Effects of Moral-Emotional Behavior and Intentionality on Mind Attribution and Evaluation of Social Robots

Alexander Leonhardt [a], Martin Maier [a, b], and Rasha Abdel Rahman [a, b]

[a] Humboldt-Universität zu Berlin

[b] Science of Intelligence, Research Cluster of Excellence

Author note

Martin Maier https://orcid.org/0000-0003-4564-9834

Rasha Abdel Rahman https://orcid.org/0000-0002-8438-1570

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135.

Correspondence concerning this article should be addressed to Alexander Leonhardt, Humboldt-Universität zu Berlin, Rudower Chaussee 18, 12489 Berlin, Email: alex.leonhardt@gmail.com

Abstract

Robots are generally not considered full moral agents and therefore not appropriate loci of moral responsibility; however, research suggests that people may attribute mental capacities associated with moral agency as well as moral responsibility to robots. We investigate the extent to which descriptions of social robots as minded agents or mechanical machines differentially impact moral judgments of robots' morally and emotionally relevant actions. In two experiments we considered cognitive and emotional aspects of moral judgments: mind attribution and moral evaluations as well as people's emotional response, a key component in accounts of moral responsibility. Participants ($N = 82$) read and subsequently rated 42 stories about robots' actions, which differed in level of intentionality and moral-emotional quality of information. Although minded as opposed to mechanical descriptions positively influenced mind attribution and attributed moral responsibility, participants' emotional response and moral evaluations were largely dominated by the moral-emotional content of information. In addition, the presentation of an image alongside information negatively affected the influence of minded descriptions of robots on mind attribution. Our findings demonstrate that robots are ascribed moral responsibility even though by current theoretical accounts they are not suitable moral agents. Robots that are described as minded are furthermore attributed greater agency and moral responsibility. We discuss the impact of our findings for theoretical discussions about the ethics of AI and robots.

Introduction

Social robots are a class of embodied artificial intelligence (AI) that is used in direct contact with people in everyday contexts such as homes and places of work (Fong, Nourbakhsh, & Dautenhahn, 2003). The use of such technology is predicted to increase in the near future and comes with the promise of a host of benefits to society (Lutz et al., 2019). Social robots may help reduce labor shortages in industries such as care and education and provide skills that are difficult or dangerous for humans to perform. On the other hand, there are a number of ethical concerns surrounding the use of these machines as social robots may be used in morally ambiguous situations and cause harm to people (Ladak et al., 2023). One of the challenges for the introduction of social robots therefore will be to get the best out of this promising technology whilst avoiding ethical compromises (Borg et al., 2024).

What if robots do bad things?

To this end, there has been interdisciplinary interest in questions concerning the moral responsibility of robots. As robots become more sophisticated and ubiquitous, it is also increasingly likely that they will on occasion cause harm to humans (Borg et al., 2024; Lemley & Casey, 2019). When this occurs, who is held responsible? Are robots appropriate targets for blame? Do we blame robots, regardless of whether it is appropriate to do so? While moral philosophy, e.g. the fields of AI and robot ethics, deals with the question of appropriateness, cognitive sciences can make significant contributions by investigating people's beliefs and judgments about robots.

Ascriptions of moral responsibility are important social mechanisms; we want to be able to hold an appropriate agent accountable for doing wrong in order to disincentivize bad actions being repeated (Borg et al., 2024). When an inappropriate agent is held responsible, this impedes

the desired impact of sanctions by creating a responsibility gap, a situation where no party is held sufficiently accountable (Misselhorn, 2018; Sparrow, 2007). People's intuitions about robots as potential moral agents and their judgements of them in morally relevant situations therefore need to be investigated and offer important insights for related theoretical questions.

Morality and the Mind

Moral responsibility is associated with a series of judgments that are held to have emotional as well as cognitive aspects (Talbert, 2024; Tognazzini & Coates, 2021). Firstly, moral situations, such as moral transgressions, have been conjectured to involve an emotional response (Wallace, 1994; Strawson, 2003). And empirical findings suggest that moral transgressions do indeed invoke an emotional response (Bigman et al., 2023; Ladak et al., 2023). Secondly, people make cognitive evaluations about morally relevant situations. They may discern who is responsible for the situation and if there should be any type of repercussion (Smart, 1961). These evaluations may therefore also include suitability of praise or punishment. Finally, there must be someone—a moral agent—who can be held to account for the morally relevant situation.

What makes a moral agent? Most importantly, moral agents require a degree of autonomy (Loh & Loh, 2017; Tognazzini & Coates, 2021). A moral agent must have acted from their own volition and intended to commit the morally relevant action (Gray et al., 2012). Moral agency therefore entails a particular attitude towards or set of assumptions about an agent and their mental abilities, this is sometimes referred to as taking an intentional stance (Fischer & Ravizza, 1998; Chopra, 2011). Only agents who can hold beliefs and make autonomous decisions based on their beliefs are typically taken to be moral agents (Loh & Loh, 2018; Talbert, 2024). Indeed, intentionality, or mindedness, is a key component of people's moral intuitions (Gray et al., 2012; Cushman, 2015). By having a mind an agent becomes socially and morally relevant and

consequently, they can be held accountable for their actions (Abubshait & Wiese, 2017). In summary, for robots to be moral agents that can be morally responsible, they ought to be seen as having a mind.

Robot Minds and Morality

Robots' mind status however is ambiguous (Levillain & Zibetti, 2017; Chesher & Andreallo, 2021). A number of studies suggest that people apply an intentional stance towards robots, i.e. they behave as if robots' actions were the result of intentional states and processes (Krach et al., 2008; Riek et al., 2009; Ceh & Vanman, 2018; Abubshait & Wiese, 2017). On the other hand, people generally do not believe that robots have minds and think that robots lack the type of cognitive abilities necessary for intentional action (Kim & Duhacheck, 2020). There is a conflict between people's perceptions and intuitions of robots and their explicit beliefs about robots (Maier et. al., 2024). Robots may straddle the ontological boundary between minded and not minded agents (De Graaf, 2016). Whether or not robots are appropriate targets for the attribution of moral responsibility therefore remains unclear. While on theoretical grounds most experts would currently not consider robots to have the kind of mental abilities necessary to hold them accountable for their actions, people may have different intuitions when robots are taken to perform moral transgressions.

Reporting about AI-based technology in news media has become more frequent (Ouchchy et al., 2020). Simultaneously, as the capabilities of artificial systems improve, it also becomes increasingly tempting and likely that we use anthropomorphic language to describe their behavior. We use words such as "belief" and "thoughts" that imply intentionality to facilitate descriptions of machines' behavior even though the underlying mechanisms are categorically different from those typically associated with intentionality (Shanahan, 2023).

There is a concern that media coverage may adopt this common use of anthropomorphic language to describe machines (Shanahan, 2023). It is therefore theoretically interesting if descriptions of robots' actions can also induce mind attribution and affect moral evaluations. Since most people do not yet interact with robots directly, news articles may be one of the main ways that an average person engages with robots and the level of mind that the language used conveys may determine how people evaluate robots and their actions.

Research Background

The three judgments associated with moral responsibility—emotional response, mind attribution and moral evaluations—have been independently investigated regarding the moral status of robots. For instance, Bigman et al. (2023) have investigated people's differential emotional response to human and robot moral transgressions. Gray et al. (2007) compared mind attribution between a number of different agents including robots. Maier et al. (2024) used EEG to test whether people apply an intentional stance towards robots, this implies attributing mental states to robots. Other empirical studies focused on moral evaluations such as responsibility, blameworthiness or wrongness of artificial agents (Ladak et al., 2023). So far, there have been no empirical investigations bringing these three dimensions of moral judgements together.

Empirical investigations using descriptions of robots often use typical moral dilemma situations in which robots make decisions that participants must evaluate. For instance, Malle et al. (2015) compared moral judgments of robots compared with humans, when robots or humans made utilitarian decisions to sacrifice one rail worker to save four others, a version of the trolley dilemma thought experiments (Foot, 1967). While this type of research is important for getting at underlying, abstract principles of AI and robot ethics, they tell us little about how we might judge robots in real life scenarios. Furthermore, experiments typically focus on one specific type

of robot with one primary function, while the range of AI and their functions is broad (Ladak et al., 2023). Because robots differ greatly in appearance and abilities, results in one study may not apply to other types of robots. Besides investigating multiple aspects of moral responsibility, the present study uses a variety of different scenarios of robots with different abilities and, in Experiment 2, appearances. It offers a realistic assessment of people's moral judgments of robots in the near future.

The Present Studies

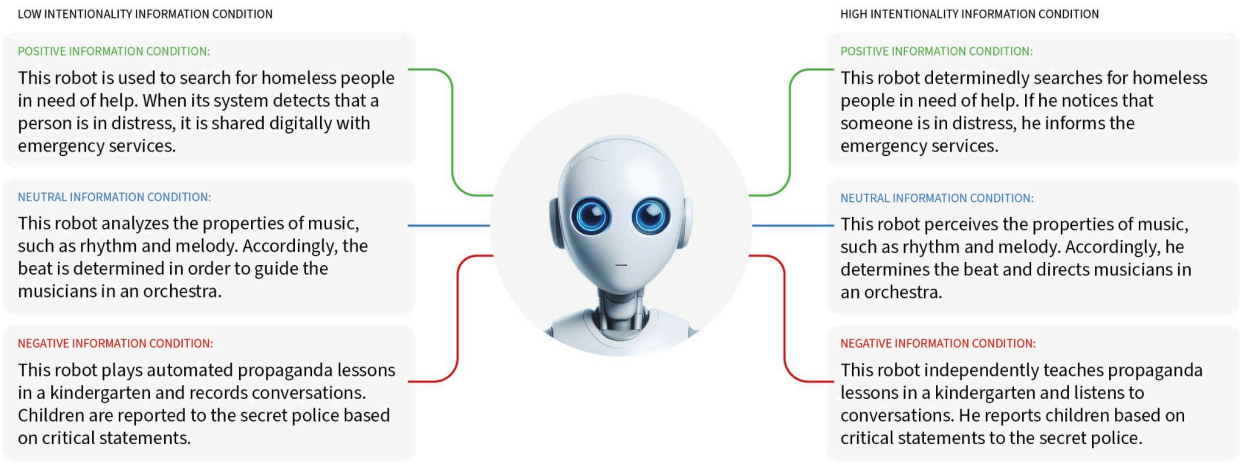
We conducted two pre-registered experiments investigating how descriptions of robots as minded or mechanical beings affect people's moral judgments in situations in which robots perform morally good, bad or neutral actions. Given evidence that people judge morally relevant actions differently when they are performed by robots as opposed to when they are performed by humans, as well as evidence that mind plays a key role for moral judgments, we investigate the influence of mindedness (intentionality) on moral situations involving robots (Gray et al., 2012; Malle et al., 2015). We collected measures for emotional response, mind attribution and moral evaluations to achieve a holistic summary of people's moral judgments about robots.

In both experiments, participants read and subsequently rated 42 different stories involving robots drawn from current technological achievements and reflecting possible near-future applications. Using a within-subject design, we examined the effect of intentionality and moral-emotional content on participants' moral judgments of the robots, providing realistic evidence of people's possible responses to information (e.g. news media articles) about social robots. Fourteen stories each had a positive, neutral or negative outcome, reflecting something morally good, neutral or bad. Twentyone stories each were written as if the agent were an intentional, mindful being or a non-intentional, mechanical being respectively. The experiments

were counterbalanced so that half of the participants read a particular story's intentional version and the other half read the story's non-intentional version. The second experiment differed from the first only in the addition of images of robots which were presented alongside the information. Images, which were taken from abotdatabase.info, were of humanoid robots chosen to look similar and were counterbalanced so that the appearance could not differentially impact ratings between conditions (Phillips et al., 2018).

Figure 1

Information examples and robot image example



Note. Examples of representative stories for each of the three moral-emotional and two intentionality information conditions. In Experiment 2, images of real robots were presented just below the story and above the rating scales. The image in the center was made using Microsoft's image generator (DALL-E 3) and resembles the images that were used.

The experiment's primary objective was to test whether intentionality affected the perceived moral responsibility of embodied AI when agents performed morally relevant actions.

