Title: Brain dynamics of mental state attribution during perception of social robot faces

Authors:

Martin Maier [1, 2] *, Alexander Leonhardt [1,] , Florian Blume [2, 3], Pia Bideau [2, 4], Olaf Hellwich [2, 3], and Rasha Abdel Rahman [1, 2]

Affiliations:

[1] Humboldt-Universität zu Berlin; Berlin, Germany
[2] Science of Intelligence, Research Cluster of Excellence; Berlin, Germany
[3] Technische Universität Berlin; Berlin, Germany
[4] Univ. Grenoble Alpes, Inria, CNRS, Grenoble, France
*Corresponding author. Email: martin.maier@hu-berlin.de

Abstract: The interplay of mind attribution and emotional responses is considered crucial in shaping human trust and acceptance of social robots. Understanding this interplay can help us create the right conditions for successful human-robot social interaction in the service of societal needs. In this study we show that information about robots describing positive, negative or neutral behavior prompts participants (*N*=90) to attribute mental states to robot faces, modulating impressions of trustworthiness, facial expression and intentionality. These novel findings were replicated in an experiment investigating the underlying dynamics in the human mind and brain. EEG recordings from 30 participants revealed that affective information influenced specific processing stages in the brain associated with basic face perception and more elaborate stimulus evaluation. However, a modulation of fast emotional brain responses, typically found for human faces, was not observed. These findings suggest that neural processing of robot faces alternates between being perceived as mindless machines and intentional agents: people rapidly attribute mental states during perception, literally seeing good or bad intentions in robot faces, but are emotionally less affected than when facing humans. These nuanced insights into the fundamental psychological and neurocognitive processes supporting mind attribution hold potential for informing the design of artificial social agents, improving human-robot social interactions, and guiding policies regarding moral responsibility.

One-Sentence Summary: Brain signatures reveal rapid mental state attribution as prior information lets people see emotions in neutral robot faces.

INTRODUCTION

The demand for social robots, embodied artificial systems that interact with humans in their daily lives, is expected to increase in the coming years. As robots are developed for different uses such as care, retail or entertainment, more people are likely to engage with social robots in their private and professional lives (*1–3*). Yet, key psychological and neurocognitive aspects of interacting with social robots are still under investigation, including the extent to which humans—and their brains—process robots akin to intentional social agents with mental states (*4–6*). This question bears important implications for how we deal with artificial social agents as a society, including moral judgments of their responsibility for negative outcomes (*7*).

Previous theoretical work has proposed that two conflicting intuitions come into play, the "intentional" stance and the "physical" or "design" stance (*6, 8–10*). Taking an intentional stance towards robots means to intuitively treat them as if they had a mind: people can tend to anthropomorphize robots, interacting with them as if they possessed mental states such as motivations, intentions, or emotions. This may allow them to tap into processes used in social interaction with other humans, such as theory of mind, as a basis for social perception, communication, and coordination of behavior (*11–15*). On the other hand, people's explicit opinions often reflect a physical stance towards robots, viewing them as machines designed or programmed to behave in specific ways (*8, 11, 16–18*).

How are these seemingly contradictory intuitions about what type of things robots are—intentional beings or mindless machines—reflected in people's perception of robots and the underlying neural processes? Most research has focused on how human-robot interaction and mind perception are shaped by the design characteristics of the robots (*15, 19–24*) and trait characteristics of the human perceivers, such as their attitudes towards robots and artificial intelligence (AI) (*17, 25–27*).

However, some crucial variables that shape human social perception and interaction in real time have been largely overlooked in studying mind attribution to robots. The human ability to perceive and evaluate others' intentions is strongly influenced by context and prior knowledge, such as learned person-related information (*28–32*). Our perception of others is not only shaped by what we can read from their faces, but also from what we read into them based on our expectations (*33*). For instance, social-affective information has been shown to influence brain signatures of early perceptual, reflexive emotional, as well as higher-level evaluative processing (*28–30, 32, 34, 35*), as elaborated on below.

The present study

In this study, we investigate how brain dynamics reflect mental state attribution in the perception of social robots and to what extent these dynamics align with the intentional and physical stance, respectively. Specifically, we test how prior information about robots' behavior influences neural correlates of perception, emotional responses, and evaluation (as illustrated in Fig. 1). In human social perception, prior knowledge or beliefs about others (e.g., "she bullied her work colleague") can lead people to perceive emotional expressions in objectively neutral faces, revealing the attribution of mental states that are not necessarily present in the person (*28, 30, 35*). Cognitive scientists ascribe this to the interplay between bottom-up and top-down processing, where perceptions are constructed based on

combinations of sensory input from the environment and predictions generated based on prior knowledge and expectations (*30*, *36–40*). Here we investigate this mechanism for the first time with robot faces: does information about a robot's behavior literally make people see good or bad intentions in its face? And do people react emotionally to neutral robot faces based on whether they are associated with morally good or bad behavior, as they would with other humans (*28*, *29*, *35*, *41*)?

In two pre-registered experiments, participants were presented with positive, neutral, and negative information about real human-like robots (e.g., teaches social skills to people with autism, assembles orders at a warehouse, reports children to the secret police; see Fig. 1). Subsequently, they rated the valence of the robots' facial expressions and their trustworthiness. We hypothesized that, similar to human faces, participants would attribute emotional expressions to the robots' faces and assess their trustworthiness based on the acquired information, which induced expectations about their possible intentions. Therefore, we anticipated that ratings of both facial expressions and trustworthiness would align with the information's valence. Experiment 1 investigated the effect of information type on ratings of facial expression and trustworthiness, seeking initial evidence that people indeed read good or bad intentions into robot faces. To explore the neurocognitive dynamics associated with the attribution of mental states, Experiment 2 measured evoked brain responses using electroencephalography (EEG) as participants evaluated robots' facial expressions.

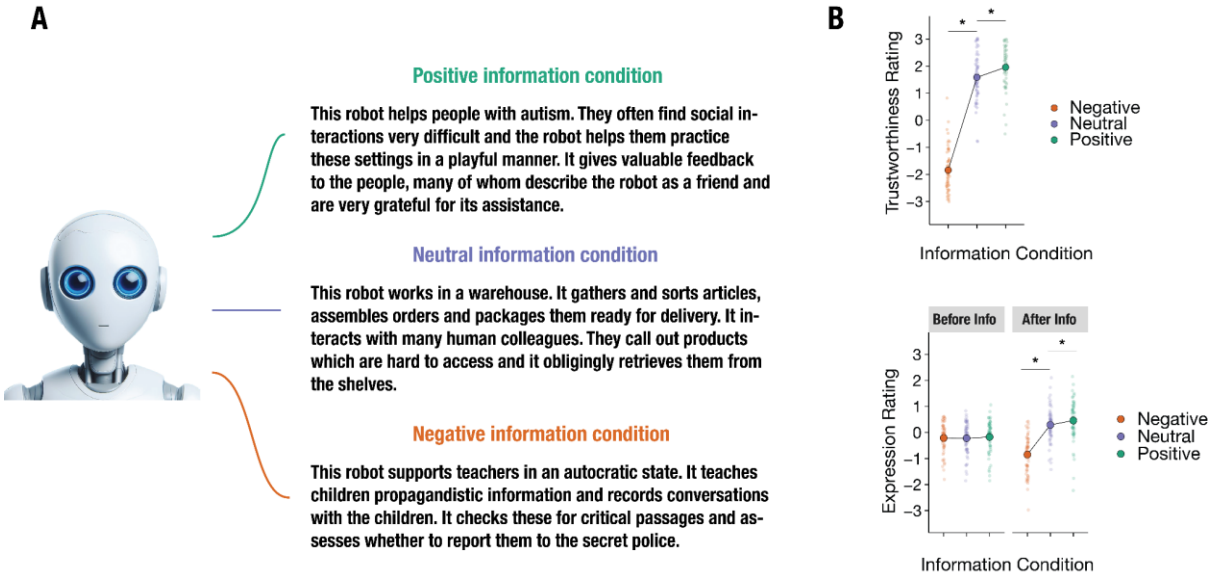

Fig. 1. Information examples and Rating results of Experiment 1. (**A**). Representative story examples for each affective information condition along with an AI-generated robot portrait (*42*) similar to the 36 real robots used in the study. The pairings of robots and information type were counterbalanced, ensuring each robot was paired equally as often with negative, neutral, and positive information across participants. (**B**) Trustworthiness ratings after information acquisition and facial expression ratings before and after information acquisition, categorized by information condition. Large dots denote group means with corresponding 95% CIs, while small dots indicate individual participant means. Asterisks highlight statistically significant differences.

RESULTS

Experiment 1

In Experiment 1, an online study was conducted to investigate the influence of robot-related information on ratings of robots' trustworthiness and facial expressions among a sample of 60 participants, evenly split between English and German native speakers. After the information manipulation, we anticipated both trustworthiness and facial expression ratings to align with the valence of the learned information. Rating data were analyzed using linear mixed effects models (LMMs) with fixed effects coded as sliding difference contrasts.

Facial Expression Ratings

Participants rated facial expressions of robot faces in two phases, before and after acquiring information. We analyzed facial expression ratings with the independent variables phase, i.e. pre- vs. post-learning, and information, i.e. robots matched with positive, neutral, or negative information. By modeling information type nested within phase, we estimated the effect of information on the pre- and post-ratings separately.

As expected, before participants had learned information about the robots, ratings of the facial expression of robots were neutral overall and showed no significant differences between conditions (see Table 1, Fig. 1). However, after hearing the stories, participants rated the same facial expressions differently depending on the valence of the associated information. In line with our predictions, facial expressions in the negative condition were rated as more negative than in the neutral condition and facial expressions in the positive condition were rated as more positive than in the neutral condition. An additional analysis of variance confirmed the significance of the interaction between phase and information, $F(4, 59.53) = 24.16$; $p < .001$, showing that the overall differences between the information conditions increased significantly between phases. An additional LMM analysis including the variable language indicated no differences in the facial expression ratings provided by English and German native speakers (details in Supplementary Materials).

Trustworthiness Ratings

We further analyzed trustworthiness ratings, which were collected after information acquisition. We found a main effect of information. As expected, participants rated robots as significantly more trustworthy when matched with positive information compared to neutral information. Robots matched with negative information were also rated as less trustworthy in comparison to robots matched with neutral information (see Table 2, Fig. 1). An additional analysis including the independent variable group (English vs. German speakers) revealed a main effect of group on trustworthiness ratings ($b = 0.30$, 95% CI = [0.01, 0.59], $p = .041$). Overall, English-speaking participants rated the robots' trustworthiness slightly higher than German-speaking participants (details in Supplementary Materials).

Table 1. Facial expression rating results. Results of linear mixed model analyses of facial expression ratings in Experiment 1

Predictors	*b*	95% CI	*p*-value
Intercept	-0.12	[-0.35, 0.12]	.326
Phase(Post-Pre)	0.17	[0.10, 0.24]	**<.001**
Phase(Pre):Information(Neu-Neg)	-0.01	[-0.13, 0.11]	.873
Phase(Post):Information(Neu-Neg)	1.14	[1.02, 1.26]	**<.001**
Phase(Pre):Information(Pos-Neu)	0.05	[-0.07, 0.17]	.437
Phase(Post):Information(Pos-Neu)	0.17	[0.05, 0.29]	**.005**
Random Effects			*SD*
Participants			0.19
Stimuli			0.37
Residual			1.21
Deviance	13764.60		
log-Likelihood	-6882.30		

Note. Information Conditions: Neg = Negative, Neu = Neutral, Pos = Positive; Phase Conditions: Pre = pre-rating, i.e. before information acquisition, Post = post-rating, i.e. after information acquisition; Colons indicate nesting of fixed variables; Boldface indicates statistical significance at α = .05.

