1.432	−0.238	**<0.001**	-14.777	-9.158
Agreement	-27.775	2.048	−0.384	**<0.001**	-31.794	-23.756
Gender	−0.271	1.498	−0.005	0.856	-3.212	2.669
Age	−0.127	0.068	−0.056	0.063	−0.260	0.007
Religiosity	1.905	0.910	0.060	**0.037**	0.119	3.690
NFC	-1.361	0.828	−0.048	0.101	-2.986	0.264
Machiavellianism	−0.172	0.553	−0.012	0.756	-1.258	0.913
Psychopathy	1.279	0.581	0.077	**0.028**	0.138	2.420
Narcissism	−0.025	0.510	−0.002	0.960	-1.027	0.976

$R^2 = 0.225$, $F_{9,980} = 31.30$, $p < .001$

Agreement = agreement with agent's moral choice, NFC = Need for Cognition

depending on agent type. In the case of *human* agents, women endorsed the suggested course of action (i.e., instrumental harm to achieve a greater good) to a lesser degree than men. This result mirrors findings from previous moral dilemma research with human agents. Women have been reported to be less likely to choose 'utilitarian' options, i.e., to opt for actions that required harm to one or a few persons for the benefit of multiple individuals (e.g., Armbruster et

Table 2 Results of regression analysis on predicting *trust* in human agents

	B	Std.-Error	β	*p*	95% CI lower bound	95% CI upper bound
(Constant)	3.498	0.452		<0.001	2.611	4.384
Dilemma Type	0.100	0.109	0.029	0.359	−0.114	0.315
Agreement	1.296	0.156	0.258	**<0.001**	0.989	1.603
Gender	0.050	0.114	0.014	0.663	−0.175	0.274
Age	0.003	0.005	0.021	0.531	−0.007	0.013
Religiosity	−0.034	0.069	−0.016	0.622	−0.171	0.102
NFC	0.095	0.063	0.048	0.132	−0.029	0.218
Machiavellianism	0.004	0.042	0.004	0.932	−0.079	0.086
Psychopathy	−0.109	0.044	−0.095	**0.015**	−0.196	−0.022
Narcissism	0.011	0.039	0.011	0.771	−0.065	0.087

$R^2 = 0.079$, $F_{9,959} = 9.05$, $p < .001$

Agreement = agreement with agent's moral choice, NFC = Need for Cognition

Table 3 Results of regression analysis on predicting *blame* of artificial agents

	B	Std.-Error	β	*p*	95% CI lower bound	95% CI upper bound
(Constant)	60.541	8.692		<0.001	43.483	77.600
Dilemma Type	-5.712	1.619	−0.106	**<0.001**	-8.888	-2.535
Agreement	-25.296	2.315	−0.323	**<0.001**	-29.838	-20.753
Gender	0.319	1.810	0.006	0.860	-3.233	3.871
Age	−0.189	0.076	−0.077	**0.014**	−0.339	−0.039
Religiosity	2.145	1.042	0.063	**0.040**	0.101	4.189
NFC	-1.826	1.008	−0.059	0.070	-3.804	0.152
Machiavellianism	−0.019	0.623	−0.001	0.976	-1.240	1.203
Psychopathy	−0.009	0.654	0.000	0.990	-1.292	1.275
Narcissism	1.568	0.576	0.096	**0.007**	0.438	2.698
ATI	−0.211	1.009	−0.007	0.834	-2.192	1.770
NARS 1	4.101	1.470	0.109	**0.005**	1.216	6.985
NARS 2	−0.832	1.316	−0.024	0.528	-3.414	1.751
NARS 3	-1.408	1.162	−0.043	0.226	-3.689	0.873

$R^2 = 0.163$, $F_{13,979} = 14.47$, $p < .001$

Agreement = agreement with agent's moral choice, NFC = Need for Cognition, ATI = Affinity for Technology Interaction, NARS = Negative Attitudes toward (1) Situations and Interactions with Robots, (2) Social Influence of Robots, and (3) Emotions in Interaction with Robots

Table 4 Results of regression analysis on predicting *trust* in artificial agents ($R^2 = 0.071$)

	B	Std.-Error	β	*p*	95% CI lower bound	95% CI upper bound
(Constant)	5.423	0.695		<0.001	4.058	6.788
Dilemma Type	−0.200	0.129	−0.050	0.121	−0.454	0.053
Agreement	1.244	0.183	0.218	**<0.001**	0.884	1.604
Gender	0.048	0.144	0.012	0.741	−0.236	0.331
Age	0.004	0.006	0.020	0.556	−0.009	0.016
Religiosity	0.122	0.084	0.048	0.146	−0.043	0.286
NFC	−0.050	0.080	−0.022	0.532	−0.206	0.106
Machiavellianism	−0.047	0.051	−0.040	0.353	−0.147	0.052
Psychopathy	−0.070	0.052	−0.054	0.179	−0.173	0.032
Narcissism	−0.002	0.047	−0.002	0.966	−0.093	0.089
ATI	0.005	0.081	0.002	0.954	−0.154	0.164
NARS 1	−0.143	0.117	−0.052	0.222	−0.373	0.087
NARS 2	0.008	0.105	0.003	0.942	−0.198	0.213
NARS 3	−0.208	0.092	−0.085	**0.025**	−0.390	−0.027