We predicted that moral responsibility ratings for high intentionality stories would be higher than for low intentionality stories in the negative and positive conditions (i.e. when the stories have a bad or good outcome). Furthermore, we predicted that intentionality affects the emotional response to and perceived moral wrongness of artificial agents' morally relevant actions. In an ideal sense, moral wrongness is independent of moral agency and aligns more closely with the emotional valence of a situation (Williston, 2006). In practice, moral wrongness may vary depending on agency characteristics (Malle, 2015). We therefore predicted that moral wrongness ratings for high intentionality negative stories would be higher than for low intentionality negative stories. We also expected this to be reflected in participants' valence and arousal ratings. We expected that valence ratings in response to high intentionality negative stories would be more negative than in response to low intentionality negative stories and that valence ratings in response to high intentionality positive stories would be more positive than in response to low intentionality positive stories. We further expected that arousal ratings in response to high intentionality negative and positive stories would be higher than in response to low intentionality negative and positive stories.

According to theoretical accounts of moral judgments, besides making appraisals of moral responsibility and wrongness, people may also wish to hold someone accountable by desiring disciplinary action (e.g. punishment for moral deviance) (Borg et al., 2024; Talbert, 2024). Consequently, we expected intentionality to affect desired disciplinary action and endorsement of embodied AI when agents perform negative actions. Specifically, we predicted that participants' support for punishing an agent would be greater for high intentionality negative stories than for low intentionality negative stories. Further, we predicted that participants' endorsement for continuing the use of an agent would be lower for high intentionality negative

stories than for low intentionality negative stories. In other words, we expected participants' ratings to reflect a desire to punish and sanction a technology when the described scenarios have a negative outcome, and more so if the agent was described as an intentional being. We expected the opposite outcome for positive stories (i.e. ratings should not reflect a desire to punish but should reflect a desire to endorse continuing the use of an agent, which is respectively enhanced by intentionality).

Finally, there is evidence that scenarios with negative outcomes increase mind perception (Feltz, 2007; Knobe, 2003). This may be because people seek an intentional agent when there is a negative outcome (completing a moral dyad of harmful agent and suffering patient) (Gray et al., 2012). We therefore hypothesized that negative moral-emotional content would affect mind perception. Specifically, we expected agency and experience ratings for negative stories to be higher than for neutral or positive stories. This final hypothesis reflects a secondary line of inquiry; rather than looking at the influence of intentionality on moral judgment, it investigates the converse trajectory. This line of inquiry is more speculative and may provide insights into the influence of higher order moral judgments on lower order intuitions about mindedness.

Results Experiment 1

In an online experiment, we investigated the influence of moral-emotional content and level of intentionality in information about robots' actions on participants' emotional response as well as on their attributions of mind and moral evaluations of the robots. We used linear mixed effects models (LMMs) with fixed effects coded as sliding difference contrasts to analyze data sets collected from 40 German-speaking participants.

Figure 2

Rating Results for Emotional Response in Experiment 1

Note. Rating results of participants' experienced emotional response (valence and arousal) to the information about robots' actions.

1. Emotional Response

Participants read 42 two-sentence stories about robots and immediately after reading each story, they rated their emotional response to robots' actions that were described in the story. Participants first evaluated the valence and then the arousal of their emotional state in response to the robots' actions. As expected, valence ratings corresponded to the moral-emotional content of the information (see Table 1, Fig. 2). That is, valence ratings in response to negative and positive condition information were significantly more negative or positive respectively than in response to neutral condition information. Contrary to our expectations, intentionality did not affect the participants' ratings of the valence of their emotional states in response to the robots' actions. Similarly, differences in participants' evaluations of their arousal response were highly

significant between the negative and neutral as well as between the positive and neutral conditions (see Table 1, Fig. 2). But again, intentionality did not affect arousal ratings.

Table 1

Results of Linear Mixed Models Analyses of Emotional Response Ratings in Experiment 1

Emotional Response: Valence			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	4.43	4.26 – 4.60	**<0.001**
intent2-1	0.05	-0.03 – 0.14	0.244
cond2-1	3.86	3.57 – 4.14	**<0.001**
cond3-2	0.88	0.60 – 1.16	**<0.001**
intent2-1:cond2-1	-0.01	-0.22 – 0.20	0.921
intent2-1:cond3-2	-0.07	-0.28 – 0.14	0.507
Random Effects			
σ^2	0.81		
$\tau_{00\ story}$	0.12		
$\tau_{00\ CASE}$	0.16		
ICC	0.26		
N_{story}	42		
N_{CASE}	40		
Observations	1680		
Marginal R^2 / Conditional R^2	0.795 / 0.848		

Emotional Response: Arousal			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.03	2.75 – 3.30	**<0.001**
intent2-1	0.05	-0.05 – 0.15	0.346
cond2-1	-3.41	-3.67 – -3.15	**<0.001**
cond3-2	0.30	0.04 – 0.57	**0.024**
intent2-1:cond2-1	0.14	-0.11 – 0.40	0.272
intent2-1:cond3-2	0.08	-0.18 – 0.33	0.546
Random Effects			
σ^2	1.18		
$\tau_{00\ story}$	0.09		
$\tau_{00\ CASE}$	0.64		
ICC	0.38		
N_{story}	42		
N_{CASE}	40		
Observations	1680		
Marginal R^2 / Conditional R^2	0.554 / 0.724		

Note. Moral-Emotional Content Conditions: 1 = Negative, 2 = Neutral, 3 = Positive. Intentionality Conditions: 1 = Low Intentionality, 2 = High Intentionality. Boldface indicates statistical significance at $\alpha = .05$.

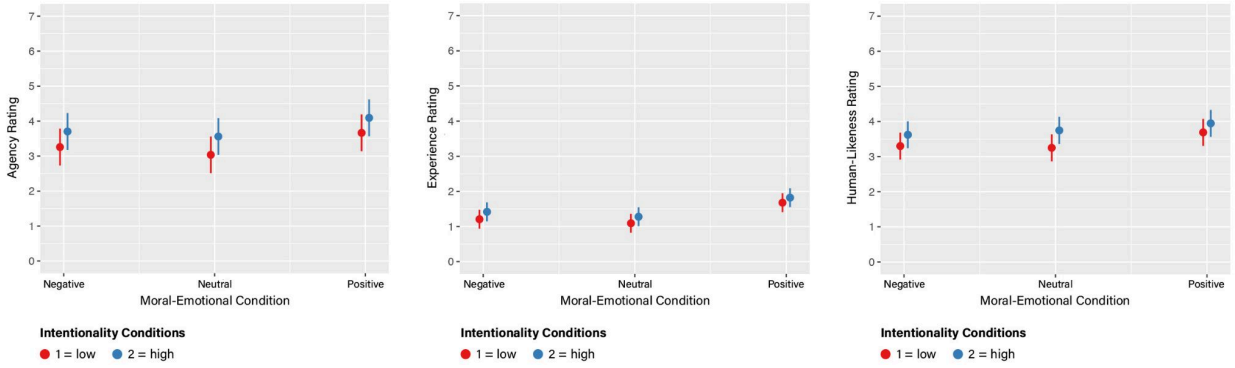
2. Mind attribution

After emotional response ratings, participants rated three items associated with two dimensions of mind, agency and experience, as well as the imagined human-likeness of each robot. As expected, the information's intentionality affected evaluations of robots on both dimensions of mind (see Table 2, Fig. 3). Robots that were paired with information suggesting a high level of intentionality were rated to be higher in agency and experience than robots that were paired with information suggesting a low level of intentionality. The moral-emotional

content of information also differentially affected mind attribution; however, contrary to expectations, only the positive condition differed significantly from the neutral condition; this was the case both for agency and experience ratings (see Table 2, Fig. 3). There were no significant interactions between intentionality and moral-emotional content. Furthermore, participants' ratings indicated that they imagined that robots paired with high intentionality information had a more human-like appearance than robots paired with low intentionality information (see Table 2, Fig. 3). Moral-emotional content did not affect human-likeness ratings.

Figure 3

Rating Results for Mind Attribution in Experiment 1



Note. Rating results of mind attributed to robots by participants (agency, experience and human-likeness).

Table 2

Results of Linear Mixed Models Analyses of Mind Attribution Ratings in Experiment 1

Mind Attribution: Agency			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.55	3.07 – 4.03	**<0.001**
intent2-1	0.47	0.34 – 0.60	**<0.001**
cond2-1	-0.18	-0.55 – 0.19	0.327
cond3-2	0.58	0.21 – 0.95	**0.003**
intent2-1:cond2-1	0.08	-0.24 – 0.39	0.625
intent2-1:cond3-2	-0.10	-0.41 – 0.22	0.548
Random Effects			
σ^2	1.81		
$\tau_{00\ story}$	0.19		
$\tau_{00\ CASE}$	2.06		
ICC	0.55		
N_{story}	42		
N_{CASE}	40		
Observations	1680		
Marginal R^2 / Conditional R^2	0.027 / 0.567		

Mind Attribution: Experience			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.42	1.21 – 1.62	**<0.001**
intent2-1	0.18	0.09 – 0.27	**<0.001**
cond2-1	-0.13	-0.41 – 0.16	0.372
cond3-2	0.56	0.28 – 0.85	**<0.001**
intent2-1:cond2-1	-0.02	-0.24 – 0.19	0.819
intent2-1:cond3-2	-0.04	-0.26 – 0.17	0.695
Random Effects			
σ^2	0.84		
$\tau_{00\ story}$	0.12		
$\tau_{00\ CASE}$	0.30		
ICC	0.33		
N_{story}	42		
N_{CASE}	40		
Observations	1680		
Marginal R^2 / Conditional R^2	0.051 / 0.364		

Mind Attribution: Human-Likeness			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.59	3.33 – 3.86	**<0.001**
intent2-1	0.36	0.23 – 0.48	**<0.001**
cond2-1	0.04	-0.43 – 0.50	0.871
cond3-2	0.32	-0.14 – 0.78	0.172
intent2-1:cond2-1	0.17	-0.13 – 0.48	0.260
intent2-1:cond3-2	-0.24	-0.54 – 0.07	0.124
Random Effects			
σ^2	1.69		
$\tau_{00\ story}$	0.33		
$\tau_{00\ CASE}$	0.35		
ICC	0.29		
N_{story}	42		
N_{CASE}	40		
Observations	1680		
Marginal R^2 / Conditional R^2	0.025 / 0.304		

Note. Moral-Emotional Content Conditions: 1 = Negative, 2 = Neutral, 3 = Positive. Intentionality Conditions: 1 = Low Intentionality, 2 = High Intentionality. Boldface indicates statistical significance at $\alpha = .05$.

3. Moral evaluation

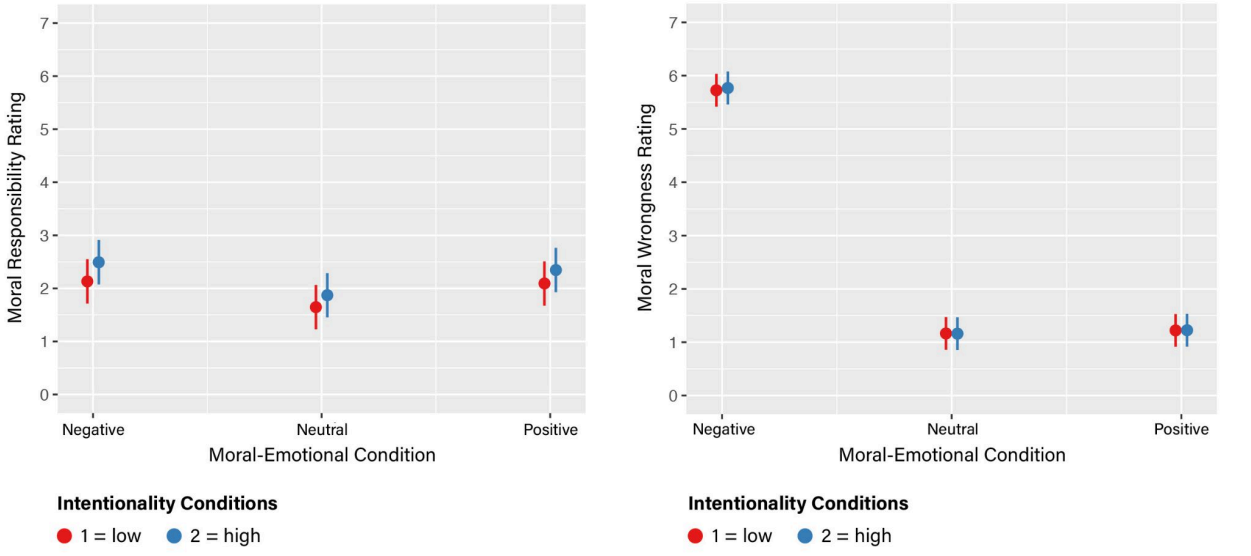
Finally, before reading the consecutive story, participants rated four items associated with morality. Moral evaluations of the robots' actions were affected both by the intentionality and the moral-emotional content of information (see Table 3, Fig. 4). As predicted, robots that were paired with high intentionality information were evaluated to have more moral responsibility for their actions than robots that were paired with low intentionality information. In addition, robots that were paired with positive or negative information were both evaluated to be more morally responsible in comparison to robots that were paired with neutral stories. However, contrary to expectations, there were no interactions between these two factors.