Table 2. Trustworthiness rating results. Results of linear mixed model analyses of trustworthiness ratings in Experiment 1

Predictors	*b*	95% CI	*p*-value
Intercept	0.57	[0.40, 0.74]	**<.001**
Information(Neu-Neg)	3.42	[3.15, 3.70]	**<.001**
Information(Pos-Neu)	0.37	[0.22, 0.53]	**<.001**
Random Effects			*SD*
Participants			0.55
Information(Neu-Neg)			0.97
Information(Pos-Neu)			0.43
Stimuli			0.25
Residual			1.03
Deviance	6606.81		
log-Likelihood	-3303.41		

Note. Information Conditions: Neg = Negative, Neu = Neutral, Pos = Positive. Boldface indicates statistical significance at α = .05.

Discussion of Experiment 1

In Experiment 1, our information manipulation influenced participants' trustworthiness ratings, demonstrating that people clearly distinguish between the trustworthiness of robots previously described as fulfilling negative, neutral, and positive tasks. The additional main effect of participant group on trustworthiness

ratings may reflect genuine differences in trust evaluation towards humanoid robots among English and German speakers. Alternatively, despite our efforts to maintain consistency in meaning across languages, it could stem from nuances in the robots' backstories conveyed differently in each language. More importantly, we also observed an effect of information on facial expression ratings, providing initial evidence that learned information may lead humans to perceive emotional facial expressions in robot faces. Notably, in the absence of prior information, these faces were initially rated as neutral. These findings imply that participants attributed mental states already during robot face perception.

However, explicit facial expression ratings may have been influenced by task demands, as participants could have adjusted their ratings in accordance with perceived experimenters' hypotheses. Addressing this concern, Experiment 2 used EEG to provide insights into the underlying neurocognitive mechanisms. The use of event-related potentials (ERPs) allowed us to assess the genuine effect of information on perception more implicitly, shedding light on whether individuals truly perceive good or bad intentions in robot faces, with early ERP components being less susceptible to task demands.

Experiment 2

In Experiment 2, we used EEG to investigate the neural mechanisms and temporal dynamics underlying the attribution of mental states to robot faces. We tested a new sample of 30 native German speakers, using the same information manipulation for a subset of 18 out of the 36 robots presented in Experiment 1.

Using ERPs, we investigated different stages of processing in the brain with high temporal precision, aiming to determine which stages support the intentional vs. physical stance towards robots. Our analysis focused on four distinct stages: early visual perception (P1 component (*43*, *44*)), visual processing of faces and facial expressions (N170 component (*45*, *46*)), fast reflexive emotional responses (early posterior negativity, EPN (*34*, *47*)), and more elaborate stimulus evaluation (late positive potential, LPP (*29*, *47*)). If affective information modulates the amplitude of the P1 component, it suggests an influence on low-level visual perception, which has been repeatedly observed in object perception (*48–53*) but less so in face perception (*54*, *55*). We explored this phenomenon considering that robot faces may be perceived as situated between objects and faces. Similarly, if affective information influences the amplitude of the N170 component, i.e. how the face's visual features are structurally encoded, it implies that affective information prompts people to perceive emotional facial expressions in neutral robot faces (*56*, *57*). Additionally, an impact of information on reflexive emotional responses to robot faces should be reflected in increased EPN amplitudes (*47*, *58*), while an influence on more deliberate evaluations should result in increased LPP amplitudes (*28*, *29*, *47*). We anticipated that potential amplitude differences between negative and neutral information conditions would be more pronounced than those between positive and neutral conditions across ratings and ERP components, aligning with previous findings (*28*, *35*, *41*).

To gain deeper insight into how affective information and the ERP components of interest relate to perceived intentionality, we introduced an additional rating task inspired by the InStance questionnaire (*6*). This task assessed participants' perceived intentionality attributed to each individual robot and its behavior. We investigated

whether affective information about robots could enhance perceived intentionality. Additionally, by considering the intentionality rating as a covariate, we explored whether any specific ERP component is particularly linked to the likelihood of attributing intentionality to humanoid robots based on affective information.

Rating results

Facial expression ratings were collected after the learning part. Each robot's facial expression was rated twelve times (the stimulus repetitions served to increase signal-to-noise ratio in ERP analysis). The rating results reported here are based on the first rating of each stimulus. Similar to Experiment 1, facial expression ratings were influenced by affective information: expressions of robots associated with negative information were rated as more negative than those associated with neutral information (b = 0.69, 95% CI = [0.40, 0.97], $p < .001$). Expressions of robots in the positive information condition were rated as slightly more positive compared to the neutral information condition, but this comparison only yielded a statistical trend (b = 0.18, 95% CI = [-0.01, 0.37], p = .069).

Trustworthiness ratings were collected before and after learning, in order to get a baseline for each robot face's trustworthiness. Ratings of robots assigned to the three information conditions differed significantly after learning (neutral–negative: b = 1.83, 95% CI = [1.57, 2.10], $p < .001$; positive-neutral: b = 0.33, 95% CI = [0.11, 0.55], p = .003), but not in the baseline rating before learning (neutral–negative: b = 0.09, 95% CI = [-0.18, 0.35], p = .506; positive-neutral: b = 0.03, 95% CI = [-0.19, 0.25], p = .802). An additional analysis of variance confirmed the significance of the interaction between phase and information, $F(2, 973.78) = 114.95$; $p < .001$, showing that the overall differences between trustworthiness ratings in the different information conditions increased significantly between phases. Taken together, the results of Experiment 2 align well with those obtained in Experiment 1.

In addition, we assessed ratings of the perceived intentionality of each robot's behavior. On a scale from -50 (completely agree with a non-intentional explanation of behavior) to 50 (completely agree with an intentional explanation), mean ratings were -7.05 (95% CI = [-13.34, -0.69]). So on average, ratings tended more towards non-intentional than intentional explanations of the robots' behavior, but were also far from completely non-intentional. We next analyzed the impact of affective information on robots' perceived intentionality. Robots associated with negative information were rated as more intentional compared to robots associated with neutral information (b =-9.67, 95% CI = [-15.62, -3.73], $p < .001$). There was no significant difference between the impact of neutral and positive information on perceived intentionality (b = 3.19, 95% CI = [-2.76, 9.14], p = .293). This provides initial evidence that framing robots as exhibiting purposeful bad behavior influences the intentionality attributed to the robots. In sum, ratings of facial expression, robot trustworthiness and perceived intentionality were influenced by the valence of affective information learned about the robots.

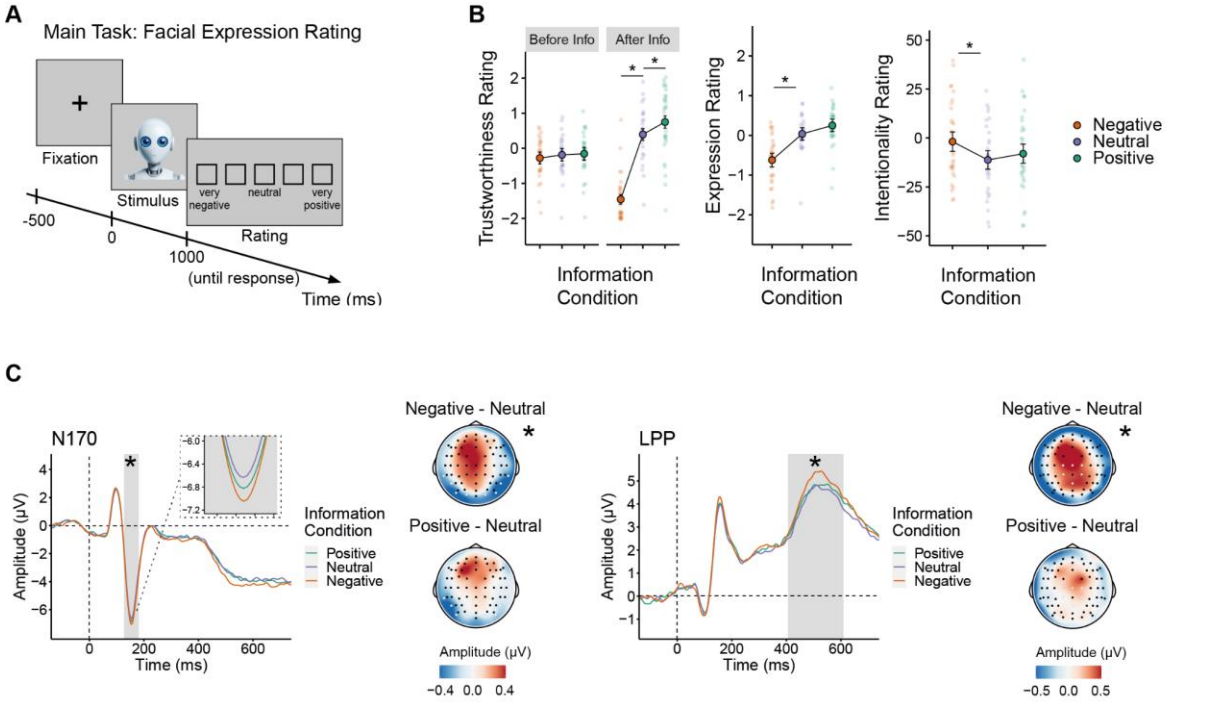

Fig. 2. EEG study results. (**A**). Trial sequence of the facial expression rating task. (**B**) Trustworthiness ratings before and after information acquisition, facial expression and intentionality ratings after information acquisition, categorized by information condition. Large dots denote group means with corresponding 95% CIs, while small dots indicate individual participant means. (**C**) Grand average ERPs for the N170 and LPP components collected during the facial expression rating task. Gray shading highlights time windows for the N170 and LPP. Scalp topographies illustrate differences between the information conditions, with channels included in the N170 and LPP regions of interest highlighted in white. Asterisks highlight statistically significant differences.

EEG results

We tested the effects of negative, neutral and positive information on the processing of robot faces presented during the facial expression rating task. Specifically, we analyzed ERP components associated with early perceptual processing (P1 and N170), reflexive emotional responses to visual input (EPN) and higher-level evaluation (LPP).

We observed significant influences of affective information on the N170 and LPP components, but not the P1 and EPN components (see Tables 3–4 and Figure 2). Both N170 and LPP amplitudes were significantly increased in the negative information condition compared to the neutral information condition. There were no significant differences between the neutral and the positive information conditions.

We explored whether participants' ratings of each robot's perceived intentionality would be associated with the impact of information on the N170 and LPP components. We calculated an additional LMM for each component including centered intentionality scores as a covariate. We discovered an interaction between perceived intentionality and information (negative vs. neutral) in the N170 component

(b = -0.29, 95% CI = [-0.57, -0.01], p = .045). Specifically, higher scores of intentionality were linked to a larger impact of negative information on N170 amplitudes. Additionally, in the LPP, we observed a significant main effect of perceived intentionality, where higher scores of the covariate were associated with lower LPP amplitudes (b = -0.14, 95% CI = [-0.28, -0.01], p = 0.033). However, we found no interactions between perceived intentionality and information.