$R^2 = 0.071$, $F_{13,922} = 5.34$, $p < .001$

Agreement = agreement with agent's moral choice, NFC = Need for Cognition, ATI = Affinity for Technology Interaction, NARS = Negative Attitudes toward (1) Situations and Interactions with Robots, (2) Social Influence of Robots, and (3) Emotions in Interaction with Robots

al. 2021; Banerjee et al. 2010; Bjorklund 2003; Capraro and Sippel 2017; Fumagalli et al. 2010) although findings are not entirely consistent (e.g., Brannon et al. 2019; Seyedsayamdost 2015). Notably, a similar effect of gender was not present in the low-stakes dilemmas. Previous moral dilemma research has predominantly used scenarios that can be classified as high-stakes dilemmas similar to the ones used in study 1 where lives are on the line. Contrariwise, in low-stake scenarios there are no life-or-death consequences associated with any of the two response options. Although choices in these scenarios may result in negative outcomes, their impact is considerably less severe which might explain the lack of difference between the ratings of men and women. Furthermore, in high-stakes scenarios with *artificial* agents the opposite pattern regarding gender was observed: women were more likely to endorse the suggested course of action compared to men. Again, no gender difference of this kind was found for the low-stakes scenarios.

4.2 Blaming and Trusting Moral Agents

Unlike the assessment of moral appropriateness, blame and trust ratings differed between human and artificial agents. Overall, *human* agents received considerably more blame but were also more trusted. Blame requires attribution of moral responsibility which in turn depends on various legal and psychological capacities (e.g., the capacity to act, legal capacity, autonomy, liability, explainability, and moral agency) most of which are usually not ascribed to artificial agents (Meyer et al. 2023). Our finding of lower blame ratings for artificial agents (for making the same decisions as their human counterparts) indicate that they are held less responsible. Contrary to Malle et al. (2015), and to our own initial hypothesis, we did not find higher blame ratings of artificial agents in inaction conditions. Differences in blame ratings of human vs. artificial agents varied and were least pronounced in the low-stakes inaction condition but never veered in the opposite direction. Notably, response differences to human vs. artificial moral agents have been reported to be modulated by additional factors, including an artificial agent's physical appearance (Laakasuo et al. 2021) or their perceived expertise and experience (Bigman and Gray 2018). Furthermore, research into human moral agents highlighted the importance of dilemma content features for judging agents' decisions as morally right or wrong (Christensen et al. 2014; Christensen and Gomila 2012). While, for instance, Malle et al. (2015) used a scenario modelled after the *trolley dilemma*, we investigated moral decisions in medical and administrative contexts. Thus, the response pattern regarding blame and trust found in our study might also be partly due to dilemmas features and might differ in other moral scenarios. Furthermore, we did not provide participants with pictures of the robots or other clues regarding their features. Since the look of a robot has been recently shown to affect humans' responses to the robot's moral decisions (Laakasuo et al. 2021), the (different) way participants might have imagined them to look like, may have had an effect on their responses in this study as well.

Blame and trust were also affected by dilemma and decision type. In both high- and low-stakes dilemmas, actions resulted generally in higher blame and reduced trust ratings compared to inactions, which is in line with previous findings on omission bias. The latter refers to a typical preference of not taking action in moral dilemmas when both acting and doing nothing are expected to result in adverse outcomes, particularly when the consequences are perceived as similarly harmful (cf. Jamison et al. 2020). Furthermore, average blame ratings were higher in high-stakes compared to low-stakes scenarios, although there was no difference in trust ratings between the two dilemma types. Interaction effects indicate that these general tendencies are partly modified by other factors. For instance, in high-stakes dilemmas trust in human compared to artificial agents was only higher in the inaction condition, while there was no difference in trust when agents took action. Contrariwise, there was no interaction of this kind in low-stakes scenarios. Regression analysis additionally confirmed dilemma type as a predictor for blame ratings. However, the by far strongest and most consistent predictor for blame of and trust in both types of agents was agreement with agents' moral choices. Unsurprisingly, agents whose choices tended to be in line with those of the participants received on average less blame and were perceived as more trustworthy. If this finding proves to be robust, it would have significant implications for people's expectations of the behavior of artificial agents. Previously, Malle et al. (2015) reported that participants, who considered it *impermissible* to sacrifice one individual to save four persons, blamed human and artificial agents considerably more for choosing to act than for refraining to act. However, blame ratings of participants, who found it *permissible* to sacrifice the one person, did not significantly differ between agents who decided to act and those who refrained from doing so (Malle et al. 2015).