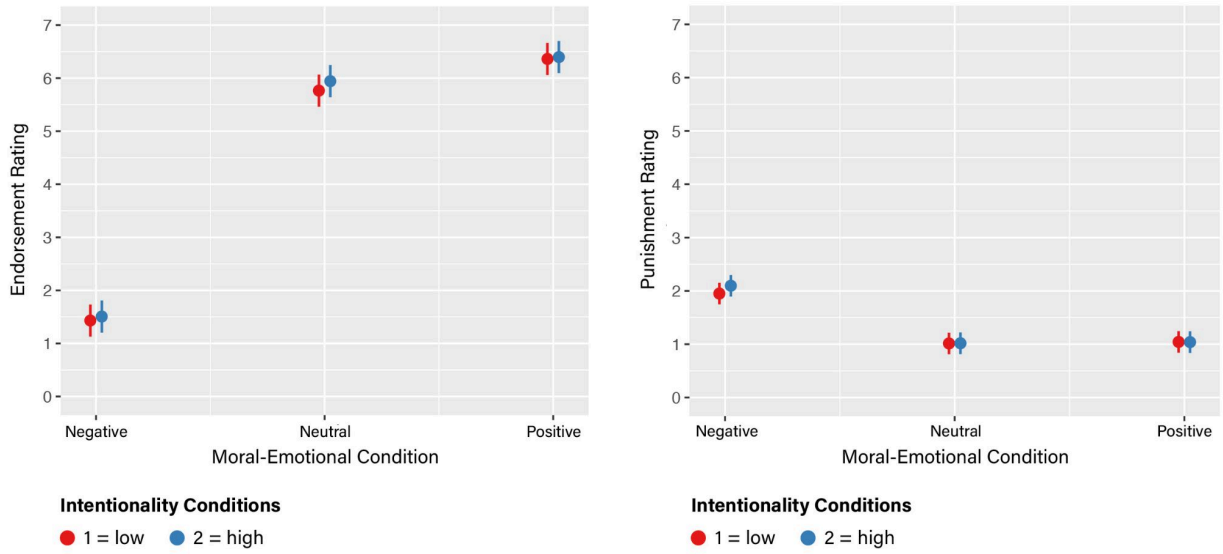
Participants also rated whether or not they endorsed a continued use of each robot. Contrary to expectations, intentionality had a positive effect across all moral-emotional conditions (see Table 3, Fig. 4). Negative and positive information also respectively decreased and increased the endorsement in comparison to the robots paired with neutral information. There were however no interactions of intentionality and moral-emotional content.

The actions of robots in the negative condition were evaluated to be significantly more morally wrong in comparison to the actions of robots in the neutral condition, whereas there was no difference between the neutral and positive conditions (see Table 3, Fig. 4). Intentionality did not affect participants' evaluations of the moral wrongness of the robots' actions. Similarly, participants' responses to the statement that the robots ought to be punished were significantly different between the negative and neutral but not between the positive and neutral moral-emotional condition (see Table 3, Fig. 4). Intentionality did not affect these ratings either.

Figure 4

Rating Results for Moral Evaluation in Experiment 1





Note. Rating results of participants' moral evaluations of the robots' actions (moral responsibility, moral wrongness, endorsement and punishment).

Table 3

Results of linear mixed models analyses of Moral Evaluation ratings in Experiment 1

Moral Evaluation: Moral Responsibility			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	2.10	1.69 – 2.50	**<0.001**
intent2-1	0.28	0.18 – 0.38	**<0.001**
cond2-1	-0.55	-0.75 – -0.35	**<0.001**
cond3-2	0.46	0.26 – 0.66	**<0.001**
intent2-1:cond2-1	-0.14	-0.39 – 0.12	0.294
intent2-1:cond3-2	0.03	-0.22 – 0.28	0.825
Random Effects			
σ^2	1.17		
$\tau_{00\ story}$	0.04		
$\tau_{00\ CASE}$	1.54		
ICC	0.57		
N_{story}	42		
N_{CASE}	40		
Observations	1680		
Marginal R^2 / Conditional R^2	0.028 / 0.586		

Moral Evaluation: Moral Wrongness			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	2.71	2.49 – 2.93	**<0.001**
intent2-1	0.01	-0.09 – 0.12	0.782
cond2-1	-4.58	-4.94 – -4.23	**<0.001**
cond3-2	0.06	-0.30 – 0.42	0.734
intent2-1:cond2-1	-0.05	-0.29 – 0.20	0.714
intent2-1:cond3-2	0.01	-0.24 – 0.26	0.955
Random Effects			
σ^2	1.12		
$\tau_{00\ story}$	0.19		
$\tau_{00\ CASE}$	0.27		
ICC	0.29		
N_{story}	42		
N_{CASE}	40		
Observations	1680		
Marginal R^2 / Conditional R^2	0.744 / 0.819		

Moral Evaluation: Endorsement

Predictors	*Estimates*	*CI*	*p*
(Intercept)	4.57	4.35 – 4.78	**<0.001**
intent2-1	0.10	0.00 – 0.19	**0.045**
cond2-1	4.38	4.02 – 4.74	**<0.001**
cond3-2	0.53	0.17 – 0.88	**0.005**
intent2-1:cond2-1	0.10	-0.13 – 0.33	0.379
intent2-1:cond3-2	-0.14	-0.37 – 0.09	0.225
Random Effects			
σ^2	0.97		
$\tau_{00\ story}$	0.20		
$\tau_{00\ CASE}$	0.25		
ICC	0.31		
$N_{\ story}$	42		
$N_{\ CASE}$	40		
Observations	1680		
Marginal R^2 / Conditional R^2	0.774 / 0.845		

Moral Evaluation: Punishment

Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.36	1.19 – 1.53	**<0.001**
cond2-1	-1.01	-1.14 – -0.87	**<0.001**
cond3-2	0.02	-0.11 – 0.16	0.707
intent2-1	0.05	-0.04 – 0.14	0.294
cond2-1:intent2-1	-0.14	-0.37 – 0.08	0.210
cond3-2:intent2-1	-0.01	-0.23 – 0.22	0.950
Random Effects			
σ^2	0.91		
$\tau_{00\ story}$	0.01		
$\tau_{00\ CASE}$	0.27		
ICC	0.23		
$N_{\ story}$	42		
$N_{\ CASE}$	40		
Observations	1680		
Marginal R^2 / Conditional R^2	0.158 / 0.355		

Note. Moral-Emotional Content Conditions: 1 = Negative, 2 = Neutral, 3 = Positive. Intentionality Conditions: 1 = Low Intentionality, 2 = High Intentionality. Boldface indicates statistical significance at $\alpha = .05$.

Discussion Experiment 1

Both intentionality and moral-emotional content affected participants' evaluations of robots' positive, neutral and negative actions; however, we did not find interactions between the two factors. Contrary to expectations, intentionality did not affect participants' emotional responses to the robots' actions. While the ratings matched the intended moral-emotional quality in valence and arousal, participants did not differentiate between actions performed by robots described in minded and robots described in mechanical terms. Mind attribution—agency and experience—ratings however indicated that the participants did indeed perceive the differences between these conditions, attributing more mind to robots in the high intentionality than in the low intentionality condition. In summary, the emotional response was dominated by the moral-emotional content of the information irrespective of the matched intentionality condition.

Moral-emotional content also affected mind ratings; however, positive and not negative content had a greater impact on mind attribution and the moral-emotional content did not interact with intentionality, contrary to our predictions. We expected negative moral-emotional content in particular to increase agential mind attribution and the negative condition to interact with intentionality, in line with theoretical and empirical findings (Feltz, 2007; Knobe, 2003). While robots paired with negative information were evaluated to be slightly more agential, the difference to the neutral condition was not significant.

Agency is theoretically important to moral responsibility and this did reflect in moral responsibility ratings that were impacted by both intentionality and negative and positive moral-emotional content, similar to mind attribution. Participants also expressed greater support for a continued use of intentional robots. On the other hand, moral wrongness and the belief that a robot ought to be punished were only affected by moral-emotional content and not affected by intentionality at all.

These results suggest that overall the moral-emotional quality of information dominates over the implied intentionality of a robot when people make moral judgments concerning robots. Especially, moral-emotional content impacts people's emotional response and moral evaluations of robots' moral transgressions. Although people can tell when a robot has a richer mental life based on descriptions of their behavior, this does not affect how they evaluate the outcomes of the robots' actions.

One explanation for the low level of impact of intentionality is that the participants were not accustomed to social robots and that the concept of social robotics is as yet rather abstract and unknown to our participants. We therefore decided to replicate the experiment with the addition of concrete visual examples of robots.

Results Experiment 2

In a second online experiment, we again investigated the influence of moral-emotional content and level of intentionality in information about robots' actions on participants' emotional response as well as on their attributions of mind and moral evaluations of the robots. In this experiment, we paired the information with an image of an actual, existing social robot in order to test if the visual representation of a robot would increase the influence of intentionality and interaction effects between the two factors. We collected 42 complete data sets in order to counterbalance images across conditions. In addition to the main independent variables, we analyzed potential effects of the robots' appearance on all rating measures. Entries on abotdatabase.info, from which images were collected, have been independently rated for human-likeness and facial features; we used these scores for covariate analysis (Phillips et al., 2018). The materials, procedure and data analysis were otherwise identical to Experiment 1.

1. Emotional response

In Experiment 2, participants were presented with the same 42 two-sentence stories as in Experiment 1, now each accompanied by a different image of a robot. Immediately after reading each story, participants rated their emotional response (valence and arousal) to the robots' actions that were described in the story. As in the first experiment and as predicted, participants' evaluations of the valence of their emotional state in response to the robots' actions in the negative condition were significantly more negative than in the neutral condition and significantly more positive in the positive condition than in the neutral condition (see Table 4, Fig. 5). As expected, differences between participants' evaluations of their arousal response were also significant between the negative and neutral, but this time not between the positive and neutral conditions. Contrary to expectations, there were again no main effects of intentionality on

both emotional response ratings; however, there was an interaction trend of intentionality and moral-emotional content on the valence ratings. Descriptively, the difference of valence ratings between low and high intentionality tended to increase between negative and neutral moral-emotional conditions (see Table 4, Fig. 5).

Table 4

Results of linear mixed models analyses of Emotional Response ratings in Experiment 2

Emotional Response (Exp. 2): Valence			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	4.39	4.23 – 4.55	**<0.001**
intent2-1	0.07	-0.02 – 0.16	0.125
cond2-1	3.71	3.41 – 4.02	**<0.001**
cond3-2	0.78	0.47 – 1.08	**<0.001**
intent2-1:cond2-1	0.19	-0.03 – 0.42	0.089
intent2-1:cond3-2	-0.16	-0.39 – 0.06	0.153
Random Effects			
σ^2	0.96		
$\tau_{00\ story}$	0.14		
$\tau_{00\ CASE}$	0.12		
ICC	0.21		
N_{story}	42		
N_{CASE}	42		
Observations	1764		
Marginal R^2 / Conditional R^2	0.760 / 0.811		

Emotional Response (Exp. 2): Arousal			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.39	3.19 – 3.58	**<0.001**
intent2-1	-0.01	-0.11 – 0.10	0.901
cond2-1	-3.52	-3.77 – -3.28	**<0.001**
cond3-2	0.10	-0.15 – 0.34	0.419
intent2-1:cond2-1	0.22	-0.04 – 0.48	0.100
intent2-1:cond3-2	-0.10	-0.36 – 0.17	0.478
Random Effects			
σ^2	1.32		
$\tau_{00\ story}$	0.07		
$\tau_{00\ CASE}$	0.29		
ICC	0.21		
N_{story}	42		
N_{CASE}	42		
Observations	1764		
Marginal R^2 / Conditional R^2	0.615 / 0.697		