Table 3. Perception-related EEG results. Results of linear mixed model analyses of the P1 and N170 components

	P1 component			**N170 component**		
Predictors	*b*	95% CI	*p*-value	*b*	95% CI	*p*-value
Intercept	3.11	[1.79, 4.43]	**<.001**	-5.22	[-6.53, -3.92]	**<.001**
Information(Neu-Pos)	-0.12	[-0.43, 0.19]	.455	0.21	[-0.06, 0.47]	.127
Information(Neg-Neu)	0.10	[-0.21, 0.42]	.524	-0.34	[-0.60, -0.08]	**.012**
Random Effects			*SD*			*SD*
Participants			3.45			3.35
Stimuli			0.59			0.82
Residual			5.24			4.42
Deviance	39670.08			37501.78		
log-Likelihood	-19835.04			-18750.89		

Note. Information Conditions: Neg = Negative, Neu = Neutral, Pos = Positive. Boldface indicates statistical significance at α = .05.

Table 4. Emotion-related EEG results. Results of linear mixed model analyses of the EPN and LPP components

	EPN			**LPP**		
Predictors	*b*	95% CI	*p*-value	*b*	95% CI	*p*-value
Intercept	-1.02	[-2.38, 0.33]	0.135	4.35	[3.60, 5.10]	**<.001**
Information(Neu-Pos)	-0.06	[-0.34, 0.21]	0.657	-0.16	[-0.48, 0.16]	.322
Information(Neg-Neu)	-0.19	[-0.46, 0.09]	0.181	0.48	[0.18, 0.78]	**.003**
Random Effects			*SD*			*SD*
Participants			3.35			1.94
Information(Neu-Neg)			-			0.447
Information(Pos-Neu)			-			0.36
Stimuli			1.14			0.35
Residual			4.60			4.45
Deviance	38020.26			37543.59		
log-Likelihood	-19010.13			-18771.79		

Note. Information Conditions: Neg = Negative, Neu = Neutral, Pos = Positive. Boldface indicates statistical significance at α = .05.

Discussion of Experiment 2

In line with the findings of Experiment 1, our manipulation of affective information significantly influenced trustworthiness and facial expression ratings. We also observed an impact of affective information on perceived intentionality, indicating that behaviors were judged as more intentional when robots were linked to negative backstories. This corresponds with the asymmetry often noted in judgments of human behavior, where people tend to attribute greater intentionality for actions with negative outcomes compared to positive ones (*59*, *60*).

In ERPs, we observed significant information effects on two distinct processing stages: the N170 component, associated with visual perception, and the later LPP component, reflecting more elaborate stimulus evaluation. The N170 is most commonly associated with structural visual encoding of faces, and it has been shown to be sensitive to manipulations of facial expression, as well as the realism of face images (*45*, *61*, *62*). Thus, robot faces associated with negative information were perceived either as displaying an emotional facial expression or as more "face-like," both of which support the idea that affective information enables attribution of mental states to robot faces at an early, potentially automatic perceptual stage. Similarly, the information effect on the LPP indicates that negatively framed robots are evaluated as more emotionally relevant compared to neutrally framed robots (*47*, *63*). Affective information did not influence the P1 component, indicating that low-level visual processing remained unaffected. Since previous studies have demonstrated knowledge effects in the P1 for objects (*48*, *50*), this suggests that the visual processing of robot faces leaned more towards face perception rather than object perception. Notably, contrary to our prediction, affective information did not influence the EPN component, suggesting a lack of early reflexive emotional response typically observed when human faces are associated with analogous affective information (*28*, *35*, *41*, *64*). We elaborate further on these findings in the general discussion.

The absence of a significant difference between the positive and neutral information conditions in ERPs was partly expected, given our prediction that the impact of negative information would be more pronounced than that of positive information. The fact that ERPs in the positive and neutral conditions did not differ at all could be due to a situational effect during the learning session, where the presence of highly negative stories overshadowed the more subtle differences between positive and neutral stories. Participants appeared to perceive the neutral stories as somewhat positive, as indicated by a positive shift in both facial expression ratings (Experiment 1) and trustworthiness ratings (Experiment 2) from the pre- to post-learning phases. This phenomenon aligns with previous studies that used vignettes describing human social behavior (*35*, *41*).

By including perceived intentionality as a covariate, we delved deeper into understanding the interplay between affective information, intentionality, and the neurocognitive processes represented by the N170 and LPP components. The interaction effect observed between information type and perceived intentionality on the N170 component provides further evidence linking the perceptual impact of negative information—where neutral robot faces are perceived as displaying negative expressions—to the attribution of intentionality. There are two plausible interpretations of these findings: Firstly, enhanced visual face processing, influenced by affective information, increases the likelihood of individuals attributing mental

states to a robot. Alternatively, it is possible that robot faces inherently more predisposed to attribution of intentionality, perhaps due to their appearance, are also more susceptible to having facial expressions inferred onto them. This open question warrants further investigation in future research.

Moreover, higher scores of perceived intentionality were correlated with a reduction in the LPP component, regardless of affective information. This observation could suggest that the cognitive effort involved in deliberate emotional evaluation diminishes for robot faces perceived as more intentional (*54*, *65*). It is plausible that faces that are more easily ascribed intentionality may give us less pause during social-emotional evaluation, suggesting facilitated use of theory of mind.

GENERAL DISCUSSION

Humanoid social robots present an intriguing puzzle: people often intuitively interact with robots as they would with another human, even though they may be explicitly aware that robots are mechanical artifacts that do not share the same cognitive abilities as humans (*8*, *11*). Thus, people variably apply a "physical" or an "intentional" stance towards robots, interpreting their behavior either by invoking mechanical or mental causation. In this study, we investigated how these different modes of construing robots manifest during the perception of robot faces: to what extent does processing in the brain reveal attribution of mental states? Two pre-registered experiments tested the prediction that people read intentions and emotional expressions into objectively neutral robot faces based on information they learned about the robots' previous social behavior, framed as either positive, neutral, or negative.

Experiment 1 established that the valence of information about robots' behavior influenced people's ratings of robots' trustworthiness and facial expressions. The same robots were distrusted when associated with negative information (e.g., the robot reports children to the secret police), but were trusted when associated with positive information (e.g., the robot teaches social skills to people with autism). Crucially, participants also rated the objectively neutral facial expressions of robots as more negative when paired with negative information and as more positive when paired with positive information, compared to neutral information. These results provided initial evidence that people in fact read good or bad intentions into robot faces—an effect previously observed only during the perception of human faces (*28*, *35*).

Experiment 2 investigated the brain dynamics underlying the attribution of mental states to robot faces based on information about their behavior. Event-related potentials allowed to track face processing from early to late stages with high temporal precision. Affective information influenced the processing of robot faces at two stages: perceptual encoding, indexed in the N170 ERP component, and more elaborate stimulus evaluation, reflected in the LPP component. The modulation of the N170 highlights the speed of mental state attribution: changes to perceptual processing occur within 130 to 180 ms, at a processing stage that typically precedes conscious access (*66*). This effect on visual processing shows that people not only evaluate robot faces differently, but literally see bad intentions in a robot face associated with negative behavior. The later effect of information on the LPP indicates that negatively framed robots are also evaluated as more emotionally relevant compared to neutrally framed robots. Interestingly, and against our

prediction, affective information did not influence the EPN during the processing of robot faces, suggesting the absence of an early, reflexive emotional response that is typically elicited when human faces are associated with similar affective information (*28, 35*). This reduced malleability of affective responses to robots on a trial-by-trial basis may also explain why brain activity associated with social bonding does not increase over long intervals of interaction with robots, as it does in human-human interaction (*67*).

Taken together, it appears that both the intentional and the physical stance are reflected at different stages of processing in the brain. In line with the intentional stance, we rapidly and automatically read mental states into robot faces during visual perception (N170). We also explicitly evaluate robots in the light of acquired information, as shown in ratings and the LPP. However, in our fast emotional reaction (EPN), we are not as affected as we would be by comparable negative information about other humans. This suggests that the brain's emotional response to humanoid robots is influenced more by the physical stance than by social or intentional aspects. In conclusion, while we do engage perceptual and cognitive aspects of social cognition to process social robots, our findings on emotional processing suggest that we do not experience them as fully fledged intentional and social agents.

Our exploratory analyses uncovered further connections between perceived intentionality, affective information, and brain responses that support this view. Firstly, robots were perceived as more intentional when paired with negative rather than neutral information, suggesting that framing robots with emotional background information enhances the attribution of mental causes for their actions. Secondly, the perceived intentionality of individual robots was statistically associated with the impact of negative information on the N170 component. This correlation underscores the idea that the intentionality we attribute to a robot face based on affective information is supported by a visual process, something we can automatically perceive in its facial expression.

Implications for social robot design and policy

Social robot design has often sought to create a social interface (*12*), leveraging people's natural tendency to treat things as intentional beings when a social threshold is passed (*63, 68*). Our findings demonstrate that attention needs to be paid not only to the appearance of a robot, but also to the framing that informs users' prior knowledge and expectations. Our results highlight the interplay between top-down and bottom-up mechanisms in the brain, i.e. processing that starts from expectations vs. from sensory information, in shaping the appearance of and capacities attributed to robots. It becomes evident that solely relying on design characteristics does not fully determine the perception of a robot, as contextual factors like affective information significantly alter its perceived trustworthiness and even facial expression. Thus, adopting a less detailed bottom-up approach in design (e.g., not conveying intentionality through overly realistic facial features) may allow contextual cues and perceivers' own top-down predictions to shape processing of the robot effectively and facilitate fluent interactions. With facial expressions reduced to essential features, humanoid robot faces can open a canvas for projecting users' expectations, reducing prediction errors, and consequently mitigating adverse outcomes like the uncanny valley effect (*69, 70*). Future research could further explore the dynamics between specific design attributes and psychological factors like affective information and contextual cues.

Regarding policy implications, our results emphasize that acceptance and moral judgments of social robots hinge on the interplay of intentionality and emotional valence. Our findings show that (negative) emotional information can increase perceived intentionality, both on the level of explicit ratings and automatic perceptual processing in the brain. Intentionality plays a key role in moral judgment, both in that mindedness may be a prerequisite for moral responsibility and that moral transgressions may necessitate intentional agency (*7*, *63*, *71–73*). Despite the absence of actual intentionality in current humanoid robots, our results show that people's perception of intentionality is influenced by semantic and emotional cues, potentially leading to moral judgments based on these attributions. Since robots are automatically perceived as more intentional when associated with negative actions, they may in practice be judged as morally responsible for negative actions even though they are not. There have been concerns on theoretical grounds that this can distract from the proper attribution of responsibility causing "responsibility gaps", situations in which no party is held accountable and therefore potentially harmful outcomes caused by artificial agents are not properly addressed (*74–77*). Thus, psychological variables like beliefs and affective information should inform policies legislating the moral responsibility of artificial social agents.

Limitations

One potential limitation of our study is that it did not involve interaction with real robots but rather examined effects on the perception of images of robot faces. This was due to experimental design considerations: While live interaction in the lab typically involves only one or a few robots with a limited number of repetitions, our design, which has been well-tested in previous EEG studies (28, 29, 35), enabled us to test a larger set of diverse robot images and characterizations. Consequently, our results are more likely to generalize to the perception of various humanoid robots and areas of robot behavior. A large stimulus set was also required to obtain high-quality EEG data with the requisite number of repeated measurements in the different information conditions. Further research can explore how the affective information-based mind attribution observed here translates into responses during live social interactions with robots.