4.3 Individual Differences in Blaming and Trusting Moral Agents

Religiosity emerged as a further predictor for blaming human and artificial agents alike with higher trait levels being associated with more blame. Generally, religion and religiosity have been argued to be key factors in moral attitudes and judgements (Cohen 2015; Graham et al. 2016). Previously, religiosity has been linked to deontological judgments and increased emotions while processing moral

dilemmas (Szekely et al. 2015). Our findings of higher blame ratings (for human agents) are in line with this. However, religious values vary substantially within and between cultures contributing to considerable differences in morality (Graham et al. 2016). Furthermore, there are also associations between religion and religiosity, respectively, and attitudes towards robots and AI. Giger et al. (2017) reported an association between religiousness and negative attitudes towards interactions with robots. Religiosity has also been linked to a more fearful attitude towards robots (Katz and Halpern 2014). These connections between religiosity and more negative attitudes towards robots or AI might be partly responsible for our finding of more blame for artificial agents by individuals with higher religiosity scores. This notion is supported by positive correlations between religiosity and the NARS subscales in our samples, in particular with subscale 2 (negative attitudes towards social influence of robots; Supplemental Table 4). However, it should be noted that distinct religious belief systems are likely to differently affect acceptance of and attitudes towards robots and AI (Halpern and Katz 2012; MacDorman et al. 2009; Shaw-Garlock 2009). Recently, Ikari et al. (2023) investigated moral care for robots in US participants with predominantly Abrahamic beliefs and in Japanese participants with Shinto-Buddhist traditions. They found higher moral care for robots in Japan. Furthermore, more pronounced religious beliefs were linked to less moral care in the American but not the Japanese sample. Additionally, lower scores in anthropocentrism and higher ones in animism were also linked to increased moral care (Ikari et al. 2023). Given the cultural background of our participants, the association of religiosity and increased blame of artificial agents is in line with similar findings in Western samples (e.g., Giger et al. 2017).

Of the Dark Triad traits, psychopathy in particular showed associations with blame and trust ratings. Psychopathy was associated with more blame of and less trust in *human* agents while narcissism was related to blame of *artificial* agents. Higher levels of trait psychopathy have been associated with harsher punishment of (human) moral agents in fictional moral dilemmas despite concomitantly being linked to reduced inappropriateness ratings of moral transgressions and increased understanding emotions towards the moral agent (Behnke et al. 2020). Psychopathy is also linked to increased vengefulness, and both psychopathy and narcissism have been found to predict reduced forgiveness (Giammarco and Vernon 2014). Shared features of traits belonging to the Dark Triad (Paulhus and Williams 2002) or the dark core of personality (Moshagen et al. 2018) are behavioral tendencies for self-promotion and maximizing one's own interests while ignoring, accepting, or maliciously causing detriment to others combined with beliefs that justify such behavior (Moshagen et al. 2018; Paulhus and Williams 2002). The key is the disregard of others linked to social malevolence, which together with reduced forgiveness might have manifested here in an inclination to ascribe greater blame and indicate less trust.

Contrary to initial expectations, NFC did not predict responses to moral agents, although it only just missed the significance level for blame (human agents: p=.101, artificial agents: p=.070). NFC reflects the tendency to engage in and enjoy effortful cognitive activities (Cacioppo and Petty 1982). Although previous findings are not entirely consistent, reasoning abilities and propensities (i.e., deliberate thinking styles like NFC) have been linked to increased preferences for utilitarian moral choices and optimization of overall welfare without being associated with reduced harm aversion (e.g., Patil et al. 2021). Cognitive abilities resp. motivation have been proposed to contribute to differences in processing moral problems with studies showing selective impairment of utilitarian judgements by cognitive load (Conway and Gawronski 2013; Timmons and Byrne 2019). Based on findings that NFC is negatively associated with punitive responses, Sargent (2004) suggested that higher NFC levels might result in an increased ability and/or willingness to invest cognitive effort to reflect on specifics and constraints of dilemma settings and protagonists. Although associations of NFC with reduced blame did not reach significance in our study, they are, on a descriptive level, in line with reported links of NFC and reduced support for punishment (Sargent 2004). Both are likely due to a more in-depth cognitive analysis of a dilemma situation and its protagonists.

Finally, negative attitudes towards situations and interactions with robots (NARS subscale 1) was a predictor of blaming artificial agents, while negative attitudes towards emotions in interaction with robots (NARS subscale 3) was a predictor of how much they were trusted. The other NARS subscales as well as Affinity for Technology Interaction (ATI) did not predict blame of resp. trust in artificial agents. NARS scores have been linked to actual behavior towards robots (e.g., time talking with them or touching them; Nomura et al. 2008) and evaluation of robot behavior styles (Syrdal et al. 2009). NARS scores were found to be negatively associated with complying with a robot's request in a VR setting (Babel et al. 2022), although in another real-life experiment, NARS scores did not correlate with complying with requests made by a geminoid robot (Aroyo et al. 2018). Thus, despite some inconsistent findings, self-reported attitudes towards robots appear to be linked to behavioral responses to robots, which is echoed in our findings on their associations with blame and trust. Contrary to NARS, Affinity for Technology Interaction (ATI) showed no association with responses to artificial agents, which might be due to

the more general scope of the questionnaire which is not tailored to interactions with robots/AI but focuses on dealing with technology in comparatively broad terms.