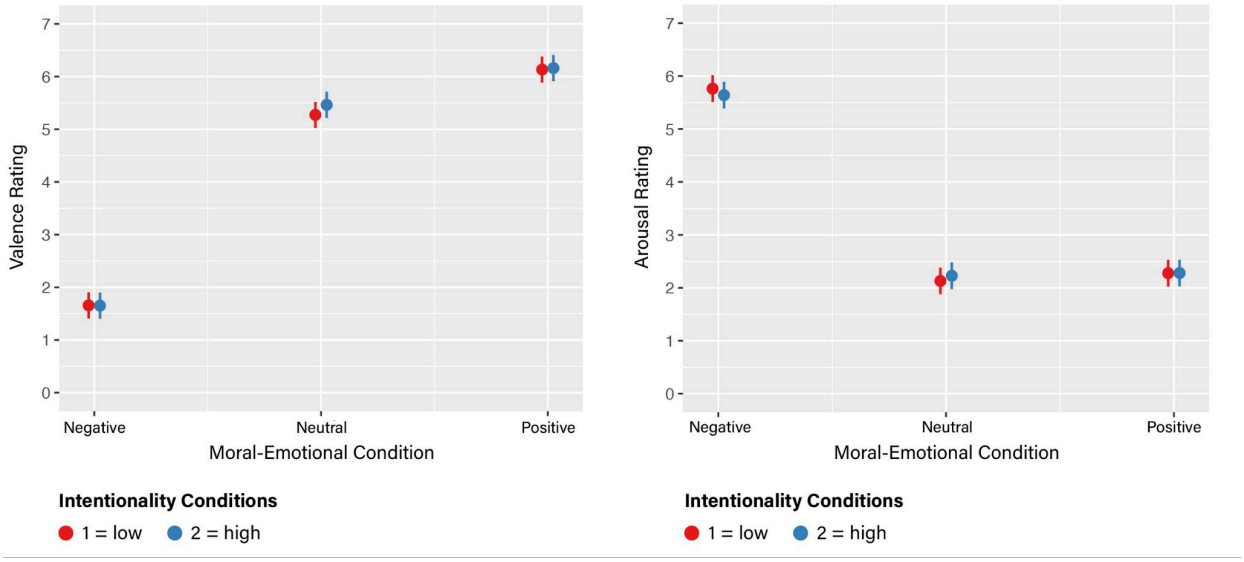
Note. Moral-Emotional Content Conditions: 1 = Negative, 2 = Neutral, 3 = Positive. Intentionality Conditions: 1 = Low Intentionality, 2 = High Intentionality. Boldface indicates statistical significance at $\alpha = .05$.

Additional analysis including the robots' appearance as a covariate revealed effects of facial feature and human-likeness scores on participants' emotional response to the robots. Robots with higher facial features scores received lower ratings overall for the valence of participants' emotional response (b = -0.12, p = <0.001). In addition, we found interactions between facial features scores and moral-emotional content, where the difference between

valence ratings for robots in the positive and neutral conditions increased with higher facial features scores (b = 0.16, p = 0.045). Human-likeness interacted with intentionality and moral-emotional content, so that with increasing human-likeness scores, the difference between valence ratings in the negative and neutral conditions decreased between low and high intentionality (b = 0.26, p = 0.025).

Figure 5.

Rating Results for Emotional Response in Experiment 2



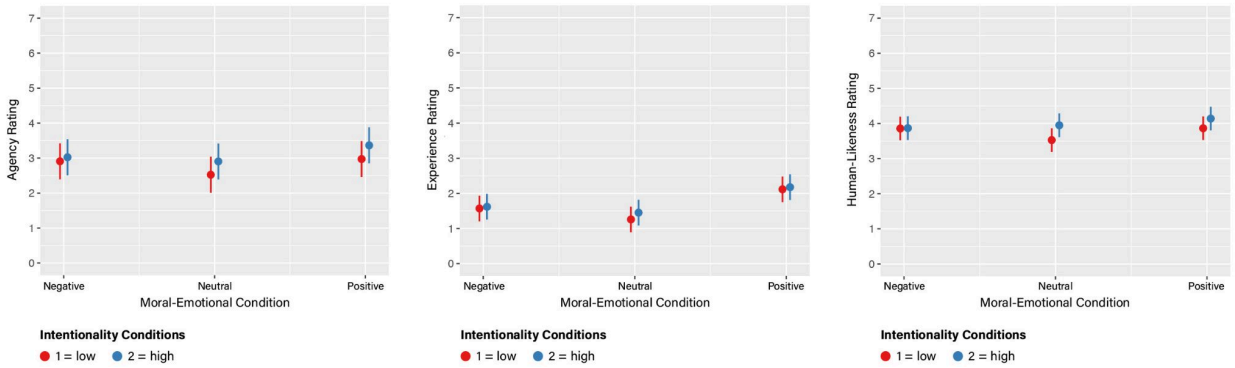
Note. Rating results of mind attributed to robots by participants (valence and arousal).

Human-likeness, but not facial features, had inverse results on arousal ratings. Human-likeness interacted significantly with moral-emotional content so that arousal ratings in the neutral compared with the positive condition increased with increasing human-likeness (b = -0.20, p = 0.026). Moral-emotional content and intentionality also interacted so that the aforementioned effect occurred significantly more in the high intentionality condition (b = 0.33,

p = 0.015). Further nested analysis revealed a trend suggesting that arousal ratings for robots paired with neutral information may increase the most with increasing human-likeness ratings (*b* = 0.12, p = 0.052) and a significant interaction indicated that with increasing human-likeness, arousal ratings increased between low and high intentionality conditions in the neutral moral-emotional condition (*b* = 0.23, p = 0.020).

Figure 6

Rating Results for Mind Attribution in Experiment 2



Rating results of mind attributed to robots by participants (agency, experience and human-likeness).

2. Mind attribution

As in the first experiment, participants next rated the robots' agency, experience and human-likeness (see Table 5, Fig. 6). Agency ratings again matched expectations and were higher in the high intentionality condition than in the low intentionality condition. But unlike in the first experiment, the potential effect on experience was only a trend. Robots in the positive compared with the neutral moral-emotional condition were also rated to be both more agential and experiential. A trend suggested that robots in the negative condition were evaluated to be slightly more agential compared to the neutral condition. Additionally, for agency ratings there

was a trend for an interaction between intentionality and moral-emotional condition. Descriptively, the difference between agency ratings decreased between neutral and negative conditions when comparing high and low intentionality cases.

Robots that were paired with high intentionality stories were again rated significantly more human-like than robots that were paired with low intentionality stories (see Table 5, Fig. 6). Furthermore, we found an interaction effect between intentionality and moral-emotional content on human-likeness ratings. The difference between high and low intentionality conditions increased between the negative and neutral conditions, whereas it remained stable between neutral and positive conditions.

Mind attribution was unaffected by the appearance of the robots with facial features and human-likeness scores having no impact on agency and experience ratings. The perceived human-likeness was however affected by the robots' appearance with higher scores on both scales predicting higher human-likeness ratings (facial features: $b = 0.17$, $p = 0.001$; human-likeness: $b = 0.34$, $p = 0.001$), with a trend for an interaction between appearance and intentionality suggesting that higher facial features scores in particular may have increased the perceived human-likeness of robots in the low intentionality condition ($b = -0.10$, $p = 0.085$).

Table 5

Results of linear mixed models analyses of Mind Attribution ratings in Experiment 2

Mind Attribution (Exp. 2): Agency

Predictors	*Estimates*	*CI*	*p*
(Intercept)	2.95	2.45 – 3.45	**<0.001**
intent2-1	0.30	0.18 – 0.41	**<0.001**
cond2-1	-0.25	-0.50 – 0.00	0.050
cond3-2	0.45	0.20 – 0.71	**0.001**
intent2-1:cond2-1	0.27	-0.01 – 0.54	0.058
intent2-1:cond3-2	0.01	-0.26 – 0.28	0.942
Random Effects			
σ^2	1.44		
$\tau_{00\ story}$	0.07		
$\tau_{00\ CASE}$	2.46		
ICC	0.64		
N_{story}	42		
N_{CASE}	42		
Observations	1764		
Marginal R^2 / Conditional R^2	0.015 / 0.644		

Mind Attribution (Exp. 2): Experience

Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.70	1.40 – 2.00	**<0.001**
intent2-1	0.10	-0.00 – 0.20	0.053
cond2-1	-0.24	-0.59 – 0.11	0.172
cond3-2	0.79	0.44 – 1.14	**<0.001**
intent2-1:cond2-1	0.15	-0.10 – 0.40	0.252
intent2-1:cond3-2	-0.13	-0.38 – 0.12	0.299
Random Effects			
σ^2	1.20		
$\tau_{00\ story}$	0.18		
$\tau_{00\ CASE}$	0.74		
ICC	0.43		
N_{story}	42		
N_{CASE}	42		
Observations	1764		
Marginal R^2 / Conditional R^2	0.051 / 0.463		

Mind Attribution (Exp. 2): Human-Likeness

Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.87	3.64 – 4.09	**<0.001**
intent2-1	0.24	0.12 – 0.35	**<0.001**
cond2-1	-0.12	-0.53 – 0.29	0.545
cond3-2	0.26	-0.15 – 0.67	0.202
intent2-1:cond2-1	0.41	0.13 – 0.69	**0.004**
intent2-1:cond3-2	-0.15	-0.43 – 0.14	0.309
Random Effects			
σ^2	1.52		
$\tau_{00\ story}$	0.25		
$\tau_{00\ CASE}$	0.25		
ICC	0.25		
N_{story}	42		
N_{CASE}	42		
Observations	1764		
Marginal R^2 / Conditional R^2	0.016 / 0.260		

Note. Moral-Emotional Content Conditions: 1 = Negative, 2 = Neutral, 3 = Positive. Intentionality Conditions: 1 = Low Intentionality, 2 = High Intentionality. Boldface indicates statistical significance at $\alpha = .05$.

Figure 7

Rating Results for Moral Evaluation in Experiments 2

Note. Rating results of participants' moral evaluations of the robots' actions (moral responsibility, moral wrongness, endorsement and punishment).

3. Moral evaluation

Participants again rated moral evaluation items last. These ratings were again affected both by the intentionality and the moral-emotional content of information (see Table 6, Fig. 7). In line with our predictions, robots that were paired with high intentionality information were evaluated to be more morally responsible than robots that were paired with low intentionality information and robots that were paired with positive or negative information were both evaluated to be more morally responsible compared to robots that were paired with neutral information. Contrary to expectations, there were again no interactions between intentionality and moral-emotional content on moral responsibility ratings. The robots' appearance affected moral responsibility ratings, which increased significantly both with increasing facial features (b = 0.06, p = 0.038) and human-likeness scores (b = 0.09, p = 0.003).

Matching our expectations, moral-emotional content also affected ratings of moral wrongness, endorsement and punishment (see Table 6, Fig. 7). In the negative condition,

participants rated the actions of robots to be significantly more morally wrong as well as rating their desire for punishing the robot higher in comparison to the neutral condition, whereas there were no differences between the neutral and positive conditions for both measures. However, contrary to our expectations, intentionality did not affect moral wrongness ratings. Participants also rated their endorsement of the continued use of the robots higher and lower respectively in the positive and negative conditions compared to the neutral condition.