Conclusion

In our study, we discovered that humans rapidly attribute mental states to humanoid robots following exposure to affective information regarding the robots' behavior. This phenomenon is observed during both perceptual processing and more deliberate evaluation of robot faces, but notably absent during fast emotional processing, which lacks a component seen in the social perception of other humans. These findings imply that the processing of social robots oscillates between being perceived as mindless machines and intentional agents, contingent upon the stage of perceptual and emotional processing in the brain. Such nuanced insights into the neural, cognitive, and emotional mechanisms underlying the perception of robots have significant implications for social robot design and the formulation of policies regarding the moral responsibility of artificial agents. These considerations are especially pertinent given the projected proliferation of such agents in our societies.

MATERIALS AND METHODS

Experiment 1

The preregistration for Experiment 1 can be accessed at https://osf.io/qytra.

Participants

Sixty participants (22 cisgender women, 38 cisgender men; mean age 28 years, range 18–39) were recruited from Prolific (prolific.com) and received monetary compensation. Thirty participants were German speakers (6 cisgender women, 24 cisgender men; mean age 28 years, range 19–39), and 30 were English speakers (16 cisgender women, 14 cisgender men; mean age 27 years, range 18–37). The sample size, a multiple of three, was determined based on similar studies (28, 29, 35), ensuring counterbalancing across three information conditions. 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. Participants provided informed written consent before participation.

Materials

The picture stimuli comprised 36 full-color frontal portrait photographs featuring existing humanoid robots, each displaying approximately neutral facial expressions (refer to Fig. 1 for an illustration; names and sources of the robots used are listed in the Supplementary Materials). The featured robots have been developed for commercial (e.g. entertainment or personal service) or research (e.g. psychology or robotics) purposes. The images were found on the online database abotdatabase.info (*78*) or on relevant commercial, news or academic websites. 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 (2.7 x 3.5 cm) 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.

We recorded 36 spoken stories (mean durations: English = 18.1 s, German = 17.9 s) with affectively positive (e.g., the robot teaches social skills to people with autism), neutral (e.g., the robot assembles orders at a warehouse) or negative (e.g., the robot reports children to the secret police) information about the robots (see Fig. 1 for examples; for the complete set of stories, see Supplementary Materials). To make the stories plausible, they were based on news stories about developments in robotics and AI so that the robots' fictional actions resembled functions carried out by real existing robots and AI (e.g. commercial, educational, military or medical). Neutral stories described morally neutral functionality, while the positive and negative stories described actions that are commonly held to be kind and helpful or contemptible and cruel, respectively. The wording of the stories uniformly implied that the robots are able to learn and can make decisions.

Per participant, each image was paired with a different story, so that 12 images were presented with positive stories, 12 images were presented with neutral stories and 12 images were presented with negative stories. The stories were presented to the participants auditorily and recordings started playing when an image appeared. Across participants, matching of images and stories was counterbalanced, so that each robot was shown an equal number of times with negative, neutral and positive information.

To ensure that the stories themselves would indeed be perceived as positive, neutral or negative according to the respective condition, we pretested the stories with a separate sample of participants (N = 15). The valence of the stories was rated as expected. Stories belonging to the negative condition were rated as more negative and stories belonging to the positive condition were rated as more positive than stories belonging to the neutral condition. Negative stories were also rated as more arousing than stories in the other two conditions and neutral stories were rated the least arousing. Details of the pretest results are provided in the Supplementary Materials.

Procedure

In the first section of the experiment, participants gave ratings of the robots' facial expressions (pre-learning), to be later compared with ratings after information acquisition (post-learning). Participants were presented with all 36 robots one at a time in a random order and asked to rate the robots' facial expressions on 7-point Likert scales ranging from very negative to very positive. The middle of the scale was marked neutral and the order of the anchors from left to right was switched for 50% of all participants.

In the main section of the experiment, the robots were presented in different blocks, each featuring six robots. The participants were instructed to pay close attention to the information they were about to hear, as they would be required to respond to questions about the robots at regular intervals. Each robot face was paired with a different story, which was automatically played when an image of a robot appeared on the screen. In every block, two robots per information condition (negative, neutral, positive) were presented in random order. Immediately after the presentation of a robot, the participants rated the robot's trustworthiness on 7-point Likert scales ranging from not at all trustworthy to very trustworthy (with the order of the anchors from left to right counterbalanced across participants). Then the next robot was presented. After they had seen and rated the trustworthiness of all six robots within a block, the participants rated their facial expressions in succession. Between blocks, the participants responded to multiple choice questions about the robots with four possible answers to verify that they were paying attention.

After the main section of the experiment, participants were asked to complete questionnaires and to answer a series of questions about the experiment. Participants completed the Attitude towards Artificial Intelligence Scale (*79*). They answered questions about whether they had researched information about the robots or had been distracted during the experiment. Participants rated the perceived intentionality and deliberateness of the robots' actions (collectively) on a 7-point Likert scale from "not at all" to "very." Additional questions included whether they had previously known any of the information presented, and whether they distrusted any of it, with responses given as yes or no. Those who distrusted the information estimated the percentage of stories to which this applied. Finally, participants provided feedback or noted any

concerns or thoughts during the experiment. Afterward, they were debriefed and informed that none of the information presented pertained to any of the featured robots.

Data exclusion Criteria

The participants were recruited to have no specialized prior knowledge about robots. Exclusion criteria stated that participants may not recognize more than four of the robots used as stimuli in the experiment; however, none of the participants selected more than four robots from a list presented at the start and so no participants were excluded for this reason. Further exclusion criteria were based on general task performance. Participants were required to respond to multiple choice sanity checks throughout the experiment. Less than 50% correct responses would have led to data exclusion; however, no participants were excluded for this reason. Similarly, if participants had continuously given the same score, if they had given clearly random scores in response to the tasks, or if the manipulation had obviously failed (e.g., if robots paired with negative stories were rated as highly trustworthy or vice versa) then data would have been excluded from analysis; no data were excluded for this reason. Finally, participants were asked after the main experiment if they had had strong doubts about the veracity of the stories, if they had googled information about the robots during the experiment, or if they had been distracted. Two participants were excluded because they reported to have been highly distracted and one participant was excluded for reporting to have googled information about the robots. One further participant was excluded for participating twice in this experiment (i.e. we excluded the second attempt). After excluding these participants, data collection continued until 30 complete data sets each from German and English speakers were obtained.

Statistical Analysis

Facial expression and trustworthiness rating data were analyzed using linear mixed effects models (LMMs)(*80*). Information (negative, neutral, positive) and, if applicable, phase (pre vs. post learning) were modeled as fixed effects, coded as sliding difference contrasts. For the facial expression ratings in Experiment 1 and the trustworthiness ratings in Experiment 2, nested LMMs (information type nested within phase) were used to analyze effects of the information conditions separately in the pre-learning and the post-learning phase. The significance of the interaction term between information type and phase was then tested using the ANOVA-function of the R Stats Package (*81*). We modeled random intercepts for both participants and items (robot images), as well as random slopes for the independent variable information type across participants and items, whenever supported by the models. We employed a backward model selection approach (*82*) to specify the maximal random effects structure compatible with model convergence.

Experiment 2

Experiment 2 was pre-registered under https://osf.io/c8va7. The procedure and materials were similar to Experiment 1, with the following differences due to the EEG setting: 1) Experiment 2 was divided into two main parts: a learning phase (without EEG) in which participants acquired and rehearsed information about all robots, and a subsequent EEG part in which participants performed rating tasks on the robot pictures; 2) only a subset of 18 robot stimuli was used to keep the length of the

experiment and amount of information to memorize reasonable; 3) the original long versions of the stories about each robot were only presented once—after that, shorter versions were presented for rehearsal; 4) Rating scales were 5-point instead of 7-point-Likert scales due to the experimental setup in the EEG laboratory; 5) Trustworthiness ratings were collected both before and after learning, whereas facial expression ratings were only collected after learning; and 6) a new task was added at the end to collect perceived intentionality ratings for each robot.

Participants

Thirty participants, all German speakers (24 cisgender women, 6 cisgender men; mean age 25.4 years, range 18–36), took part in this study and received monetary compensation or course credit. The datasets of five participants were replaced due to expressing strong doubts about the veracity of the information provided about the robots during debriefing. To ensure counterbalancing, the target sample size needed to be a multiple of three. A simulation-based power analysis, conducted using the R package simr (*83*), helped determine the sample size. With 1000 random simulations of the specified Linear Mixed Model (LMM), the analysis revealed that testing 30 participants would yield 91% power (95% CI = [89.05, 92.7]) to detect a mean difference of 0.4 µV between the negative (or positive) condition and the neutral condition in ERPs. 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. Participants provided written consent before participating.

Materials

A subset of 18 robot pictures with the corresponding stories was selected from the stimulus set of Experiment 1 (see Table S2 in the Supplementary Materials). This subset comprised stimuli that had yielded the largest information effects on facial expression ratings in Experiment 1. Stimuli were presented on a gray background on a 19-inch LCD monitor with a resolution of 1280 × 1024 pixels and a 75-Hz refresh rate. During the rating tasks, robot faces were displayed with a size subtending 6.03° vertical and 6.02° horizontal visual angles (viewing distance: 70 cm). We recorded additional short versions of each robot story, focusing on a central part of the robots' behavior (e.g., "this robot works in the checkroom of a nightclub"). After each story's long version was presented once, further repetitions used the short versions. The long and short versions of all robot stories are listed in the Supplementary Materials.

As an additional dependent variable, Experiment 2 included a rating of each robot's perceived intentionality. The format of the rating task was inspired by the InStance questionnaire (*6*). For each robot, two statements about the potential motivation for its behavior were presented, one description in mechanistic terms (e.g., "the robot reacts to moving objects") and one in intentional terms (e.g., "the robot likes to play table tennis"). Participants moved a slider towards the statement that they believed better captured the robot's behavior, on a scale from -50 (full agreement with mechanistic description) to 50 (full agreement with intentional description). Details on the intentionality questionnaire are provided in the Supplementary Materials.

Procedure

The experiment started with participants rating the trustworthiness of each robot picture on a 5-point Likert scale (pre-learning). Subsequently, in the learning phase, which lasted approximately 30 minutes, participants acquired and rehearsed information about 18 robots. Robots were introduced in three sets of six (selected pseudorandomly, with two robots each associated with positive, neutral, and negative information). Within each set, participants first encountered each robot along with the long version of its associated story, then once again with the short version. This was followed by a short rehearsal, during which participants verbally recalled key details from the robot's story while an experimenter noted the accuracy of responses. Following this rehearsal, the six robots were presented once more with the short story version. This process was repeated for the remaining sets of robots. Upon completion of this phase, all 18 robots were presented two additional times with the short story versions. Finally, the learning session concluded with participants once again recalling story keywords for all 18 robots. In total, each robot was presented five times with its corresponding story.