Overall, our findings emphasize several general tendencies when responding to artificial compared to human agents. While humans were generally more blamed compared to artificial agents for the same decisions, they were nevertheless more trusted. Also, high-stakes scenarios resulted in higher blame ratings compared to low-stakes dilemmas. However, responses to agents' decisions also showed considerable individual variance and several person variables emerged as predictors of blame and trust. The most important and consistent ones were moral appropriateness ratings, i.e., whether there was agreement or not with an action that an agent was suggested to take. Further predictors included dilemma type and religiosity for blame of both human and artificial agents, and psychopathy for blaming and trusting specifically human agents. For artificial agents, negative attitudes towards robots were additional predictors for blame and trust, respectively. Still, the amount of explained variance (R^2 = 0.163-0.225 for blame and 0.071-0.79 for trust; see also Tables 1, 2, 3 and 4) indicates that there are other predictors, particularly for trust, that were not part of the study and require further investigation. In summary, our findings highlight the fact that people apply partly different standards when judging the moral decisions of robots and/or AI than when judging human decisions. As artificial agents are increasingly integrated into societies, awareness and understanding of such standards is crucial to ensure a successful adoption of these technologies. Satisfying human-robot interactions depend, among other things, on meeting human expectations regarding what type of decisions artificial agents should be allowed to make and in what direction their decision should lean. In turn, these expectations are shaped by several (potentially interacting) factors. Recently, a framework has been introduced (Five Factor of Social Appropriateness (FASA); Wullenkord et al. 2023) which integrates various aspects of social appropriateness (i.e., relations between interacting agents, standards of customary practice, type of action, situational context, individual specifics) and can be applied human-robot interactions. Our research falls into this framework as it investigates several of these aspects in the moral domain. We investigated the role of multiple variables, including situational factors (e.g., dilemma type, agent type) and personal variables (e.g., age, gender, personality traits, attitudes towards technology and robots). Our findings further underline the fact that responses to moral choices of artificial agents depend on multiple factors. They also show that additional variables need to be considered as parts of the variance remain unexplained. Hence, the road to an actual integration of artificial moral agents in societies will be a complicated one. Furthermore, the process is rather dynamic: as people learn more about robots and AI (whose capabilities in turn continue to increase), their attitudes and expectations might also change, albeit not necessarily in a linear fashion. Furthermore, human perceptions and expectations of artificial agents, combined with an acknowledgement of the real limits of robots and/or AI, are also key factors in informing future legislation (cf. Meyer et al. 2023). To be effective, the respective laws to be devised will need to reflect people's sense of what is just. In particular, information on perceived guilt and liability of artificial entities is important in this context.

4.4 Limitations

The study has several limitations. While we were able to recruit sufficiently large samples, the online format adds limits to data quality control. Furthermore, individuals who register on platforms like Prolific to partake in research studies are usually better educated. Accordingly, in both samples about half of our participants reported to hold a university degree and more than additional 30% had graduated from high school ('Abitur/Matura'; see supplemental Table 1). Also, to ensure thorough comprehension of the dilemma texts, only German native speakers and individuals who spoke German at a sufficiently high level could participate. Furthermore, while the age range of our samples is rather large (18–72 years), more than 80% of participants were under 40 years old. Regarding the scenarios used, content variety was limited in particular for high-stakes dilemmas, all of which were set in a medical context. As dilemma content features can affect responses (Christensen et al. 2014; Christensen and Gomila 2012), scenarios with different settings might lead to different results. Furthermore, we only investigated responses to 'generic' types of agents without further specifications or modifications of their characteristics. However, physical features (Laakasuo et al. 2021) and perceived experience and expertise of artificial agents (Bigman and Gray 2018) might modulate responses to their moral decisions. Furthermore, participants completed both the questionnaires and gave their responses to the moral dilemmas in one session. Therefore, potential carry-over effects of the questionnaires on later responses in the dilemma situations cannot be completely ruled out, as participants may have been inclined to align their responses. However, alignment tendencies could also occur in the opposite direction if the dilemmas had been presented first. In general, traits like Need for Cognition or the Dark Triad were hypothesized to be potential influence factors on the responses to the dilemma situations. Thus, we considered a biasing effect on the answers to the questionnaires by a dilemma experiment conducted beforehand as the less favorable variant and

decided to present the questionnaires first. Thus, there are several design factors limiting generalizability of our findings and further research is needed to investigate the interplay of agent and dilemma features.

5 Conclusion

In sum, our study provides further insight into how moral decisions of artificial in contrast to human agents are evaluated and which variables might affect individual differences in those responses. While (dis)agreement with one's own moral choice preferences proofed to be the most important predictor for blaming and trusting other moral agents, several additional predictors emerged. Although studies investigating responses to artificial moral agents cannot sufficiently address the question of whether these agents are actually capable of making moral choices (cf. Meyer et al. 2023), findings nevertheless yield important information. Understanding human perception of and responses to artificial agents in situations with moral implications is essential for optimizing these interactions with due consideration of their psychological and legal limitations. For instance, findings on broader response differences to action vs. inaction in moral conflicts might inform decisions on which 'default response mode' of artificial entities might be preferable in certain types of moral dilemmas. However, findings on the importance of inter-individual differences (e.g., religiosity, attitudes towards robots) also point to future implementation problems as the question arises of how to adjust an artificial entity's actions appropriately to people with different preferences and expectations. With the presence of artificial agents increasing, ensuring smooth encounters with humans becomes ever more important for future societal functioning.

Funding This work was funded by the Deutsche Forschungsgemeinschaft (DFG), Grant CRC 1410.

Data Availability All materials, data, and analyses of this study are available online (https://osf.io/cmqbs/).