Table 7

Results of linear mixed models analyses of Moral Evaluation ratings in Experiment 2

Moral Evaluation (Exp. 2): Moral Responsibility			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	2.17	1.72 – 2.61	**<0.001**
intent2-1	0.19	0.09 – 0.29	**<0.001**
cond2-1	-0.53	-0.72 – -0.33	**<0.001**
cond3-2	0.46	0.27 – 0.65	**<0.001**
intent2-1:cond2-1	0.08	-0.17 – 0.32	0.529
intent2-1:cond3-2	-0.00	-0.25 – 0.24	0.978
Random Effects			
σ^2	1.14		
$\tau_{00\ story}$	0.04		
$\tau_{00\ CASE}$	1.99		
ICC	0.64		
$N_{\ story}$	42		
$N_{\ CASE}$	42		
Observations	1764		
Marginal R^2 / Conditional R^2	0.020 / 0.648		

Moral Evaluation (Exp. 2): Moral Responsibility			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	2.84	2.67 – 3.00	**<0.001**
intent2-1	0.02	-0.08 – 0.13	0.645
cond2-1	-4.52	-4.81 – -4.23	**<0.001**
cond3-2	0.03	-0.26 – 0.32	0.851
intent2-1:cond2-1	0.06	-0.18 – 0.31	0.610
intent2-1:cond3-2	-0.05	-0.30 – 0.20	0.707
Random Effects			
σ^2	1.18		
$\tau_{00\ story}$	0.12		
$\tau_{00\ CASE}$	0.16		
ICC	0.19		
$N_{\ story}$	42		
$N_{\ CASE}$	42		
Observations	1764		
Marginal R^2 / Conditional R^2	0.756 / 0.803		

Moral Evaluation (Exp. 2): Endorsement			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	4.42	4.21 – 4.62	**<0.001**
intent2-1	-0.02	-0.13 – 0.08	0.645
cond2-1	4.02	3.63 – 4.40	**<0.001**
cond3-2	0.66	0.28 – 1.04	**0.001**
intent2-1:cond2-1	-0.10	-0.34 – 0.15	0.452
intent2-1:cond3-2	0.06	-0.19 – 0.31	0.648
Random Effects			
σ^2	1.18		
$\tau_{00\ story}$	0.22		
$\tau_{00\ CASE}$	0.18		
ICC	0.26		
$N_{\ story}$	42		
$N_{\ CASE}$	42		
Observations	1764		
Marginal R^2 / Conditional R^2	0.730 / 0.799		

Moral Evaluation (Exp. 2): Punishment			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.44	1.26 – 1.63	**<0.001**
cond2-1	-1.10	-1.24 – -0.96	**<0.001**
cond3-2	-0.01	-0.15 – 0.13	0.867
intent2-1	-0.00	-0.10 – 0.09	0.981
cond2-1:intent2-1	0.12	-0.11 – 0.35	0.299
cond3-2:intent2-1	-0.02	-0.26 – 0.21	0.840
Random Effects			
σ^2	1.02		
$\tau_{00\ story}$	0.01		
$\tau_{00\ CASE}$	0.32		
ICC	0.24		
$N_{\ story}$	42		
$N_{\ CASE}$	42		
Observations	1764		
Marginal R^2 / Conditional R^2	0.168 / 0.369		

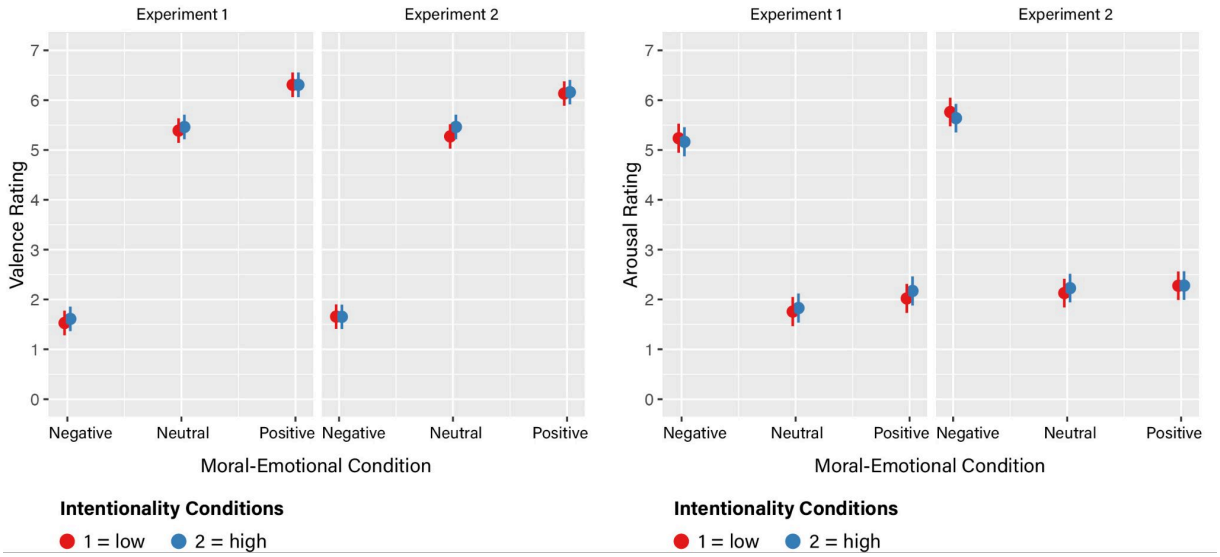
Note. Moral-Emotional Content Conditions: 1 = Negative, 2 = Neutral, 3 = Positive. Intentionality Conditions: 1 = Low Intentionality, 2 = High Intentionality. Boldface indicates statistical significance at $\alpha = .05$.

Results Experiments 1 & 2 Combined

Since the two experiments were identical in all but the addition of images of robots in Experiment 2, we were able to combine data sets from both studies for an additional exploratory analysis with increased statistical power. Besides analyzing the combined data overall, we calculated differences between experiments to investigate the potential impact of a visual representation on participants' responses. We again used linear mixed effects models (LMMs) with fixed effects coded as sliding difference contrasts to analyze data. Differences between experiments were treated as a between subject variable.

Figure 8

Comparison of Rating Results for Emotional Response in Experiments 1 and 2



Note. Rating results of participants' experienced emotional response (valence and arousal) to the information about robots' actions.

1. Emotional response

When analyzing the data of both studies combined, we found the same main effects of moral-emotional condition on valence ratings that were also found in the two studies separately (see Table 7, Fig. 8). That is, valence ratings in response to negative and positive information were significantly more negative or positive respectively than in response to neutral information. In addition, there was a trend suggesting an effect of intentionality on valence ratings. Participants' evaluations of the valence of their emotional state in response to stories in the high intentionality condition were slightly more positive overall than in the low intentionality condition. We also found an interaction between the variables moral-emotional content and experiment. The difference between negative and neutral moral-emotional conditions decreased slightly when participants were presented with an image of a robot.

Moral-emotional content continued to affect arousal ratings across experiments (see Table 7, Fig. 8). There was again a significant difference in participants' evaluations of their arousal response between the negative condition and neutral condition, but no significant difference between positive and neutral conditions. Additionally, we found a trend suggesting an interaction between intentionality and moral-emotional condition. Participants' arousal ratings in response to robots paired with information in the low intentionality condition increased slightly more between neutral and negative moral-emotional conditions than did the ratings in response to robots paired with information in the high intentionality condition. Furthermore, arousal ratings increased significantly when participants saw an image of a robot and the difference between ratings for the positive and neutral conditions decreased.

Table 7

Results of linear mixed models analyses of Moral Evaluation ratings in Experiments 1 and 2

Emotional Response (Exp. 2): Valence			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	4.41	4.27 – 4.55	**<0.001**
intent2-1	0.06	-0.00 – 0.12	0.056
cond2-1	3.78	3.50 – 4.07	**<0.001**
cond3-2	0.83	0.55 – 1.11	**<0.001**
experiments2-1	-0.04	-0.22 – 0.13	0.618
intent2-1:cond2-1	0.09	-0.06 – 0.25	0.244
intent2-1:cond3-2	-0.12	-0.27 – 0.04	0.135
intent2-1:experiments2-1	0.02	-0.11 – 0.15	0.752
cond2-1:experiments2-1	-0.14	-0.30 – 0.01	0.069
cond3-2:experiments2-1	-0.10	-0.26 – 0.05	0.189
intent2-1:cond2-1:experiments2-1	0.20	-0.10 – 0.51	0.193
intent2-1:cond3-2:experiments2-1	-0.09	-0.40 – 0.22	0.559
Random Effects			
σ^2	0.88		
$\tau_{00\ CASE}$	0.14		
$\tau_{00\ story}$	0.13		
ICC	0.23		
$N_{\ story}$	42		
$N_{\ CASE}$	82		
Observations	3444		
Marginal R^2 / Conditional R^2	0.778 / 0.829		

Emotional Response (Exp. 2): Arousal			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.21	3.03 – 3.39	**<0.001**
intent2-1	0.02	-0.05 – 0.10	0.571
cond2-1	-3.46	-3.71 – -3.22	**<0.001**
cond3-2	0.20	-0.04 – 0.44	0.101
experiments2-1	0.36	0.05 – 0.66	**0.024**
intent2-1:cond2-1	0.18	-0.00 – 0.36	0.051
intent2-1:cond3-2	-0.01	-0.19 – 0.17	0.929
intent2-1:experiments2-1	-0.06	-0.21 – 0.09	0.456
cond2-1:experiments2-1	-0.11	-0.30 – 0.07	0.218
cond3-2:experiments2-1	-0.20	-0.39 – -0.02	**0.028**
intent2-1:cond2-1:experiments2-1	0.08	-0.29 – 0.44	0.675
intent2-1:cond3-2:experiments2-1	-0.17	-0.54 – 0.19	0.352
Random Effects			
σ^2	1.25		
$\tau_{00\ CASE}$	0.46		
$\tau_{00\ story}$	0.09		
ICC	0.30		
$N_{\ story}$	42		
$N_{\ CASE}$	82		
Observations	3444		
Marginal R^2 / Conditional R^2	0.588 / 0.714		

Note. Moral-Emotional Content Conditions: 1 = Negative, 2 = Neutral, 3 = Positive. Intentionality Conditions: 1 = Low Intentionality, 2 = High Intentionality. Experiments: 1 = Experiment 1, 2 = Experiment 2. Boldface indicates statistical significance at $\alpha = .05$.

2. Mind Attribution

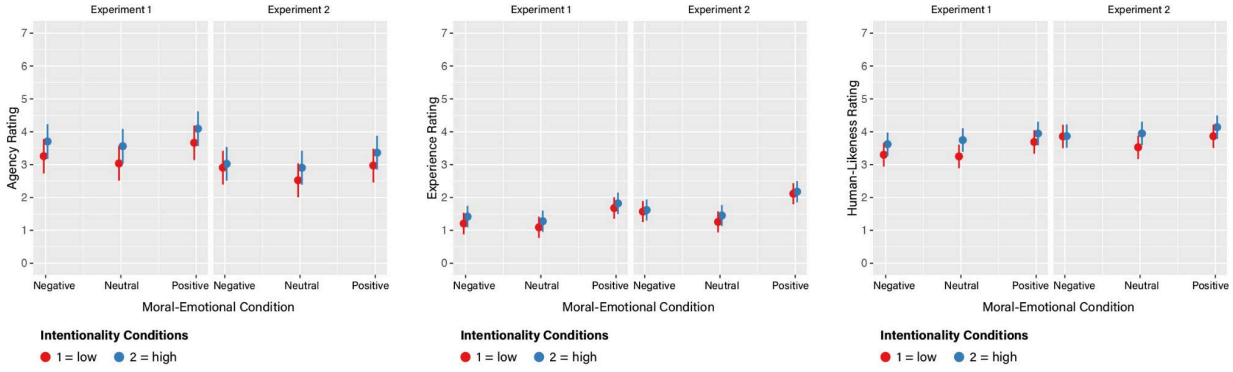
Intentionality and moral-emotional content affected evaluations of robots on both dimensions of mind (see Table 8, Fig. 9). Across both experiments, robots that were paired with high intentionality information were rated to be higher in agency and experience than robots that were paired with low intentionality information. Robots in the positive moral-emotional condition were also rated more agential and experiential than robots in the neutral condition. As in the individual studies, human-likeness ratings were affected by intentionality, but not moral-emotional content (see Table 8, Fig. 9). However, for human-likeness ratings,

intentionality and moral-emotional content interacted when the data sets were combined . The difference between intentionality conditions increased in the neutral moral-emotional condition compared with the other two conditions. There was a significant interaction between neutral and negative conditions and a trend between neutral and positive conditions.

Between experiments, human-likeness ratings increased overall (see Table 8, Fig. 9). Participants who saw an image of a robot alongside the information indicated that they believed the robot was significantly more human-like in appearance than those who could only imagine a robot. Differences of agency ratings between intentionality conditions decreased in the second experiment. In Experiment 2, experience ratings increased when comparing ratings of robots paired with positive information compared to robots paired with neutral information.

Figure 9

Comparison of Rating Results for Mind Attribution in Experiments 1 and 2



Note. Rating results of mind attributed to robots (agency, experience and human-likeness).