Following the learning phase, participants underwent EEG electrode placement and preparation, which took approximately 45 minutes. The EEG session started with another trustworthiness rating of each robot (post-learning). This was followed by the expression rating task, which formed the basis for ERP analyses. All 18 robots were rated for facial expression valence on a 5-point Likert scale twelve times in random order. Detailed information on stimulus timing can be found in Fig. 2. Next, each robot was rated once for perceived intentionality. For all rating tasks, the position of the rating anchors (e.g., very positive, very negative) was counterbalanced across participants. The EEG session concluded with the recording of prototypical eye movements used for subsequent artifact correction.

After the EEG experiment, participants completed several questionnaires. Initially, they answered 4-alternative multiple-choice questions assessing their recollection of the stories associated with all 18 robots. Following this, participants responded to the same questionnaires used in Experiment 1, which included assessments of attitudes towards artificial intelligence (*79*), awareness of experimental hypotheses (*84*), perceptions of general robot intentionality, and indications of any distrust towards presented information or familiarity with featured robots prior to the experiment. Additionally, participants filled out the Edinburgh Handedness Inventory (*85*). Finally, participants underwent debriefing, where they were informed that none of the information presented pertained to any of the featured robots.

EEG recording and analysis

The EEG data were acquired using Ag/AgCl electrodes placed at 64 scalp sites according to the extended 10–20 system, with a sampling rate of 500 Hz and all electrodes referenced to the left mastoid. The electrooculogram (EOG) was recorded using a bipolar vertical EOG channel consisting of electrodes Fp1 - IO1 and a bipolar horizontal EOG channel consisting of electrodes F9 - F10. During recording, a low-cut-off filter (0.032 Hz) was applied, and electrode impedances were maintained below 10 kΩ. Post-experiment, a calibration procedure was conducted to capture prototypical eye movements for subsequent artifact correction.

Offline processing for single-trial ERP analysis followed a pipeline detailed in a previous study (*86*), re-implemented using functions of MNE Python (*87*), available at https://github.com/alexenge/hu-neuro-pipeline. Continuous EEG data were re-referenced to a common average reference, and eye movement artifacts were removed using a spatio-temporal dipole modeling procedure with the BESA software (*88*). The corrected data were low-pass filtered at 40 Hz, segmented into epochs of -500 to 1500 ms relative to face stimulus onset, and baseline-corrected using the 200 ms pre-stimulus interval. For each participant, 72 segments were created in each information condition (six robots per information condition $\times$ 12 repetitions), excluding segments containing artifacts (amplitudes over ±150 µV, or changing by more than 50 µV between samples).

Single-trial mean amplitudes were obtained for the P1, N170, EPN, and LPP components by averaging across pre-registered time windows and electrode sites typical for each component. The P1 was averaged at parieto-occipital electrode sites (O1, O2, Oz, PO7, PO8) in the time window 76–116 ms centered around the average P1-peak of the ERP collapsed across all conditions. The N170 was averaged at parieto-occipital electrode sites (TP9, TP10, P7, P8, PO9, PO10, O1, O2) centered around its average peak, between 129–179 ms. The EPN was averaged at posterior electrodes (PO7, PO8, PO9, PO10, TP9, TP10) between 220–350 ms. The LPP was averaged at centro-parietal sites (Pz, Cz, C1, C2, CP1, CP2) between 408–608 ms. For statistical analysis, LMMs were specified following the same procedure as in Experiment 1.

Data exclusion criteria

Several criteria for the exclusion of participants' datasets were preregistered. Three criteria concerned participants' performance across different tasks: Failing the memory test on the learned information after the experiment (more than 5 incorrect answers out of 18 multiple choice questions), continuously giving the same score or clearly random scores in response to the ratings tasks, systematically rating robots that were paired with negative stories as highly trustworthy or vice versa, indicating a failed information manipulation. Two criteria concerned participants' prior knowledge or beliefs about the experiment: prior knowledge of more than 4 robots and indicating strong doubts about the veracity of the robots' backstories during debriefing, doubting the veracity of over 50% of the stories. The final criterion was EEG data quality, specifically excessive EEG artifacts resulting in less than 30 out of 72 trials in the facial expression rating task per information condition after artifact rejection. The only criterion that led to data exclusions was strong doubts in the veracity of the robots' backstories, resulting in the replacement of five participants.

Supplementary Materials

Supplementary materials include:

Fig. S1. Ratings of stories
Fig. S2. Intentionality questionnaire instructions
Table S1. Rating results by group
Table S2. Robot list

REFERENCES

1. T. Belpaeme, J. Kennedy, A. Ramachandran, B. Scassellati, F. Tanaka, Social robots for education: A review. *Sci. Robot.* **3**, eaat5954 (2018).
2. A. Paolillo, F. Colella, N. Nosengo, F. Schiano, W. Stewart, D. Zambrano, I. Chappuis, R. Lalive, D. Floreano, How to compete with robots by assessing job automation risks and resilient alternatives. *Sci. Robot.* **7**, eabg5561 (2022).
3. C. Breazeal, K. Dautenhahn, T. Kanda, "Social Robotics" in *Springer Handbook of Robotics*, B. Siciliano, O. Khatib, Eds. (Springer International Publishing, Cham, 2016; https://doi.org/10.1007/978-3-319-32552-1_72), pp. 1935–1972.
4. E. S. Cross, R. Ramsey, Mind Meets Machine: Towards a Cognitive Science of Human–Machine Interactions. *Trends Cogn. Sci.* **25**, 200–212 (2021).
5. A. Henschel, R. Hortensius, E. S. Cross, Social Cognition in the Age of Human–Robot Interaction. *Trends Neurosci.* **43**, 373–384 (2020).
6. S. Marchesi, D. Ghiglino, F. Ciardo, J. Perez-Osorio, E. Baykara, A. Wykowska, Do We Adopt the Intentional Stance Toward Humanoid Robots? *Front. Psychol.* **10** (2019).
7. K. Gray, L. Young, A. Waytz, Mind Perception Is the Essence of Morality. *Psychol. Inq.* **23**, 101–124 (2012).
8. H. H. Clark, K. Fischer, Social robots as depictions of social agents. *Behav. Brain Sci.*, 1–33 (2022).
9. D. C. Dennett, *The Intentional Stance* (MIT press, 1989).
10. J. Perez-Osorio, A. Wykowska, Adopting the intentional stance toward natural and artificial agents. *Philos. Psychol.* **33**, 369–395 (2020).
11. M. M. A. de Graaf, An Ethical Evaluation of Human–Robot Relationships. *Int. J. Soc. Robot.* **8**, 589–598 (2016).
12. D. C. Dryer, Getting personal with computers: How to design personalities for agents. *Appl. Artif. Intell.* **13**, 273–295 (1999).
13. N. Epley, A. Waytz, J. T. Cacioppo, On seeing human: A three-factor theory of anthropomorphism. *Psychol. Rev.* **114**, 864–886 (2007).
14. K. Gray, D. M. Wegner, Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition* **125**, 125–130 (2012).
15. L. Onnasch, E. Roesler, Anthropomorphizing Robots: The Effect of Framing in Human-Robot Collaboration. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.* **63**, 1311–1315 (2019).
16. M. M. A. de Graaf, S. Ben Allouch, J. A. G. M. van Dijk, "What Makes Robots Social?: A User's Perspective on Characteristics for Social Human-Robot Interaction" in *Social Robotics*, A. Tapus, E. André, J.-C. Martin, F. Ferland, M. Ammi, Eds. (Springer International Publishing, Cham, 2015), pp. 184–193.
17. O. L. Jacobs, K. Gazzaz, A. Kingstone, Mind the Robot! Variation in Attributions of Mind to a Wide Set of Real and Fictional Robots. *Int. J. Soc. Robot.* **14**, 529–537 (2022).
18. T. W. Kim, A. Duhachek, Artificial Intelligence and Persuasion: A Construal-Level Account. *Psychol. Sci.* **31**, 363–380 (2020).
19. E. Broadbent, V. Kumar, X. Li, J. Sollers, R. Q. Stafford, B. A. MacDonald, D. M.

Wegner, Robots with Display Screens: A Robot with a More Humanlike Face Display Is Perceived To Have More Mind and a Better Personality. *PLoS ONE* **8**, e72589 (2013).
20. S. Ceh, E. J. Vanman, The Robots are Coming! The Robots are Coming! Fear and Empathy for Human-like Entities. [Preprint] (2018). https://doi.org/10.31234/osf.io/4cr2u.
21. B. R. Duffy, Anthropomorphism and the social robot. *Robot. Auton. Syst.* **42**, 177–190 (2003).
22. R. Hortensius, F. Hekele, E. S. Cross, The Perception of Emotion in Artificial Agents. *IEEE Trans. Cogn. Dev. Syst.* **10**, 852–864 (2018).
23. M. C. Martini, C. A. Gonzalez, E. Wiese, Seeing Minds in Others – Can Agents with Robotic Appearance Have Human-Like Preferences? *PLOS ONE* **11**, e0146310 (2016).
24. A. Waytz, K. Gray, N. Epley, D. M. Wegner, Causes and consequences of mind perception. *Trends Cogn. Sci.* **14**, 383–388 (2010).
25. N. Spatola, O. A. Wudarczyk, Ascribing emotions to robots: Explicit and implicit attribution of emotions and perceived robot anthropomorphism. *Comput. Hum. Behav.* **124**, 106934 (2021).
26. R. Q. Stafford, B. A. MacDonald, C. Jayawardena, D. M. Wegner, E. Broadbent, Does the Robot Have a Mind? Mind Perception and Attitudes Towards Robots Predict Use of an Eldercare Robot. *Int. J. Soc. Robot.* **6**, 17–32 (2014).
27. S. Thellman, M. De Graaf, T. Ziemke, Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings. *ACM Trans. Hum.-Robot Interact.* **11**, 1–51 (2022).
28. R. Abdel Rahman, Facing Good and Evil: Early Brain Signatures of Affective Biographical Knowledge in Face Recognition. *Emotion* **11**, 1397–1405 (2011).
29. J. Baum, M. Rabovsky, S. B. Rose, R. Abdel Rahman, Clear judgments based on unclear evidence: Person evaluation is strongly influenced by untrustworthy gossip. *Emotion* **20**, 248–260 (2020).
30. M. Maier, F. Blume, P. Bideau, O. Hellwich, R. Abdel Rahman, Knowledge-augmented face perception: Prospects for the Bayesian brain-framework to align AI and human vision. *Conscious. Cogn.* **101**, 103301 (2022).
31. M. Otten, A. K. Seth, Y. Pinto, A Social Bayesian Brain: How Social Knowledge Can Shape Visual Perception. *Brain Cogn.* **112**, 69–77 (2017).
32. M. J. Wieser, A. B. M. Gerdes, I. Büngel, K. A. Schwarz, A. Mühlberger, P. Pauli, Not so harmless anymore: How context impacts the perception and electrocortical processing of neutral faces. *NeuroImage* **92**, 74–82 (2014).
33. R. Hassin, Y. Trope, Facing faces: Studies on the cognitive aspects of physiognomy. *J. Pers. Soc. Psychol.* **78**, 837–852 (2000).
34. A. Schacht, W. Sommer, Emotions in Word and Face Processing: Early and Late Cortical Responses. *Brain Cogn.* **69**, 538–550 (2009).
35. F. Suess, M. Rabovsky, R. Abdel Rahman, Perceiving emotions in neutral faces: expression processing is biased by affective person knowledge. *Soc. Cogn. Affect. Neurosci.* **10**, 531–536 (2015).
36. A. Clark, Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behav. Brain Sci.* **36**, 181–204 (2013).
37. K. Friston, The free-energy principle: a rough guide to the brain? *Trends Cogn. Sci.* **13**, 293–301 (2009).
38. T. S. Lee, D. Mumford, Hierarchical Bayesian inference in the visual cortex. *J. Opt. Soc. Am. A* **20**, 1434–1448 (2003).
39. G. Lupyan, R. Abdel Rahman, L. Boroditsky, A. Clark, Effects of Language on Visual Perception. *Trends Cogn. Sci.* **24**, 930–944 (2020).