Declarations

Conflict of Interest The authors have no relevant financial or non-financial interests to disclose.

References

Armbruster D, Kirschbaum C, Strobel A (2021) Androgenic morality?? Associations of sex, oral contraceptive use and basal testosterone levels with moral decision making. Behav Brain Res 408:113196. https://doi.org/10.1016/j.bbr.2021.113196

Aroyo AM, Kyohei T, Koyama T, Takahashi H, Rea F, Sciutti A, Yoshikawa Y, Ishiguro H, Sandini G (2018) *Will People Morally Crack under the Authority of a Famous Wicked Robot? 2018* 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), Nanjing, China. https://doi.org/10.1109/ROMAN.2018.8525744

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

Babel F, Vogt A, Kraus PH, Angerer J, Seufert F, T., Baumann M (2022) Step aside! Vr-Based evaluation of adaptive robot conflict resolution strategies for domestic service robots. Int J Social Robot 14:1239–1260. https://doi.org/10.1007/s12369-021-00858-7

Banerjee K, Huebner B, Hauser M (2010) Intuitive moral judgments are robust across variation in gender, education, politics and religion: A Large-Scale Web-Based study. J Cognition Cult 10(3–4):253–281. https://doi.org/10.1163/156853710x531186

Behnke A, Strobel A, Armbruster D (2020) When the killing has been done: exploring associations of personality with Third-Party judgment and punishment of homicides in moral dilemma scenarios. PLoS ONE 15(6):e0235253. https://doi.org/10.1371/journal.pone.0235253

Bendor J, Swistak P (2001) The evolution of norms. Am J Sociol 106(6):1493–1545. https://doi.org/10.1086/321298

Bigman YE, 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

Bjorklund F (2003) Differences in the justification of choices in moral dilemmas: effects of gender, time pressure and dilemma seriousness. Scand J Psychol 44(5):459–466. https://doi.org/10.1046/j.1467-9450.2003.00367.x

Bless H, Wänke M, Bohner G, Fellhauer RF, Schwarz N (1994) Need for cognition: eine Skala Zur erfassung von engagement und freude Bei Denkaufgaben. Z Für Sozialpsychologie 25(2):147–154

Boyd R, Gintis H, Bowles S, Richerson PJ (2003) The evolution of altruistic punishment. Proc Natl Acad Sci USA 100(6):3531–3535. https://doi.org/10.1073/pnas.0630443100

Brannon SM, Carr S, Jin ES, Josephs RA, Gawronski B (2019) Exogenous testosterone increases sensitivity to moral norms in moral dilemma judgements. Nat Hum Behav 3(8):856–866. https://doi.org/10.1038/s41562-019-0641-3

Bretschneider M, Mandl S, Strobel A, Asbrock F, Meyer B (2022) Social perception of embodied digital Technologies—a closer look at bionics and social robotics. Gruppe Interaktion Organisation Z Für Angewandte Organisationspsychologie (GIO) 1–16. https://doi.org/10.1007/s11612-022-00644-7

Cacioppo JT, Petty RE (1982) The need for cognition. J Personal Soc Psychol 42(1):116–131. https://doi.org/10.1037/0022-3514.42.1.116

Cai Z, Fan X, Du J (2017) Gender and attitudes toward technology use: A Meta-Analysis. Comput Educ 105:1–13. https://doi.org/10.1016/j.compedu.2016.11.003

Capraro V, Sippel J (2017) Gender differences in moral judgment and the evaluation of Gender-Specified moral agents. Cogn Process 18(4):399–405. https://doi.org/10.1007/s10339-017-0822-9

Christensen JF, Gomila A (2012) Moral dilemmas in cognitive neuroscience of moral Decision-Making: A principled review. Neurosci Biobehav Rev 36(4):1249–1264. https://doi.org/10.1016/j.neubiorev.2012.02.008

Christensen JF, Flexas A, Calabrese M, Gut NK, Gomila A (2014) Moral judgment reloaded: A moral dilemma validation study. Front Psychol 5:607. https://doi.org/10.3389/fpsyg.2014.00607

Cohen AB (2015) Religion's profound influences on psychology: morality, intergroup relations, Self-Construal, and enculturation.

Curr Dir Psychol Sci 24(1):77–82. https://doi.org/10.1177/0963721414553265

Conway P, Gawronski B (2013) Deontological and utilitarian inclinations in moral decision making: A process dissociation approach. J Personal Soc Psychol 104(2):216–235. https://doi.org/10.1037/a0031021

Ellemers N, van der Toorn J, Paunov Y, van Leeuwen T (2019) The psychology of morality: A review and analysis of empirical studies published from 1940 through 2017. Personality Social Psychol Rev 23(4):332–366. https://doi.org/10.1177/1088868318811759

Franke T, Attig C, Wessel D (2019) A personal resource for technology interaction: development and validation of the affinity for technology interaction (Ati) scale. Int J Human–Computer Interact 35(6):456–467. https://doi.org/10.1080/10447318.2018.1456150

Fried BH (2012) What does matter?? The case for killing the trolley problem (or letting it Die). Philosophical Q 62(248):505–529. https://doi.org/10.1111/j.1467-9213.2012.00061.x