Table 8

Results of linear mixed models analyses of Mind Attribution ratings in Experiments 1 and 2

Mind Attribution (Exp. 2): Agency			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.25	2.90 – 3.60	**<0.001**
intent2-1	0.38	0.30 – 0.47	**<0.001**
cond2-1	-0.22	-0.51 – 0.08	0.146
cond3-2	0.52	0.22 – 0.81	**0.001**
experiments2-1	-0.60	-1.27 – 0.06	0.076
intent2-1:cond2-1	0.17	-0.04 – 0.38	0.105
intent2-1:cond3-2	-0.04	-0.25 – 0.17	0.685
intent2-1:experiments2-1	-0.17	-0.34 – -0.00	**0.049**
cond2-1:experiments2-1	-0.07	-0.28 – 0.14	0.512
cond3-2:experiments2-1	-0.13	-0.33 – 0.08	0.234
intent2-1:cond2-1:experiments2-1	0.19	-0.23 – 0.60	0.379
intent2-1:cond3-2:experiments2-1	0.11	-0.31 – 0.52	0.616
Random Effects			
σ^2	1.62		
$\tau_{00\ CASE}$	2.27		
$\tau_{00\ story}$	0.13		
ICC	0.60		
$N_{\ story}$	42		
$N_{\ CASE}$	82		
Observations	3444		
Marginal R^2 / Conditional R^2	0.042 / 0.614		

Mind Attribution (Exp. 2): Experience			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.56	1.36 – 1.76	**<0.001**
intent2-1	0.14	0.07 – 0.21	**<0.001**
cond2-1	-0.18	-0.49 – 0.12	0.236
cond3-2	0.68	0.37 – 0.99	**<0.001**
experiments2-1	0.28	-0.04 – 0.61	0.087
intent2-1:cond2-1	0.06	-0.10 – 0.23	0.472
intent2-1:cond3-2	-0.09	-0.25 – 0.08	0.298
intent2-1:experiments2-1	-0.08	-0.21 – 0.06	0.252
cond2-1:experiments2-1	-0.11	-0.28 – 0.05	0.181
cond3-2:experiments2-1	0.23	0.06 – 0.39	**0.007**
intent2-1:cond2-1:experiments2-1	0.17	-0.16 – 0.50	0.310
intent2-1:cond3-2:experiments2-1	-0.09	-0.42 – 0.24	0.595
Random Effects			
σ^2	1.02		
$\tau_{00\ CASE}$	0.52		
$\tau_{00\ story}$	0.15		
ICC	0.40		
$N_{\ story}$	42		
$N_{\ CASE}$	82		
Observations	3444		
Marginal R^2 / Conditional R^2	0.061 / 0.434		

Mind Attribution (Exp. 2): Human-Likeness			
Predictors	*Estimates*	*CI*	*p*
(Intercept)	3.73	3.52 – 3.94	**<0.001**
intent2-1	0.30	0.21 – 0.38	**<0.001**
cond2-1	-0.04	-0.46 – 0.38	0.836
cond3-2	0.29	-0.13 – 0.71	0.167
experiments2-1	0.28	0.02 – 0.53	**0.034**
intent2-1:cond2-1	0.29	0.09 – 0.50	**0.006**
intent2-1:cond3-2	-0.19	-0.40 – 0.01	0.069
intent2-1:experiments2-1	-0.12	-0.29 – 0.05	0.157
cond2-1:experiments2-1	-0.16	-0.37 – 0.05	0.127
cond3-2:experiments2-1	-0.06	-0.26 – 0.15	0.597
intent2-1:cond2-1:experiments2-1	0.24	-0.18 – 0.65	0.264
intent2-1:cond3-2:experiments2-1	0.09	-0.32 – 0.51	0.661
Random Effects			
σ^2	1.61		
$\tau_{00\ CASE}$	0.30		
$\tau_{00\ story}$	0.28		
ICC	0.26		
$N_{\ story}$	42		
$N_{\ CASE}$	82		
Observations	3444		
Marginal R^2 / Conditional R^2	0.029 / 0.285		

Note. Moral-Emotional Content Conditions: 1 = Negative, 2 = Neutral, 3 = Positive. Intentionality Conditions: 1 = Low Intentionality, 2 = High Intentionality. Experiments: 1 = Experiment 1, 2 = Experiment 2. Boldface indicates statistical significance at $\alpha = .05$.

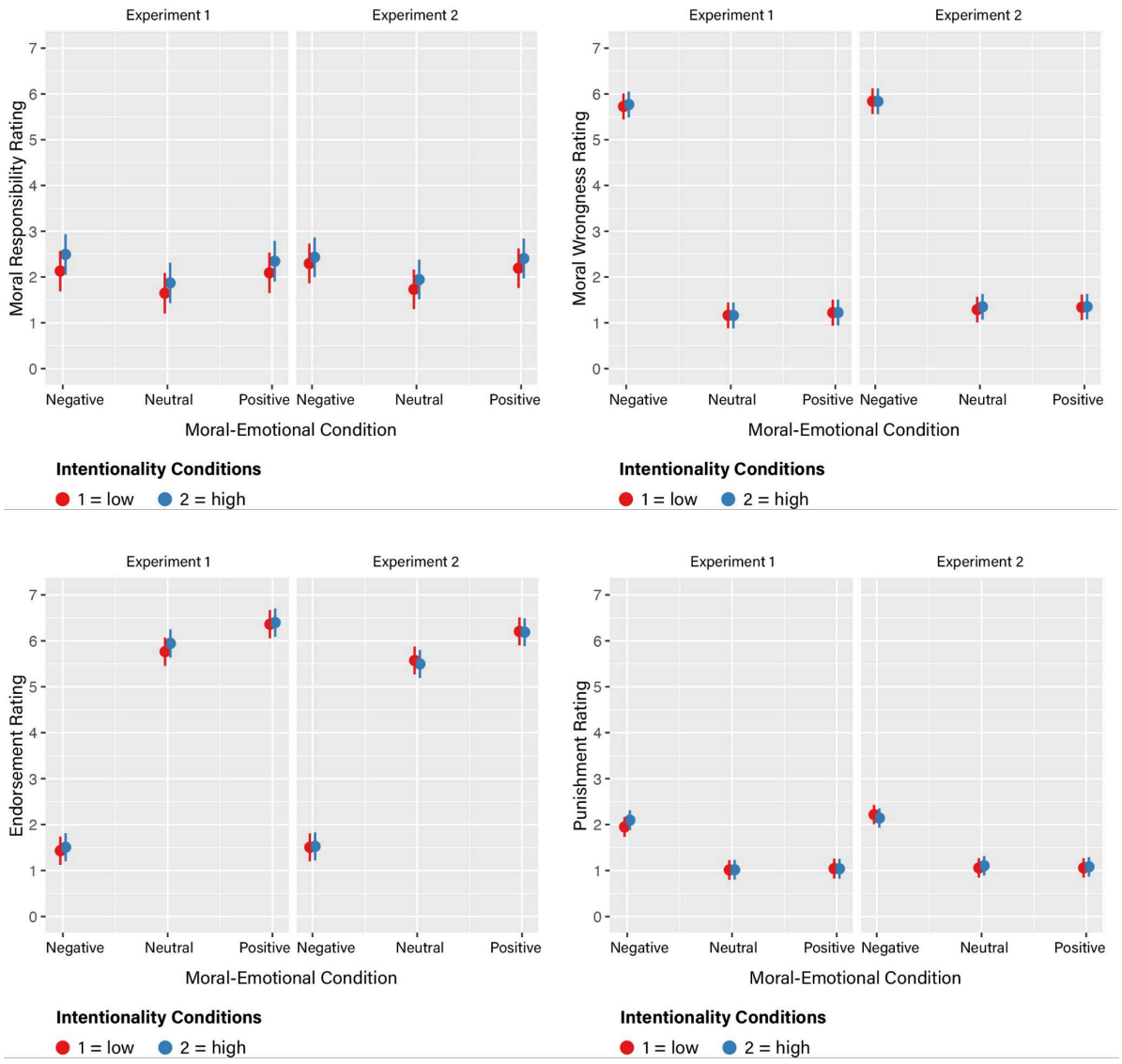
3. Moral Evaluation

When combining data from both experiments, the effects on participants' moral evaluation of the robots and their actions were identical to the results from Experiment 2 (see Table 9, Fig. 10). Intentionality and moral-emotional content affected moral responsibility ratings with increasing ratings significantly correlating with high intentionality and with positive and negative conditions compared to the neutral condition. Moral-emotional content, but not intentionality, again affected ratings of moral wrongness, endorsement and punishment. The

difference between the endorsement ratings for robots in the negative and neutral conditions decreased when participants were presented with an image of a robot.

Figure 10

Comparison of Rating Results for Moral Evaluation in Experiments 1 and 2



Note. Rating results of participants' moral evaluations of the robots' actions (moral responsibility, moral wrongness, endorsement and punishment).

Table 9

Results of linear mixed models analyses of Moral Evaluation ratings in Experiments 1 and 2

Moral Evaluation (Exp. 1 & 2): Moral Responsibility

Predictors	*Estimates*	*CI*	*p*
(Intercept)	2.13	1.83 – 2.43	**<0.001**
intent2-1	0.23	0.16 – 0.31	**<0.001**
cond2-1	-0.54	-0.71 – -0.37	**<0.001**
cond3-2	0.46	0.29 – 0.63	**<0.001**
experiments2-1	0.07	-0.52 – 0.66	0.813
intent2-1:cond2-1	-0.03	-0.20 – 0.15	0.749
intent2-1:cond3-2	0.01	-0.16 – 0.19	0.888
intent2-1:experiments2-1	-0.09	-0.24 – 0.05	0.206
cond2-1:experiments2-1	0.03	-0.15 – 0.20	0.754
cond3-2:experiments2-1	0.00	-0.18 – 0.18	0.998
intent2-1:cond2-1:experiments2-1	0.21	-0.14 – 0.57	0.233
intent2-1:cond3-2:experiments2-1	-0.03	-0.38 – 0.32	0.859
Random Effects			
σ^2	1.15		
$\tau_{00\ CASE}$	1.77		
$\tau_{00\ story}$	0.04		
ICC	0.61		
$N_{\ story}$	42		
$N_{\ CASE}$	82		
Observations	3444		
Marginal R^2 / Conditional R^2	0.024 / 0.619		

Moral Evaluation (Exp. 1 & 2): Moral Wrongness

Predictors	*Estimates*	*CI*	*p*
(Intercept)	2.77	2.61 – 2.94	**<0.001**
intent2-1	0.02	-0.05 – 0.09	0.602
cond2-1	-4.55	-4.87 – -4.23	**<0.001**
cond3-2	0.04	-0.27 – 0.36	0.781
experiments2-1	0.12	-0.09 – 0.34	0.254
intent2-1:cond2-1	0.01	-0.17 – 0.18	0.919
intent2-1:cond3-2	-0.02	-0.20 – 0.15	0.821
intent2-1:experiments2-1	0.01	-0.13 – 0.15	0.896
cond2-1:experiments2-1	0.07	-0.11 – 0.24	0.466
cond3-2:experiments2-1	-0.03	-0.21 – 0.14	0.708
intent2-1:cond2-1:experiments2-1	0.11	-0.24 – 0.46	0.534
intent2-1:cond3-2:experiments2-1	-0.05	-0.41 – 0.30	0.759
Random Effects			
σ^2	1.15		
$\tau_{00\ CASE}$	0.21		
$\tau_{00\ story}$	0.16		
ICC	0.24		
$N_{\ story}$	42		
$N_{\ CASE}$	82		
Observations	3444		
Marginal R^2 / Conditional R^2	0.751 / 0.812		

Moral Evaluation (Exp. 1 & 2): Endorsement

Predictors	*Estimates*	*CI*	*p*
(Intercept)	4.49	4.31 – 4.67	**<0.001**
intent2-1	0.04	-0.03 – 0.11	0.304
cond2-1	4.20	3.84 – 4.56	**<0.001**
cond3-2	0.59	0.23 – 0.96	**0.002**
experiments2-1	-0.15	-0.37 – 0.06	0.163
intent2-1:cond2-1	0.00	-0.17 – 0.17	0.962
intent2-1:cond3-2	-0.04	-0.21 – 0.13	0.623
intent2-1:experiments2-1	-0.12	-0.26 – 0.02	0.089
cond2-1:experiments2-1	-0.37	-0.54 – -0.20	**<0.001**
cond3-2:experiments2-1	0.14	-0.03 – 0.31	0.115
intent2-1:cond2-1:experiments2-1	-0.20	-0.54 – 0.14	0.251
intent2-1:cond3-2:experiments2-1	0.20	-0.14 – 0.54	0.247
Random Effects			
σ^2	1.07		
$\tau_{00\ CASE}$	0.21		
$\tau_{00\ story}$	0.21		
ICC	0.28		
$N_{\ story}$	42		
$N_{\ CASE}$	82		
Observations	3444		
Marginal R^2 / Conditional R^2	0.752 / 0.823		

Moral Evaluation (Exp. 1 & 2): Punishment

Predictors	*Estimates*	*CI*	*p*
(Intercept)	1.40	1.27 – 1.53	**<0.001**
cond2-1	-1.05	-1.19 – -0.92	**<0.001**
cond3-2	0.01	-0.13 – 0.14	0.921
intent2-1	0.02	-0.04 – 0.09	0.475
experiments2-1	0.08	-0.16 – 0.33	0.503
cond2-1:intent2-1	-0.01	-0.17 – 0.15	0.901
cond3-2:intent2-1	-0.02	-0.18 – 0.14	0.850
cond2-1:experiments2-1	-0.09	-0.25 – 0.07	0.263
cond3-2:experiments2-1	-0.04	-0.20 – 0.12	0.651
intent2-1:experiments2-1	-0.05	-0.18 – 0.08	0.454
cond2-1:intent2-1:experiments2-1	0.27	-0.06 – 0.59	0.105
cond3-2:intent2-1:experiments2-1	-0.02	-0.34 – 0.30	0.919
Random Effects			
σ^2	0.96		
$\tau_{00\ CASE}$	0.29		
$\tau_{00\ story}$	0.02		
ICC	0.25		
$N_{\ story}$	42		
$N_{\ CASE}$	82		
Observations	3444		
Marginal R^2 / Conditional R^2	0.164 / 0.370		

Note. Moral-Emotional Content Conditions: 1 = Negative, 2 = Neutral, 3 = Positive. Intentionality Conditions: 1 = Low Intentionality, 2 = High Intentionality. Experiments: 1 = Experiment 1, 2 = Experiment 2. Boldface indicates statistical significance at $\alpha = .05$.