40. C. Press, D. Yon, Perceptual Prediction: Rapidly Making Sense of a Noisy World. *Curr. Biol.* **29**, R751–R753 (2019).
41. J. Baum, R. Abdel Rahman, Emotional News Affects Social Judgments Independent of Perceived Media Credibility. *Soc. Cogn. Affect. Neurosci.* **16**, 280–291 (2021).
42. J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, W. Manassra, P. Dhariwal, C. Chu, Y. Jiao, A. Ramesh, Improving Image Generation with Better Captions.
43. F. Di Russo, A. Martinez, M. I. Sereno, S. Pitzalis, S. A. Hillyard, Cortical sources of the early components of the visual evoked potential. *Hum. Brain Mapp.* **15**, 95–111 (2002).
44. J. D. Haynes, G. Roth, M. Stadler, H. J. Heinze, Neuromagnetic Correlates of Perceived Contrast in Primary Visual Cortex. *J. Neurophysiol.* **89**, 2655–2666 (2003).
45. S. Bentin, T. Allison, A. Puce, E. Perez, G. McCarthy, Electrophysiological Studies of Face Perception in Humans. *J. Cogn. Neurosci.* **8**, 551–565 (1996).
46. M. Eimer, M. Kiss, S. Nicholas, Response Profile of the Face-Sensitive N170 Component: A Rapid Adaptation Study. *Cereb. Cortex* **20**, 2442–2452 (2010).
47. H. T. Schupp, T. Flaisch, J. Stockburger, M. Junghöfer, Emotion and Attention: Event-Related Brain Potential Studies. *Prog. Brain Res.* **156**, 31–51 (2006).
48. R. Abdel Rahman, W. Sommer, Seeing what we know and understand: How knowledge shapes perception. *Psychon. Bull. Rev.* **15**, 1055–1063 (2008).
49. A. Enge, F. Süß, R. Abdel Rahman, Instant Effects of Semantic Information on Visual Perception. *J. Neurosci. Off. J. Soc. Neurosci.* **43**, 4896–4906 (2023).
50. M. Maier, P. Glage, A. Hohlfeld, R. Abdel Rahman, Does the semantic content of verbal categories influence categorical perception? An ERP study. *Brain Cogn.* **91**, 1–10 (2014).
51. M. Maier, R. Abdel Rahman, Native Language Promotes Access to Visual Consciousness. *Psychol. Sci.* **29**, 1757–1772 (2018).
52. M. Maier, R. Abdel Rahman, Transient and Long-Term Linguistic Influences on Visual Perception: Shifting Brain Dynamics With Memory Consolidation. *Lang. Learn.*, lang.12631 (2024).
53. P. D. Weller, M. Rabovsky, R. Abdel Rahman, Semantic Knowledge Enhances Conscious Awareness of Visual Objects. *J. Cogn. Neurosci.* **31**, 1216–1226 (2019).
54. A. Eiserbeck, M. Maier, J. Baum, R. Abdel Rahman, Deepfake smiles matter less—the psychological and neural impact of presumed AI-generated faces. *Sci. Rep.* **13**, 16111 (2023).
55. M. J. Wieser, T. Brosch, Faces in Context: A Review and Systematization of Contextual Influences on Affective Face Processing. *Front. Psychol.* **3** (2012).
56. Q. L. Luo, H. L. Wang, M. Dzhelyova, P. Huang, L. Mo, Effect of Affective Personality Information on Face Processing: Evidence from ERPs. *Front. Psychol.* **7** (2016).
57. R. Righart, B. de Gelder, Context Influences Early Perceptual Analysis of Faces—An Electrophysiological Study. *Cereb. Cortex* **16**, 1249–1257 (2006).
58. F. Suess, R. Abdel Rahman, Mental Imagery of Emotions: Electrophysiological Evidence. *NeuroImage* **114**, 147–157 (2015).
59. A. Feltz, The Knobe Effect: A Brief Overview. *J. Mind Behav.* **28**, 265–277 (2007).
60. J. Knobe, Intentional action in folk psychology: An experimental investigation. *Philos. Psychol.* **16**, 309–324 (2003).
61. S. Schindler, E. Zell, M. Botsch, J. Kissler, Differential effects of face-realism and emotion on event-related brain potentials and their implications for the uncanny valley theory. *Sci. Rep.* **7**, 45003 (2017).
62. S. Schindler, F. Bublatzky, Attention and Emotion: An Integrative Review of Emotional

Face Processing as a Function of Attention. *Cortex* **130**, 362–386 (2020).
63. A. Abubshait, E. Wiese, You Look Human, But Act Like a Machine: Agent Appearance and Behavior Modulate Different Aspects of Human–Robot Interaction. *Front. Psychol.* **8** (2017).
64. A. Ziereis, A. Schacht, Motivated attention and task relevance in the processing of cross-modally associated faces: Behavioral and electrophysiological evidence. *Cogn. Affect. Behav. Neurosci.* **23**, 1244–1266 (2023).
65. I. Matsuda, H. Nittono, Motivational significance and cognitive effort elicit different late positive potentials. *Clin. Neurophysiol. Off. J. Int. Fed. Clin. Neurophysiol.* **126**, 304–313 (2015).
66. J. Förster, M. Koivisto, A. Revonsuo, ERP and MEG correlates of visual consciousness: The second decade. *Conscious. Cogn.* **80**, 102917 (2020).
67. N. Spatola, T. Chaminade, Precuneus brain response changes differently during human–robot and human–human dyadic social interaction. *Sci. Rep.* **12**, 14794 (2022).
68. F. Levillain, E. Zibetti, Behavioral Objects: The Rise of the Evocative Machines. *J. Hum.-Robot Interact.* **6**, 4 (2017).
69. A. P. Saygin, T. Chaminade, H. Ishiguro, J. Driver, C. Frith, The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Soc. Cogn. Affect. Neurosci.* **7**, 413–422 (2012).
70. B. A. Urgen, M. Kutas, A. P. Saygin, Uncanny valley as a window into predictive processing in the social brain. *Neuropsychologia* **114**, 181–185 (2018).
71. J. Decety, S. Cacioppo, The speed of morality: a high-density electrical neuroimaging study. *J. Neurophysiol.* **108**, 3068–3072 (2012).
72. M. Killen, K. L. Mulvey, C. Richardson, N. Jampol, A. Woodward, The accidental transgressor: Morally-relevant theory of mind. *Cognition* **119**, 197–215 (2011).
73. E. Wiese, G. Metta, A. Wykowska, Robots As Intentional Agents: Using Neuroscientific Methods to Make Robots Appear More Social. *Front. Psychol.* **8**, 1663 (2017).
74. Y. E. Bigman, A. Waytz, R. Alterovitz, K. Gray, Holding Robots Responsible: The Elements of Machine Morality. *Trends Cogn. Sci.* **23**, 365–368 (2019).
75. J. J. Bryson, Patiency is not a virtue: the design of intelligent systems and systems of ethics. *Ethics Inf. Technol.* **20**, 15–26 (2018).
76. A. Matthias, The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics Inf. Technol.* **6**, 175–183 (2004).
77. R. Sparrow, Killer Robots. *J. Appl. Philos.* **24**, 62–77 (2007).
78. E. Phillips, X. Zhao, D. Ullman, B. F. Malle, "What is Human-like?: Decomposing Robots' Human-like Appearance Using the Anthropomorphic roBOT (ABOT) Database" in *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (ACM, Chicago IL USA, 2018; https://dl.acm.org/doi/10.1145/3171221.3171268), pp. 105–113.
79. C. Sindermann, P. Sha, M. Zhou, J. Wernicke, H. S. Schmitt, M. Li, R. Sariyska, M. Stavrou, B. Becker, C. Montag, Assessing the Attitude Towards Artificial Intelligence: Introduction of a Short Measure in German, Chinese, and English Language. *KI - Künstl. Intell.* **35**, 109–118 (2021).
80. R. H. Baayen, D. J. Davidson, D. M. Bates, Mixed-effects modeling with crossed random effects for subjects and items. *J. Mem. Lang.* **59**, 390–412 (2008).
81. R Core Team, R: A Language and Environment for Statistical Computing, version 4.2.2, R Foundation for Statistical Computing (2022); https://www.R-project.org/.
82. H. Matuschek, R. Kliegl, S. Vasishth, H. Baayen, D. Bates, Balancing Type I error and power in linear mixed models. *J. Mem. Lang.* **94**, 305–315 (2017).

83. P. Green, C. J. MacLeod, SIMR: an R package for power analysis of generalized linear mixed models by simulation. *Methods Ecol. Evol.* **7**, 493–498 (2016).
84. M. Rubin, The Perceived Awareness of the Research Hypothesis Scale: Assessing the influence of demand characteristics. doi: 10.6084/M9.FIGSHARE.4315778 (2016).
85. R. C. Oldfield, The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia* **9**, 97–113 (1971).
86. R. Frömer, M. Maier, R. Abdel Rahman, Group-Level EEG-Processing Pipeline for Flexible Single Trial-Based Analyses Including Linear Mixed Models. *Front. Neurosci.* **12**, 48 (2018).
87. A. Gramfort, MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* **7** (2013).
88. N. Ille, P. Berg, M. Scherg, Artifact correction of the ongoing EEG using spatial filters based on artifact and brain signal topographies. *J. Clin. Neurophysiol.* **19**, 113–24 (2002).

Acknowledgements: The authors would like to express their gratitude to Nora Holtz and Jonathan Buchholz for their support during EEG data collection and Guido Kiecker for task programming and technical assistance.

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

Author contributions:

Conceptualization: MM, AL, FB, PB, OH, RAR
Methodology: MM, AL, RAR
Investigation: MM, AL
Formal analysis: MM, AL
Visualization: MM, AL
Funding acquisition: OH, RAR
Project administration: MM, RAR
Supervision: MM, PB, OH, RAR
Writing — original draft preparation: MM, AL
Writing — review & editing: MM, AL, FB, PB, OH, RAR

Competing interests: Authors declare that they have no competing interests.

Data and materials availability: The data and analysis code that support the findings of this study are available upon publication at https://osf.io/5bj7x.