Friesdorf R, Conway P, Gawronski B (2015) Gender differences in responses to moral dilemmas: A process dissociation analysis. Pers Soc Psychol Bull 41(5):696–713. https://doi.org/10.1177/0146167215575731

Fumagalli M, Ferrucci R, Mameli F, Marceglia S, Mrakic-Sposta S, Zago S, Lucchiari C, Consonni D, Nordio F, Pravettoni G, Cappa S, Priori A (2010) Gender-Related differences in moral judgments. Cogn Process 11(3):219–226. https://doi.org/10.1007/s10339-009-0335-2

Funk C, Tyson A, Kennedy B, Johnson C (2020) *Science and Scientists Held in High Esteem across Global Publics*

Gawronski B, Beer JS (2017) What makes moral dilemma judgments utilitarian or deontological?? Soc Neurosci 12(6):626–632. https://doi.org/10.1080/17470919.2016.1248787

Gawronski B, Armstrong J, Conway P, Friesdorf R, Hutter M (2017) Consequences, norms, and generalized inaction in moral dilemmas: the Cni model of moral Decision-Making. J Personal Soc Psychol 113(3):343–376. https://doi.org/10.1037/pspa0000086

Gesmann-Nuissl D (2018) Künstliche Intelligenz – Den Ersten schritt Vor dem Zweiten tun! Z Zum Innovations- Und Technikrecht (InTeR) 3:105–106

Giammarco EA, Vernon PA (2014) Vengeance and the dark triad: the role of empathy and perspective taking in trait forgivingness. Pers Indiv Differ 67:23–29. https://doi.org/10.1016/j.paid.2014.02.010

Giger J-C, Moura D, Almeida N, Piçarra N (2017) *Attitudes towards social robots: the role of gender, belief in human nature uniqueness, religiousness and interest in science fiction* II international Congress on interdisciplinarity in social and human sciences. Research Centre for Spatial and Organizational Dynamics, University of Algarve, Faro, Portugal

Glenn AL, Iyer R, Graham J, Koleva S, Haidt J (2009) Are all types of morality compromised in psychopathy?? J Personal Disord 23(4):384–398. https://doi.org/10.1521/pedi.2009.23.4.384

Gold N, Pulford BD, Colman AM (2014) The outlandish, the realistic, and the real: contextual manipulation and agent role effects in trolley problems. Front Psychol 5:35. https://doi.org/10.3389/fpsyg.2014.00035

Graham J, Meindl P, Beall E, Johnson KM, Zhang L (2016) Cultural differences in moral judgment and behavior, across and within societies. Curr Opin Psychol 8:125–130. https://doi.org/10.1016/j.copsyc.2015.09.007

Halpern D, Katz JE (2012) *Unveiling Robotophobia and Cyber-Dystopianism: The Role of Gender, Technology and Religion on Attitudes Towards Robots.* 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)

Hare RD (1985) Comparison of procedures for the assessment of psychopathy. J Consult Clin Psychol 53(1):7–16. https://doi.org/10.1037//0022-006x.53.1.7

Hare RD, Neumann CS (2009) Psychopathy: assessment and forensic [review]mplications [Review]. Can J Psychiatry 54(12):791–802. https://doi.org/10.1177/070674370905401202

Huber S, Huber OW (2012) The centrality of religiosity scale. (Crs) Religions 3(3):710–724. https://doi.org/10.3390/rel3030710

Ikari S, Sato K, Burdett E, Ishiguro H, Jong J, Nakawake Y (2023) Religion-Related values differently influence moral attitude for robots in the united States and Japan. J Cross-Cult Psychol 54(6–7):742–759. https://doi.org/10.1177/00220221231193369

Jamison J, Yay T, Feldman G (2020) Action-Inaction asymmetries in moral scenarios: replication of the omission Bias examining morality and blame with extensions linking to causality, intent, and regret. J Exp Soc Psychol 89:103977. https://doi.org/10.1016/j.jesp.2020.103977

Jentzsch S, Schramowski P, Rothkopf C, Kersting K (2019) The Moral Choice Machine: Semantics Derived Automatically from Language Corpora Contain Human-Like Moral Choices. In *Proceedings of the 2nd Aaai/Acm Conference on Ai, Ethics, and Society*. ACM. https://doi.org/10.1145/3306618. 33142 67

Kahane G (2015) Sidetracked by trolleys: why sacrificial moral dilemmas tell Us little (or Nothing) about utilitarian judgment. Soc Neurosci 10(5):551–560. https://doi.org/10.1080/17470919.2015.1023400

Katz JE, Halpern D (2014) Attitudes towards robots suitability for various jobs as affected robot appearance. Behav Inform Technol 33(9):941–953. https://doi.org/10.1080/0144929X.2013.783115

Kiehl KA, Hoffman MB (2011) The criminal psychopath: history, neuroscience, treatment, and economics. Jurimetrics 51:355–397

Koenig HG, Büssing A (2010) The Duke university religion index (Durel): A Five-Item measure for use in epidemological studies *Religions*. 1(1):78–85. https://doi.org/10.3390/rel1010078

Körner A, Deutsch R (2022) Deontology and utilitarianism in real life: A set of moral dilemmas based on historic events. Pers Soc Psychol Bull 1461672221103058. https://doi.org/10.1177/01461672221103058