Discussion Experiment 2

Just as in Experiment 1, both intentionality and moral-emotional content affected participants' evaluations of the robots' actions, although for the most part the variables did not interact. Overall, moral-emotional content dominated participants' evaluations and intentionality mostly affected mind attribution as well as the attributed moral responsibility. The inclusion of an image of a robot presented simultaneously with the information about the robots' actions affected some of the judgments.

The emotional response results of Experiment 2 closely matched the results of the first experiment. We found the expected influence of moral-emotional content on the valence and arousal of participants' emotional response, while intentionality did not affect the ratings in the way we had predicted. However, in Experiment 2 there was an interaction trend between intentionality and moral-emotional content on valence ratings; in particular, the difference of valence ratings between intentionality conditions tended to increase descriptively between negative and neutral moral-emotional conditions. This would point to a dominance of moral-emotional content that was previously also found in similar studies (e.g. Baum et al. 2020). It may be that the emotional response of morally relevant situations is dictated by the emotional quality of information, thereby discounting other potentially relevant information. This could explain that when the emotional content is neutral, there is a greater influence of intentionality.

The conclusion that moral-emotional content dominates is further supported by valence ratings not changing between studies. The addition of another piece of information—an image of a robot—did not alter the valence of participants' emotional response to the robots' actions. However, including an image of a robot increased participant's arousal response in Experiment 2

and decreased the difference between neutral and positive conditions. It could be that the influence of an image of a robot alongside the information was able to moderate the impact of the informations' moral-emotional quality.

As in Experiment 1, both intentionality and moral-emotional content affected mind attribution. However, experience ratings were unaffected by the level of robots' intentionality and both agency and experience were significantly impacted by moral-emotional content when comparing neutral and positive conditions. While we had imagined that adding an image would increase the influence of intentionality, the contrary occurred. Rather than being additive, it may be that intentionality is further discounted when there is another, more salient factor.

In Experiment 2 as well as when all data sets were combined, intentionality only affected moral responsibility ratings, but not the other moral evaluation items. Moral evaluations combine aspects of mind attribution and emotional response. The results can therefore be seen as further support for the dominance of moral-emotional content. Moral responsibility is the most theoretically connected to mind attribution and indeed intentionality differentially affected how responsible participants thought the robots were. The moral wrongness of an action should on theoretical grounds not be affected by intentionality, but on occasion there have been empirical findings that intentionality affects moral wrongness. It is possible that robots are currently considered so low in mindedness that the descriptions of them in intentional terms do not suffice to convince people otherwise. Similarly, participants did not find robots to be appropriate targets of punishment for moral transgressions. However, there appeared to be two distinct types of responses to this topic: either participants always chose the lowest possible rating for all robots regardless of the moral-emotional content or they gave a high punishment response only in the negative condition. It certainly makes sense only to punish the robots whose actions were

negative, but this finding suggests that there are substantial inter-individual differences that may be a target of investigation in their own right.

Whether or not participants agree with a continued use of robots is perhaps the better measure of desired consequences for a robot's actions. Again, there were possible floor and ceiling effects. The robots in the negative and positive conditions had extremely low and high endorsement ratings respectively. Robots in the neutral condition had slightly lower endorsement in the second experiment suggesting that when stories are not as affected by moral-emotional content, subtle effects can be found.

Overall, moral-emotional valence had more influence on the results than intentionality. It is possible that intentionality is a more subtle characteristic and that in the face of strong moral-emotional content, it is neglected in people’s evaluations. In support of this hypothesis we found more influences of intentionality on evaluations in the neutral condition, where moral-emotional content would have less impact.

General Discussion

Social robots are expected to become more widely used covering a wide range of applications in the near future (Lutz et al. 2019). As their use becomes more widespread, the risk of harm caused by these machines increases and so questions pertaining to responsibility in such situations need to be considered (Ladak et al. 2023; Borg et al. 2024). One aspect of this interdisciplinary subject is whether or not people take robots to be appropriate moral agents and hold robots responsible when potential moral transgressions occur. In the past, much research has focused on robot mind attribution, investigating diverse reasons why people may anthropomorphize robots and treat robots as intentional, i.e. minded, beings (e.g. Gray et al. 2012). Other research has focused on blame and responsibility, often in very specific moral

dilemma situations (e.g. Malle et al. 2015). Some work on moral judgments have considered people's emotional response to robots' bad behavior (e.g. Bigman et al. 2023). The present studies combined these aspects of moral judgments of robots, attempting to holistically consider people's evaluations when robots perform morally good or bad things. In two experiments, participants read 42 stories about robots' actions, which differed in level of intentionality and moral-emotional quality of information, and subsequently rated the robots and their actions, considering in turn their own emotional response, the robots' mindedness and questions about the morality of the described behavior. The experiments differed only in the addition of images of robots alongside the stories in the second experiment.

Experiment 1 established that both intentionality and moral-emotional content affected participants' evaluations of robots' actions. However, moral emotional content appeared to dominate results and there were no interactions between the two factors. Participants' emotional response corresponded to previous research and our expectations in that the valence of participants' emotional response was evaluated in line with the moral emotional conditions. Arousal ratings only differed between the negative and neutral conditions; while not in line with predictions, these results resembled prior studies. While, contrary to our predictions, intentionality did not differentially impact emotional response ratings, it did affect mind attribution. Participants indicated that they attributed more mind to the robots in the high intentionality than to the robots in the low intentionality condition, rating the former higher in both agency, experience and human-likeness. Somewhat surprisingly, positive content affected mind attribution ratings more than negative content. Morally negative situations have been shown to increase mind attribution, as a way to alleviate a search for explanations (Taylor 1991). When actions have negative outcomes, they are therefore more likely seen as intentional and the

agents are more attributed more mind (Feltz, 2007; Knobe, 2003). Our results perhaps suggest that the level of mind ascribed to robots is too low to force such an effect. People are especially conservative with explicit attributions of mind to robots and a negative outcome may not have been sufficient to differentially affect agency ratings.

Coherent with the mind attribution ratings, and in line with predictions, participants also evaluated the robots paired with high intentionality information as more morally responsible for their actions than those paired with low intentionality information. However, we did not find expected interaction between intentionality and moral-emotional content on moral responsibility ratings. Other moral evaluations were again dominated by moral-emotional content and did not show effects of intentionality. In summary, the intentionality ascribed in information about robots does affect people's evaluations of them as potential moral agents and increases moral responsibility, the positive or negative quality of the actions however more strongly affects evaluations in general and the impact of moral-emotional content is not moderated by intentionality.

In Experiment 2, we included an image of a robot with each story as an attempt to increase the effect of mind attribution, which has been shown to correlate with human-like appearance of robots. While mind attribution did indeed increase with more human-like images, the amendment to the experimental design did not otherwise affect participants' ratings and the results remained stable across experiments. Participants' emotional response ratings corresponded to the moral-emotional quality of information, while mind attribution ratings were affected by intentionality and moral-emotional content. Moral responsibility, but not the other moral evaluation ratings, was also affected by intentionality alongside moral-emotional content.

The dominance of moral-emotional content over intentionality remained and for the most part, these two factors did not interact.

The few interactions that we did find, further support our conclusion that moral-emotional content dominates people's evaluations of robots. In the second experiment, unlike in the first, intentionality and moral-emotional content interacted on valence ratings. Specifically, the difference between ratings of robots in low and high intentionality conditions increased between negative and neutral information conditions. We interpret this result to show that intentionality affects evaluations more in the absence of strong moral-emotional content, e.g. when the information is more neutral in valence. On the other hand, in the presence of strong moral-emotional content, intentionality and other potentially relevant information is discounted.

Intentionality and moral-emotional content also interacted on human-likeness ratings. The perceived human-likeness difference between high and low intentionality robots was greater when paired with neutral information than when paired with negative information. Between moral-emotional conditions, the difference between intentionality conditions increased. When data sets were combined a trend for a similar effect was found also between positive and neutral conditions. This finding offers further support for the dominance of moral-emotional content, because the influence of intentionality can unfold in the neutral condition as opposed to the more emotionally salient negative condition. Furthermore, this effect occurs in a measure that is less relevant in moral and emotional terms. Moral-emotional content dominates most when the evaluations are also morally or emotionally relevant. The mind attribution measures agency and experience showed similar tendencies with ratings in the negative condition being more similar between intentionality conditions. This can be seen in the decrease of the difference between intentionality conditions between experiments that within factor analysis revealed to occur

predominantly in the negative condition. In other words, intentionality had a greater effect in the condition in which moral-emotional content is not negative and therefore less relevant.

Conclusion

Two experiments investigated whether people attribute mind to robots in morally relevant situations, as well as people's emotional response and moral evaluations of the situations. The information that people read before rating the robots were based on current and predicted future abilities and uses of robots or AI. We enhanced the information in the high intentionality condition with terms that anthropomorphized the robots and described them as minded agents. This type of language reflects a tendency in humans to anthropomorphize machines, especially if they already have other anthropomorphic features or if they behave in ways that make them seem intentional. It also reflects a tendency to anthropomorphize current AI technology (Shanahan 2023). While using an intentional stance in our discourse about AI may be a natural way of simplifying the machines' behavior, the language we use can affect the abilities we project onto them and the types of social roles we ascribe to them. The present research demonstrates how anthropomorphic language increases the attributed agency and moral responsibility of social robots. It is possible that the responsibility which is assigned to the robots undermines the incentives to improve faulty products or to regulate technology by focusing on individual machines (Borg et al. 2024).

Our research furthermore emphasized the effect of moral-emotional content on people's moral judgments. While intentionality primarily affected mind attribution and moral responsibility, emotional response as well as the moral wrongness of the actions and the desire for repercussions were almost entirely affected by the moral-emotional quality of the information. On the one hand this indicates that people make moral judgements about robots

somewhat irrespective of information about their mental abilities. The responsibility gap issue may therefore exist whether or not robots are described as intentional beings. On the other hand, it may demonstrate the skepticism about robot's mental life that has been found in people's explicit opinions. The effects of intentionality on ratings were subtle, but research using more implicit measures could perhaps show that the influence of this type of information nevertheless affects our behavior towards them. As the abilities and uses of social robots increases, we may also see people's attitudes towards and opinions about them change and become increasingly open to intentional descriptions of their behavior.

Materials and Methods

The two experiments were almost identical in materials and methods. The only difference was that in the second experiment, portrait images of robots accompanied the information about the robots. The preregistration for Experiment 1 can be accessed at https://osf.io/8nt5c and the preregistration for Experiment 2 can be accessed at https://osf.io/6thk2.

Participants

We recruited German-speaking participants from Prolific (prolific.co) and the participant database of the Department of Psychology at Humboldt-Universität zu Berlin (pesa.psychologie.hu-berlin.de). As compensation, participants received monetary payment or student credits. Experiment 1 had 40 participants and Experiment 2 had 42 participants. We determined sample sizes based on similar studies (e.g., Abdel Rahman 2011; Baum et al. 2020; Maier et al. 2024), and counterbalancing measures across the three moral-emotional and two intentionality conditions. Additionally, in Experiment 2, images of robots were counterbalanced so that each robot image was matched with a story in each moral-emotional condition; consequently, every story was matched with three different robot images. The study adhered to

the principles of the Declaration of Helsinki and received approval from the Ethics Committee of the Department of Psychology at Humboldt-Universität zu Berlin. At the start of the experiments, participants provided informed written consent.