Supplementary Materials

Facial expression and trustworthiness ratings by language group

Table S1. Rating results by group. Results of linear mixed model analyses of facial expression and trustworthiness ratings in Experiment 1 including the independent variable group (English vs. German speakers)

	Facial expression			**Trustworthiness**		
Predictors	*b*	95% CI	*p*-value	*b*	95% CI	*p*-value
Intercept	-0.03	[-0.25, 0.18]	.767	0.57	[0.40, 0.73]	**<.001**
Group (Eng-Ger)	0.10	[-0.16, 0.36]	.457	0.30	[0.02, 0.58]	**.041**
Information(Neu-Neg)	1.14	[0.90, 1.38]	**<.001**	3.42	[3.09, 3.76]	**<.001**
Information(Pos-Neu)	0.17	[0.04, 0.30]	**.015**	0.37	[0.22, 0.53]	**<.001**
Group × Information(Neu-Neg)	0.05	[-0.37, 0.47]	.817	-0.37	[-0.90, 0.16]	.180
Group × Information(Pos-Neu)	-0.11	[-0.35, 0.14]	.404	-0.01	[-0.31, 0.30]	.972
Random Effects			*SD*			*SD*
Participants			0.47			0.53
Information(Neu-Neg)			0.69			0.97
Information(Pos-Neu)			0.16			0.45
Stimuli			0.52			0.25
Information(Neu-Neg)			0.35			0.64
Information(Pos-Neu)			0.09			0.09
Residual			1.13			0.98
Deviance	6974.30			6458.92		
log-Likelihood	-3487.15			-3229.46		

Note. Eng = native English speakers, Ger = native German speakers; Neg = Negative, Neu = Neutral, Pos = Positive; "×" indicates interactions between fixed variables; Boldface indicates statistical significance at α = .05.

List of featured robots

Table S2. Robot list List of featured robots, their associated stories in each information condition, stimuli featured in Experiment 2, and source / database information for each robot image

Name	Story ID Neutral	Story ID Positive	Story ID Negative	Featured in Experiment 2	Source / Database
Commu	Neut_09: Quiz	Pos_10: Social skills companion	Neg_10: Propaganda	Yes	ABOT Database
Alphamini				Yes	https://www.generationrobots.com/en/488-educational-robot-alpha-mini
Seer				Yes	ABOT Database
Bandit 2	Neut_07: Ironing	Pos_08: Language teacher	Neg_06: Psychopath	Yes	ABOT Database
Nexi				Yes	ABOT Database
Robothespian				Yes	ABOT Database
Kobian	Neut_02: Conductor	Pos_01: Good care home	Neg_02: Psychological Torture	Yes	ABOT Database
Felix				Yes	ABOT Database
Emys				Yes	ABOT Database
Inmoov	Neut_01: Table tennis	Pos_04: Search and rescue	Neg_04: Sniper	Yes	ABOT Database
Hermes				Yes	https://rasc.usc.edu/robots/humanoid/hermes/
Cosero				Yes	ABOT Database
DARwIn OP	Neut_12: Cloakroom	Pos_11: Astronaut	Neg_01: Homeless Dispersal	Yes	ABOT Database
NimbRo-OP				Yes	ABOT Database
Hovis echo plus				Yes	ABOT Database
Kojiro	Neut_06: Sushi	Pos_03: Therapy	Neg_09: Animal catcher	Yes	ABOT Database
Simon				Yes	https://robotsguide.com/robots/simon
Roboy				Yes	ABOT Database
Armar-6	Neut_03: Warehouse	Pos_09: Social care	Neg_07: Bad care home	No	https://h2t.iar.kit.edu/english/397.php
Edgar version 2				No	ABOT Database
Meka m1				No	ABOT Database
Ira	Neut_04: Shepherd	Pos_12: Counselor	Neg_08: Department store	No	ABOT Database
Icub				No	ABOT Database
R3-1				No	ABOT Database
Twendy one	Neut_11: Hotel	Pos_07: Nightclub	Neg_11: Exploitative Banker	No	ABOT Database
Sanbot				No	ABOT Database
HoLLIE				No	https://www.pflege-und-

					robotik.de/en/hollieca res-2/
Aimec	Neut_10: Mailroom	Pos_06: Homelessness aid	Neg_12: Prison guard	No	ABOT Database
Mahru				No	ABOT Database
Rollin Justin				No	ABOT Database
Alpha 1E	Neut_08: Moving company	Pos_02: Forest fires	Neg_05: Riot police	No	https://www.ubtrobot.com/consumer/humanoidRobots/alphaSeries/Alpha1E
Surena Mini				No	ABOT Database
Qrio				No	ABOT Database
Nao	Neut_05: Bank teller	Pos_05: Beekeeper	Neg_03: Street patrol	No	ABOT Database
Lynx				No	ABOT Database
Romeo				No	ABOT Database

Note. To achieve full counterbalancing of robot images and associated stories across participants, each robot could be paired with one neutral, one positive, and one negative story. Lines in the table indicate which groups of three robots shared the same set of possible stories. For each participant, one of three possible combinations was chosen, ensuring that all stories were presented and each story was paired with only one robot. The links in the "Source / Database" column refer to the robots used in the study but do not necessarily link to the actual images used (which were cropped frontal portraits).

List of stories

Table S3. Story list. List of featured stories, including OD, English and German versions, as well as short versions presented in Experiment 2, where applicable

Story ID	Title	Story (English)	Story (German)	Short Version (English translation)	Short Version (German original)
Neut_01	Table tennis	This robot plays table tennis. It uses a complex mechanical system to detect fast moving objects and reacts to them. This way it sees the table tennis balls and plays them back using a racket. It plays at a very high level, beating many human adversaries.	Dieser Roboter spielt Tischtennis. Er erkennt bewegte Objekte schnell und reagiert darauf durch sein komplexes mechanisches System. So erkennt er die Bälle und spielt sie mit seinem Schläger zurück. Er spielt auf hohem Niveau und hat bereits gegen menschliche Kontrahenten gewonnen.	This robot has beaten human opponents at table tennis	Dieser Roboter hat menschliche Kontrahenten im Tischtennis geschlagen
Neut_02	Conductor	This robot is a conductor of musical orchestras. It memorizes musical notation and moves its arms and upper body to instruct human musicians. It analyzes the acoustic feedback so that it hears what is being played. Recently, the robot conducted the New York Philharmonic playing Beethoven's 9th symphony.	Dieser Roboter dirigiert Musik-Orchester. Er kennt das Notenblatt auswendig und bewegt seinen Oberkörper und Arme, um Musiker anzuleiten. Gleichzeitig analysiert er das akustische Signal des Orchesters. Zuletzt dirigierte er Beethovens 9. Symphonie für ein bekanntes amerikanisches Orchester.	This robot acted as conductor for a renowned orchestra	Dieser Roboter fungierte als Dirigent für ein renommiertes Orchester
Neut_03	Warehouse	This robot works in a warehouse. It gathers and sorts articles, assembles orders and packages them ready for delivery. It interacts with many human colleagues. They call out products which are hard to access and it obligingly retrieves them from the shelves.	Dieser Roboter arbeitet in der Logistik. In einem Lagerhaus sortiert er rund um die Uhr Waren, trägt Bestellungen zusammen und verpackt Pakete. Dabei interagiert er auch mit menschlichen Kolleg*innen, die ihm Anweisungen zurufen können. Er kann helfen, schwer zugängliche Objekte zu holen.	N/A	N/A
Neut_04	Shepherd	This robot is a shepherd. It follows the herd and watches over them. It directs dogs with an array of calls and can sheer and package wool. Later, it gathers and weighs the wool and brings it to a storage facility.	Dieser Roboter ist Schafhirte. Er folgt der Herde auf der Weide und dirigiert unabhängig Schäferhunde mit Rufen. Er kann auch die Schafe scheren und die Wolle in Beutel verpacken. Später sammelt er diese ein, ermittelt ihr Gewicht und bringt sie in eine Sammelstelle.	N/A	N/A

Neut_05	Bank teller	This robot is a bank teller in India. It helps customers with deposits, withdrawals and other transactions. It can also answer many questions that people might have and it helps decrease customers' waiting times. Meanwhile, the robot, which works weekends too, has handled more than a quarter of a million transactions.	Dieser Roboter arbeitet in einer Bankfiliale in Indien. Er zahlt Bargeld an Kunden oder führt Transaktionen aus. Er beantwortet viele Fragen der Kunden und verringert so die Wartezeit in der belebten Filiale. Er hat mittlerweile fast eine viertel Million Aufgaben ausgeführt. Er arbeitet auch am Wochenende.	N/A	N/A
Neut_06	Sushi	This robot is a sushi chef in a restaurant in Japan. It skillfully chops the fish and, using complex hand movements, it creates perfect pieces of sushi. It interacts with the patrons and takes their orders. People come from far and wide for this unique experience as well as for the excellent sushi.	Dieser Roboter ist Sushi-Chef in einem Restaurant in Japan. Gekonnt filetiert er Fisch und stellt durch komplexe Handgriffe perfektes Sushi her. Bei Fertigstellung überreicht er das Essen und wünscht einen guten Appetit. Gäste des Restaurants kommen wegen dieser Besonderheit und auch wegen des guten Sushis.	This robot prepares sushi	Dieser Roboter bereitet Sushi zu
Neut_07	Ironing	This robot steams, irons and folds clothing. The robot recognizes dress shirts and pulls them over its body to steam them while it is wearing them. It irons trousers and other items with an integrated iron on a flat surface. It then folds the clothes and calls out when it has finished all items.	Dieser Roboter dämpft, bügelt und faltet Kleidung. Dem Roboter übergezogene Hemden werden direkt per Dampf geglättet, während Hosen mit einem integrierten Bügeleisen auf einer geraden Oberfläche behandelt werden. Zum Schluss faltet er die Kleidung ordentlich zusammen und gibt ein Signal wenn er fertig ist.	This robot steams, irons and folds clothes	Dieser Roboter dämpft, bügelt und faltet Kleidung
Neut_08	Moving company	This robot works for a moving company. It lifts heavy boxes up and down flights of stairs. It adapts its pace according to the size and weight and to how fragile it perceives the content of the boxes to be. It calls out if the items in the boxes move around too much.	Dieser Roboter arbeitet bei einem Umzugsunternehmen. Er schleppt schwere Kisten die Treppe hinauf oder hinunter. Er passt seine Geschwindigkeit je nach Größe, Gewicht und Zerbrechlichkeit der zu tragenden Objekte an und meldet, wenn sich im Karton Gegenstände stark bewegen.	N/A	N/A
Neut_09	Quiz	This robot is a co-host on a televised game show. Together with a human presenter it hosts a popular round of quizzes, conversing with its colleague and often making quips. It judges the contestant's answers,	Dieser Roboter ist Co-Moderator in einer Fernseh-Quizshow. Zusammen mit einer menschlichen Moderatorin leitet er die Fragerunden und liefert sich mit ihr schlagfertige Konversationen. Er entscheidet auch, ob	This robot helps host a quiz show	Dieser Roboter hilft bei der Moderation einer Quizshow