Körner A, Deutsch R, Gawronski B (2020) Using the Cni model to investigate individual differences in moral dilemma judgments. Pers Soc Psychol Bull 46(9):1392–1407. https://doi.org/10.1177/0146167220907203

Küfner ACP, Dufner M, Backc MD (2015) Das Dreckige Dutzend und die niederträchtigen neun. Kurzskalen Zur erfassung von narzissmus, machiavellismus und psychopathie. Diagnostica 61(2):76–91. https://doi.org/10.1026/0012-1924/a000124

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

Liao B, Slavkovik M, van der Torre L (2019) *Building Jiminy Cricket: An Architecture for Moral Agreements among Stakeholders* AIES'19, Honolulu, HI, USA. https://doi.org/10.1145/3306618.3314257

MacDorman KF, Vasudevan SK, Ho C-C (2009) Does Japan really have robot mania?? Comparing attitudes by implicit and explicit measures. AI Soc 23:485–510. https://doi.org/10.1007/s00146-008-0181-2

Malle BF (2016) Integrating robot ethics and machine morality: the study and design of moral competence in robots. Ethics Inf Technol 18:243–256. https://doi.org/10.1007/s10676-015-9367-8

Malle BF, Scheutz M, Arnold T, Voiklis C, Cusimano C (2015) 2015, March 2–5). *Sacrifice One for the Good of Many? People Apply Different Moral Norms to Human and Robot Agents* 10th ACM/IEEE International Conference on Human-Robot Interaction