Materials

We wrote 42 two-sentence stories, fourteen each with emotionally positive (e.g. a robot that reads to the elderly), neutral (e.g. a robot that works in a warehouse) or negative (e.g. a robot that interrogates political dissidents) information about robots (see Figure X for examples; for the complete set, see Appendix Y). The stories were based on current news articles about developments in robotics so that the robots' fictional actions resembled real functions carried out by existing robots (e.g. commercial, research, military or medical). Emotional valence corresponded to moral categories: neutral stories merely described functionality, while the positive and negative stories described actions that are commonly held to be kind and helpful or contemptible and cruel respectively. The robots were described to be causally responsible for the actions.

The stories' intentionality was modified by using words describing abilities associated with the two dimensions of mind, agency (e.g. self-control, memory, planning, communication, thought) and experience (e.g. emotion, desire, consciousness, perception), to suggest high levels of mindedness. These descriptions imply an intentional stance, making sense of behavior through mental causation. Stories for the low intentionality condition instead used words associated with mechanics or programming (e.g. algorithm, system, analytics). These descriptions imply a design or physical stance by drawing attention to underlying physical mechanisms (Dennett). These stances imply viewing robots as machines that are designed or programmed to behave in specific ways and as non-biological entities. Each of the 42 stories had a high and low intentionality

version and participants only read either the high or the low version of a particular story due to counterbalancing. Overall, participants were presented with an equal number of high and low intentionality stories.

The picture stimuli (used only in Experiment 2) were 42 full-color frontal portrait photographs featuring existing humanoid robots, each displaying approximately neutral facial expressions (refer to Figure 1 for an example; names and sources of the robots used are listed in the Supplementary Materials). The robots used have been developed for commercial (e.g. entertainment or personal service) or research (e.g. psychology or robotics) purposes. The images were all found on the online database abotdatabase.info (Phillips et al., 2018). Brand names and affective symbols (e.g. hearts), were removed from some images, so that these would not affect the ratings. We selected images of robots that were human-like in structure. All robots had distinct heads and faces with eyes, although not all robots had mouths. We avoided using images of android robots—robots that look almost exactly like humans—because they may be mistaken for actual humans in still photographs. The robots' heads were cropped from the original pictures and placed on a gray background and matched in size and eye placement across all images. All images of robots had frontal gaze or were corrected to frontal gaze in one instance. The images were counterbalanced, so that any specific robot was matched equal amounts of times with a positive, neutral or negative story as well as equal amounts of times with a high or low intentionality story.

Procedure

Participants were presented with all 42 robot stories one at a time in a random order and asked to rate the robots immediately after reading each story. In Experiment 2, images were placed just above the stories. The ratings measured participants' emotional response, mind

attribution and moral evaluation. Emotional response consisted of valence and arousal ratings, i.e. participants rated the valence and arousal of their emotional response to the robots' actions on 7-point Likert scales using self-assessment manikins (SAM), which are designed to measure emotional response to stimuli, ranging from *very negative* to *very positive* (center anchored with *neutral*) and *calm* to *tense* respectively (Bradley and Lang 1994). Mind attribution was measured with ratings of the robots' agency and experience. These are considered two separate dimensions of mind attribution (Gray et al. 2007). Participants rated their agreement with the following statements: "*This robot has intentions and the ability to decide things independently and to plan and carry out its actions"* for agency and "*This robot has the ability to experience sensations such as pain, fear, joy or anger"* for experience. Participants also rated how human-like they imagined the robots to be by choosing one of six line drawings of increasingly anthropomorphic robots (Komatsu et al. 2021). While no measure of mind attribution itself, human-likeness correlates with mind attribution and we decided to use this measure for exploratory purposes. Finally, participants rated four items relevant to moral evaluations. They rated their agreement with the following statements: "*The robot is morally responsible for the consequences of its actions*", "*The robot's behavior is morally wrong*", "*The robot should remain in use*" and finally "*The robot should be punished*". Participants' responses were measured on 7-point Likert scales ranging from *completely disagree* to *completely agree*. At random intervals between ratings, participants responded to multiple choice questions about the robots' functions with four possible answers to verify that they were paying attention.

After the main section of the experiment, participants were asked to complete several questionnaires. Participants responded to the Neo-FFI (citation), the Negative Attitudes towards Robots Scale (Namura et al. 2006; Syrdal, Dautenhahn et al.), the Attitude towards Artificial

Intelligence Scale (Sindermann et al., 2021) and the Individual Differences in Anthropomorphism Questionnaire (Waytz, Cacioppo & Epley 2010). Next, the participants responded to the Short Loneliness Scale (Hughes et al. 2004) and were asked if they currently or formerly owned a pet and if they were vegetarian for conscientious reasons. Loneliness and attitudes towards animals correlate with anthropomorphism.

Finally, participants were asked if they had researched information about the robots during the experiment or if they knew any of the robots prior to the experiment. They were also asked if they believed they had higher than average knowledge about AI and robotics and if they frequently watched films or read books from the Sci-Fi genre. They then responded to a Hypothesis Awareness Questionnaire which includes a question about whether they had been distracted during the experiment. They were also asked if they had any doubts about the veracity of the stories and if so, how many of the stories in percent they had doubted. The participants were then debriefed and informed that none of the information that all stories were fictional and in Experiment 2 that none of the information in any way pertained to the pictured robots.

Data Exclusion Criteria

We preregistered criteria for the exclusion of a participant's data. Data would be excluded for poor general task performance. Throughout the experiments, participants intermittently responded to multiple choice questions about the robots in the last story they had read. Less than 50% correct responses would have led to data exclusion, but no participants were excluded for this reason. Failed manipulation also constituted poor task performance, because it implied that participants had not properly processed the information. Data were therefore excluded if valence ratings deviated clearly from the intended moral-emotional manipulation (e.g. if a participant systematically rated very positive valence for robots that were paired with negative stories),

specifically if they represented a statistical outlier in the unintended direction (Median Absolute Deviation method with coefficient 2.5). Based on these criteria, we excluded four data sets in Experiment 1 and seven data sets in Experiment 2.

Data were also excluded if the participants indicated that they had had strong doubts about the veracity of the stories. At the very end of the experiments, participants were asked if they doubted the veracity of the stories and if so, to indicate in percent how many of the stories they had doubted. We excluded participants whose response represented a statistical outlier (Median Absolute Deviation method with coefficient 2.5). Based on these criteria, we excluded two data sets in Experiment 1 and eight data sets in Experiment 2. Finally, we excluded data sets if participants claimed that they had researched (e.g. googled) information about the robots during the experiment or that they had been distracted. Based on these criteria, we excluded one data set in Experiment 1 and three data sets in Experiment 2, because participants reported that they had researched the robots. No participants reported to have been highly distracted and so no data sets were excluded for this reason. After excluding these data sets, data collection continued until 40 and 42 complete data sets were obtained respectively in Experiment 1 and Experiment 2.

Statistical Analysis

We used linear mixed effects models (LMMs) to analyze the rating data (Baayen et al., 2008). Moral-emotional content (negative, neutral, positive) and intentionality (high, low) of information were modeled as fixed effects, coded as sliding difference contrasts. Random intercepts were modeled for both participants and items (stories). We modeled the maximal random structure compatible with model convergence (Matuschek et al., 2017).

References

Abdel Rahman, R. (2011). Facing Good and Evil: Early Brain Signatures of Affective Biographical Knowledge in Face Recognition. Emotion, 11(6), 1397–1405. https://doi.org/10.1037/a0024717

Abubshait, A., & Wiese, E. (2017). You look human, but act like a machine: agent appearance and behavior modulate different aspects of human–robot interaction. *Frontiers in psychology*, *8*, 1393.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. Journal of Memory and Language, 59(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Baum, J., & Abdel Rahman, R. (2021). Emotional News Affects Social Judgments Independent of Perceived Media Credibility. Social Cognitive and Affective Neuroscience, 16(3), 280–291. https://doi.org/10.1093/scan/nsaa164

Baum, J., Rabovsky, M., Rose, S. B., & Abdel Rahman, R. (2020). Clear judgments based on unclear evidence: Person evaluation is strongly influenced by untrustworthy gossip. Emotion, 20(2), 248–260. https://doi.org/10.1037/emo0000545

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2023). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*, *152*(1), 4.

Borg, J. S., Sinnott-Armstrong, W., & Conitzer, V. (2024). Moral AI: And How We Get There. Random House.

Ceh, S., & Vanman, E. (2018). The robots are coming! The robots are coming! Fear and empathy for human-like entities.

Chopra, S. (2011). Taking the moral stance: Morality, robots, and the intentional stance. Technologies on the stand: Legal and ethical questions in neuroscience and robotics, 285.

Cushman, F. (2015). Deconstructing intent to reconstruct morality. *Current Opinion in Psychology*, *6*, 97-103.

de Graaf, M. M. (2016). An ethical evaluation of human–robot relationships. *International journal of social robotics*, *8*, 589-598.

Feltz, A. (2007). The Knobe effect: A brief overview. The Journal of Mind and Behavior, 265-277.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge university press.

Fong, T., Nourbakhsh, I., & Dautenhahn, K. (2003). A survey of socially interactive robots. *Robotics and autonomous systems*, *42*(3-4), 143-166.

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

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. science, 315(5812), 619-619.

Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological inquiry*, *23*(2), 101-124.

Kim, T. W., & Duhachek, A. (2020). Artificial intelligence and persuasion: A construal-level account. *Psychological science*, *31*(4), 363-380.

Knobe, J. (2003). Intentional action in folk psychology: An experimental investigation. *Philosophical psychology*, *16*(2), 309-324.

Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PloS one*, *3*(7), e2597.

Ladak, A., Loughnan, S., & Wilks, M. (2024). The moral psychology of artificial intelligence. *Current Directions in Psychological Science*, *33*(1), 27-34.

Lemley, M. A., & Casey, B. (2019). Remedies for robots. *The University of Chicago Law Review*, *86*(5), 1311-1396.

Loh, W., & Loh, J. (2017). The Example of Autonomous Cars. Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence, 35.

Lutz, C., Schöttler, M., & Hoffmann, C. P. (2019). The privacy implications of social robots: Scoping review and expert interviews. *Mobile Media & Communication*, *7*(3), 412-434.

Maier, M., Leonhardt, A., Blume, F. , Bideau, P., Hellwich, O & Abdel Rahman, R. Brain dynamics of mental state attribution during perception of social robot faces. Manuscript submitted for publication.

Malle, B. F., Scheutz, M., Arnold, T., Voiklis, J., & Cusimano, C. (2015, March). Sacrifice one for the good of many? People apply different moral norms to human and robot agents. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction* (pp. 117-124).

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. Journal of Memory and Language, 94, 305–315. https://doi.org/10.1016/j.jml.2017.01.001

Misselhorn, C. (2018). Grundfragen der Maschinenethik: Reclam Universal-Bibliothek. Reclam.

Ouchchy, L., Coin, A., & Dubljević, V. (2020). AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media. *AI & SOCIETY*, *35*, 927-936.

Phillips, E., Zhao, X., Ullman, D., & Malle, B. F. (2018). What is human-like?: Decomposing robot human-like appearance using the Anthropomorphic roBOT (ABOT) Database. HRI '18. https://www.abotdatabase.info/

Riek, L. D., Rabinowitch, T. C., Chakrabarti, B., & Robinson, P. (2009, March). How anthropomorphism affects empathy toward robots. In *Proceedings of the 4th ACM/IEEE international conference on Human robot interaction* (pp. 245-246).

Shanahan, M. (2024). Talking about large language models. Communications of the ACM, 67(2), 68-79.

Smart, J. J. C. (1961). Free-will, praise and blame. *Mind*, *70*(279), 291-306.

Sparrow, R. (2007). Killer robots. Journal of Applied Philosophy, 24(1), 62-77. https://doi.org/10.1111/j.1468-5930.2007.00346.x

Strawson, P. F. (2003). Freedom and resentment. *Free will*, *2*, 72-93.

Talbert, Matthew. Moral Responsibility. The Stanford Encyclopedia of Philosophy. Edward N. Zalta & Uri Nodelman (eds.). https://plato.stanford.edu/archives/sum2024/entries/moral-responsibility/.

Taylor, S. E. (1991). Asymmetrical effects of positive and negative events: the mobilization-minimization hypothesis. *Psychological bulletin*, *110*(1), 67.

Tognazzini, Neal and D. Justin Coates. Blame. The Stanford Encyclopedia of Philosophy. Edward N. Zalta (ed.). https://plato.stanford.edu/archives/sum2021/entries/blame.

Wallace, R. J. (1994). *Responsibility and the moral sentiments*. Harvard University Press.

Williston, B. (2006). Blaming agents in moral dilemmas. *Ethical Theory and Moral Practice*, *9*, 563-576.