(HRI), Portland, Oregon, USA. https://doi.org/10.1145/2696454.2696458
Malle BF, Magar ST, Scheutz M (2019) Ai in the sky: how people morally evaluate human and machine decisions in a lethal strike dilemma. In: Aldinhas Ferreira MI, Silva Sequeira J, Singh Virk G, Tokhi MO, Kadar EE (eds) Robotics and Well-Being, vol 95. Springer International Publishing, pp 111–133. https://doi.org/10.1007/978-3-030-12524-0_11
Mandl S, Bretschneider M, Meyer S, Gesmann-Nuissl D, Asbrock F, Meyer B, Strobel A (2022) Embodied digital technologies: first insights in the social and legal perception of robots and users of prostheses. Front Rob AI 9787970. https://doi.org/10.3389/frobt.2022.787970
Marshall J, Watts AL, Lilienfeld SO (2018) Do psychopathic individuals possess a misaligned moral compass?? A Meta-Analytic examination of psychopathy's relations with moral judgment. Personality Disorders: Theory Res Treat 9(1):40–50. https://doi.org/10.1037/per0000226
Meyer S, Mandl S, Gesmann-Nuissl D, Strobel A (2023) Responsibility in hybrid societies: concepts and terms. AI Ethics 3:25–48. https://doi.org/10.1007/s43681-022-00184-2
Misselhorn C (2018) Artificial morality. Concepts, issues and challenges. Society 55:161–169. https://doi.org/10.1007/s12115-018-0229-y
Monsó S, Benz-Schwarzburg J, Bremhorst A (2018) Animal morality: what it means and why it matters. J Ethics 22(3):283–310. https://doi.org/10.1007/s10892-018-9275-3
Moshagen M, Hilbig BE, Zettler I (2018) The dark core of personality. Psychol Rev 125(5):656–688. https://doi.org/10.1037/rev0000111
Nomura T, Takayuki K, Suzuki T (2006) Experimental investigation into influence of negative attitudes toward robots on Human–Robot interaction. AI Soc 20:138–150. https://doi.org/10.1007/s00146-005-0012-7
Nomura T, Kanda T, Suzuki T, Kato K (2008) Prediction of human behavior in human–Robot interaction using psychological scales for anxiety and negative attitudes toward robots. IEEE Trans Robot 24(2):442–451. https://doi.org/10.1109/TRO.2007.914004
Open Roboethics Institute (2015) Last updated 24 February 2015). *Results: Should a Carebot Bring an Alcoholic a Drink? Poll Says, It Depends on Who Owns the Robot*. Retrieved 13 December 2021 from https://openroboethics.org/results-should-a-carebot-bring-an-alcoholic-a-drink-poll-says-it-depends-on-who-owns-the-robot/
Palan S, Schitter C (2018) Prolific.Ac – a subject pool for online experiments. J Behav Experimental Finance 17:22–27. https://doi.org/10.1016/j.jbef.2017.12.004
Park G, Kappes A, Rho Y, Van Bavel JJ (2016) At the heart of morality Lies Neuro-Visceral integration: lower cardiac vagal tone predicts utilitarian moral judgment. Soc Cognit Affect Neurosci 11(10):1588–1596. https://doi.org/10.1093/scan/nsw077
Patil I (2015) Trait psychopathy and utilitarian moral judgement: the mediating role of action aversion. J Cogn Psychol 27(3):349–366. https://doi.org/10.1080/20445911.2015.1004334
Patil I, Zucchelli MM, Kool W, Campbell S, Fornasier F, Calo M, Silani G, Cikara M, Cushman F (2021) Reasoning supports utilitarian resolutions to moral dilemmas across diverse measures. J Personal Soc Psychol 120(2):443–460. https://doi.org/10.1037/pspp0000281
Paulhus DL, Williams KM (2002) The dark triad of personality: narcissism, machiavellianism, and psychopathy. J Res Pers 36(6):556–563. https://doi.org/10.1016/S0092-6566(02)00505-6
Prichard JC (1835) A treatise on insanity and other disorders affecting the Mind. Sherwood, Gilbert and Piper
Reich-Stiebert N, Eyssel F (2015) Learning with educational companion robots?? Toward attitudes on education robots?, predictors of attitudes, and application potentials for education robots?. Int J Soc Robot 7:875–888. https://doi.org/10.1007/s12369-015-0308-9
Rush B (1812) Medical inquiries and observations upon the diseases of the Mind. Kimber & Richardson
Sargent MJ (2004) Less thought, more punishment: need for cognition predicts support for punitive responses to crime. Pers Soc Psychol Bull 30(11):1485–1493. https://doi.org/10.1177/0146167204264481
Seara-Cardosa A, Dolberg H, Neumann C, Roiser JP, Viding E (2013) Empathy, morality and psychopathic traits in women. Pers Indiv Differ 55(3):328–333. https://doi.org/10.1016/j.paid.2013.03.011
Seyedsayamdost H (2015) On gender and philosophical intuition: failure of replication and other negative results. Philosophical Psychol 28(5):642–673. https://doi.org/10.1080/09515089.2014.893288
Shaw-Garlock G (2009) Looking forward to sociable robots. Int J Social Robot 1:249–260. https://doi.org/10.1007/s12369-009-0021-7
Simmons J, Nelson L, Simonssohn U (2012) A 21 Word Solution. *Available at SSRN*: https://ssrn.com/abstract=2160588 or http://dx.doi.org/10.2139/ssrn.2160588
Sindermann C, Sha P, Zhou M, Wernicke J, Schmitt HS, Li M, Sariyska R, Stavrou M, Becker B, Montag C (2021) Assessing the attitude towards artificial intelligence: introduction of a short measure in german, chinese, and english Language. KI - Künstliche Intelligenz 35:109–118. https://doi.org/10.1007/s13218-020-00689-0
Singer N, Kreuzpointner L, Sommer M, Wust S, Kudielka BM (2019) Decision-Making in everyday moral conflict situations: development and validation of a new measure. PLoS ONE 14(4):e0214747. https://doi.org/10.1371/journal.pone.0214747
Spatola N, Wykowska A (2021) The personality of anthropomorphism: how the need for cognition and the need for closure define attitudes and anthropomorphic attributions toward robots. Computers Hum Bevior 122:106841. https://doi.org/10.1016/j.chb.2021.106841
Strobel A, Grass J, Pohling R, Strobel A (2017) Need for cognition as a moral capacity. Pers Indiv Differ 117:42–51. https://doi.org/10.1016/j.paid.2017.05.023
Syrdal DS, Dautenhahn K, Koay KL, Walters ML (2009) 6th – 9th April 2009). *The Negative Attitudes Towards Robots Scale and Reactions to Robot Behaviour in a Live Human-Robot Interaction Study.* Adaptive and Emergent Behaviour and Complex Systems: Proceedings of the 23rd Convention of the Society for the Study of Artificial Intelligence and Simulation of Behaviour, Edinburgh, UK
Szekely RD, Opre A, Miu AC (2015) Religiosity enhances emotion and deontological choice in moral dilemmas. Pers Indiv Differ 79:104–109. https://doi.org/10.1016/j.paid.2015.01.036
Timmons S, Byrne RM (2019) Moral fatigue: the effects of cognitive fatigue on moral reasoning. Q J Experimental Psychol 72(4):943–954. https://doi.org/10.1177/1747021818772045
Weiner B (2006) Social motivation, justice and the moral emotions: an attributional approach. Tayler & Francis
Wullenkord R, Bellon J, Gransche B, Nähr-Wagener S, Eyssel F (2023) Social appropriateness in hmi: the five factors of social appropriateness (Fasa) model. Interact Stud 23(3):360–390. https://doi.org/10.1075/is.22017.wul

Diana Armbruster is a postdoctoral researcher at the Professorship of Personality Psychology and Psychological Assessment at Chemnitz University of Technology. Her research focuses on correlates of moral decision-making, including personality traits, emotions, attitudes and biological factors.

Sarah Mandl is a postdoctoral researcher at the Professorship of Personality Psychology and Psychological Assessment at Chemnitz University of Technology. Her research focuses on (social) perception and trustworthiness of embodied digital technologies such as social robots and telepresence systems, moral decision making by artificial entities, and interindividual differences of users of said technologies.

Anne Zeiler is a senior user-experience researcher at YOUSE GmbH, Berlin. Her work focuses on systemic user-centered design and evaluation of digital products and processes across diverse domains such as medical technology, Industry 4.0, and financial services. Her work integrates social, legal, and ethical perspectives into applied research consulting for companies.

Anja Strobel is a full professor of Personality Psychology and Psychological Assessment at Chemnitz University of Technology. Her research focusses on interindividual differences related to relevant personality traits like investment traits, and morality-related personality traits, their assessment, development and practical implications.