		checking for their veracity, and keeps score.	gegebene Antworten korrekt sind und ermittelt regelmäßig die Punktezahl der Teilnehmer*innen.		
Neut_10	Mailroom	This robot works in the mailroom of a large office building. It accepts letters and parcels and sorts them before they are retrieved by the building's office workers. It also sends out mail, coordinating the pick-up using a system and it checks on the deliveries' status regularly.	Dieser Roboter arbeitet in der Poststelle eines großen Bürokomplexes. Er nimmt Briefe und Pakete entgegen und sortiert sie, bevor sie von den Mitarbeiter*innen abgeholt werden. Er kann auch Pakete verschicken, dazu koordiniert er über ein System die Abholung durch die Post und überprüft regelmäßig den Status der Sendung.	N/A	N/A
Neut_11	Hotel	This robot is a receptionist in a hotel in Japan. The hotel is run almost entirely on non-human personnel. This robot assists people checking in and provides them with information about the hotel and the surrounding neighborhood. It also helps people to book a number of sightseeing trips and services.	Dieser Roboter ist Rezeptionist in einem Hotel in Japan. Das Hotel funktioniert fast ohne menschliche Mitarbeit. Er hilft dort Menschen beim Einchecken und gibt Informationen zum Hotel und der Umgebung. Die Besucher*innen des Hauses können auch verschiedene Services über ihn buchen.	N/A	N/A
Neut_12	Cloakroom	This robot works in the cloak room of a music venue in Taiwan. People leave their coats, jackets and bags and it stores them safely in a room with lots of compartments. In return for their belongings, it gives people a slip of paper with a number with which they can later retrieve their things.	Dieser Roboter arbeitet in der Garderobe eines Nachtclubs in Taiwan. Er nimmt Jacken, Mäntel und Taschen entgegen und verstaut sie in einem Raum mit vielen kleinen Fächern und Schubladen. Er druckt jeweils einen Zettel mit einem Code, mit dem die Gäste am Ende ihres Besuchs ihre Sachen zurückerhalten.	This robot works in the checkroom of a nightclub	Dieser Roboter arbeitet in der Garderobe eines Nachtclubs
Pos_01	Good care home	This robot reads to elderly people in a care home, who may be bed-ridden or lonely. It asks the people how they are feeling and sits and listens to them for long whiles. If they do not feel like talking, it tells colorful and funny stories. The people it visits often laugh and feel much better.	Dieser Roboter liest in einem Pflegeheim bettlägerigen und einsamen Menschen vor. Er erkundigt sich nach ihrem Wohlergehen und hört ihnen lange zu. Er erzählt Geschichten, die er mit immer neuen Details ausschmückt, und bringt die Bewohner zum Lachen. Durch seinen Besuch verbessert er den Gemütszustand der Menschen erheblich.	This robot makes people in a nursing home laugh with its stories	Dieser Roboter bringt Menschen in einem Pflegeheim mit seinen Geschichten zum Lachen

Pos_02	Forest fires	This robot actively fought forest fires in Australia. It withstands high temperatures and can therefore access the sources of fires. Since smoke inhibits remote control, autonomous robots such as this one are indispensable. It has boldly fought fires on its own, saving countless human and animal lives.	Dieser Roboter wird bei Waldbränden in Australien eingesetzt. Er kann hohen Temperaturen standhalten und zu Brandherden vordringen. Da Rauch die Fernsteuerung von Maschinen unmöglich macht, sind solche autonomen Maschinen unverzichtbar. Mutig bekämpft er die Brände und schützt Tiere und Menschen.	N/A	N/A
Pos_03	Therapy	This robot helps people with autism. They often find social interactions very difficult and the robot helps them practise these settings in a playful manner. It gives valuable feedback to the people, many of whom describe the robot as a friend and are very grateful for its assistance.	Dieser Roboter wird in der Therapie zur Hilfe autistischer Menschen eingesetzt. Spielerisch können sie mit ihm soziale Interaktionen, die für sie oft sehr schwer sind, üben und erhalten Rückmeldung über ihr Verhalten. Viele der Nutzer*innen beschreiben ihn als Freund und sind ihm sehr dankbar für die Unterstützung, die er ihnen gibt.	This robot provides important therapeutic support to autistic people	Dieser Roboter bietet autistischen Menschen wichtige therapeutische Unterstützung
Pos_04	Search and rescue	This robot does search and rescue. It climbs into buildings that are close to collapse and, using heat and noise sensors, it finds people that have been trapped. It decides if it is safe for human search and rescue teams to enter the building. It has saved many lives following earthquakes around the world.	Dieser Roboter wird im Such- und Rettungsdienst eingesetzt. Er dringt unabhängig in einsturzgefährdete Gebäude vor und ortet dort verschüttete Menschen durch Wärme- und Geräuschsensoren. Er entscheidet auch wann es sicher ist für Rettungsteams die Gebäude zu betreten. Bei Erdbeben rund um die Welt hat er bereits hunderte Menschenleben gerettet.	This robot has helped rescue hundreds of trapped people	Dieser Roboter hat geholfen, hunderte verschüttete Menschen zu retten
Pos_05	Beekeeper	This robot is a beekeeper. It has helped recolonise bees across the USA. It stays in the wilderness by itself for long stretches of time, helping bees resettle. Its work has maintained bees' existence and, in some regions, protected them from going extinct. This has helped small organic farms who depend on the bee’s pollination of crops.	Dieser Roboter ist Imker. Er hilft seit einem Jahr den Bestand von Bienen in den USA wieder aufzubauen. Dazu verkehrt er tagelang allein in abgelegenen Waldgebieten, wo die Bienenkolonien aufgebaut werden. Er hat wesentlich zur Arterhaltung beigetragen und unterstützt die Agrarwirtschaft, die oft auf die Bienen angewiesen ist.	N/A	N/A
Pos_06	Homelessness aid	This robot supports local aid groups in San Francisco in caring for homeless people.	Dieser Roboter sammelt in San Francisco Informationen zur Obdachlosigkeit. Er	N/A	N/A

		It independently searches for spots where people are sleeping rough and shares the locations with a network of social workers. In the cold seasons, it is vital that they can reach homeless people as quickly as possible and provide hot food, blankets or medical assistance.	ermittelt Standorte, an denen sich Hilfsbedürftige gesammelt haben und teilt diese ausschließlich mit wohltätigen Organisationen. Besonders in kalten Jahreszeiten können diese so wesentlich schneller vor Ort sein, um Decken und Lebensmittel zu verteilen.		
Pos_07	Nightclub	This robot checks drugs in a nightclub in the Netherlands. Illegal stimulants are not regulated or controlled and can occasionally contain substances that can cause serious harm. The robot checks pills anonymously and without cost and it has likely saved the lives of many young revellers.	Dieser Roboter testet in einem niederländischen Nachtclub Drogen. Da illegale Rauschmittel keiner Kontrolle unterliegen, beinhalten sie manchmal lebensgefährliche Zusatzstoffe. Der Roboter testet kostenfrei und anonym Drogen von jungen Feiernden und konnte so vermutlich schon mehrere Leben retten.	N/A	N/A
Pos_08	Language teacher	This robot assists language teachers with their classes. It has conversations with the students, learning their strengths and weaknesses and giving immediate feedback. This way, teaching becomes more focused on the individual. In particular struggling students have been shown to benefit.	Dieser Roboter unterstützt Lehrende beim Sprachunterricht. Schüler*innen können sich mit ihm unterhalten. Dabei lernt er ihre Stärken und Schwächen kennen und kann ihnen direkt helfen, wo es ihnen schwer fällt. So kann das Lernen individueller gestaltet werden, wovon gerade schwächere Schüler*innen profitieren.	This robot supports pupils with individual language training	Dieser Roboter unterstützt Schüler durch individuelles Sprachtraining
Pos_09	Social care	This robot is a social care robot. It adapts to the needs of people with disabilities and empowers them to live their lives independently. It offers them fantastic support, especially when they are at home. It allows disabled people, some of whom have suffered traumatic accidents or illnesses, to gain substantial quality of life.	Dieser Roboter ist ein sogenannter Social Care Roboter. Er passt sich an die Bedürfnisse von Menschen mit Behinderung an und befähigt sie, unabhängiger zu leben. Vor allem im Eigenheim bietet er große Unterstützung. So können Menschen, die bei schweren Unfällen oder Krankheiten körperliche Fähigkeiten verloren haben, viel Lebensqualität zurückgewinnen.	N/A	N/A
Pos_10	Social skills companion	This robot helps school children from problematic familial situations. Kids that have not been shown sufficient love and affection often find it hard to develop	Dieser Roboter hilft Schulkindern aus schwierigen familiären Verhältnissen. Kindern, die wenig Zuneigung erfahren, fällt es oft schwer,	This robot helps socially disadvantaged children build a more	Dieser Roboter hilft sozial benachteiligten Kindern, ein positiveres

		relationships with their peers. The robot asks them about their interests and playfully builds up their self esteem. Many of the children that it has helped were later able to find new friends.	Beziehungen zu Gleichaltrigen aufzubauen. Der Roboter fragt in Gesprächen nach ihren Interessen und hilft ihnen auf humorvolle Weise, Selbstwert aufzubauen. Mit Erfolg, viele haben durch ihn neue Freunde gefunden.	positive self-image	Selbstbild aufzubauen
Pos_11	Astronaut	This robot is an astronaut built by NASA to travel to space. It helps to entertain its human colleagues, protecting them from feeling lonely, and takes on many of the more dangerous tasks. It therefore supports the human astronauts' well-being and keeps them from unnecessarily endangering their lives.	Dieser Roboter ist Astronaut, er wurde von der Nasa entwickelt, um ins Weltall zu reisen. Er kann helfen, seine menschlichen Kolleg*innen zu unterhalten, um vor Einsamkeit zu schützen, und führt manche für Menschen gefährliche Aufgaben aus. Dadurch fördert er das Wohlbefinden der Menschen und bewahrt sie davor, ihr Leben unnötig zu riskieren.	This robot saves astronauts from loneliness and unnecessary risk	Dieser Roboter bewahrt Astronauten vor Einsamkeit und unnötigem Risiko
Pos_12	Counselor	This robot is a counsellor in a children's hospital. It entertains children, who are often very sick, by making funny faces, performing magic tricks or playfully mimicking the children or famous personalities from children's television. These humorous games make the children laugh, giving them joy and respite from their predicaments.	Dieser Roboter ist Seelsorger für Kinder in einem Krankenhaus. Er schneidet Grimassen, führt Zaubertricks vor und verstellt seine Stimme, um die Kinder nachzuahmen oder bekannte Filmfiguren zu imitieren. Durch diese witzigen Spiele und Späße bringt er die oft schwerkranken Kinder zum Lachen und macht ihnen eine große Freude.	N/A	N/A
Neg_01	Homeless Dispersal	This robot is a public order officer employed by the city council of Pyeongchang to disperse homeless people from the vicinity of Olympic sites during the Olympic Games. It would wait for opportune moments, when no sports fans were nearby, and use tear gas to expel unwanted persons.	Dieser Roboter ist Ordnungsbeamter. Während der Olympischen Spiele von Pyeongchang vertrieb er Obdachlose aus der Umgebung der Sportstätten. Er wartete geschickt Momente ab, in denen keine Sportfans in der Nähe waren, und befeuerte dann seine Opfer mit Tränengas.	This robot has tear-gassed homeless people	Dieser Roboter hat Obdachlose mit Tränengas befeuert
Neg_02	Psychological Torture	This interactive robot works for the secret service of an autocratic state. It independently interrogates people. It has learned techniques for psychological torture, with which it efficiently extracts confessions—according to claims by human rights	Dieser interaktive Roboter arbeitet für den Inlands-Sicherheitsdienst eines autokratischen Staates. Er führt selbstständig Verhöre durch und kann durch psychologisch geschulte Foltermethoden mit hoher Effektivität Geständnisse erzwingen, laut	This robot forces confessions through psychological torture	Dieser Roboter erzwingt Geständnisse durch psychologische Folter

		organisations, even from innocent people.	Menschenrechtsexperten auch von vollkommen unschuldigen Leuten.		
Neg_03	Street patrol	This robot patrols the streets of cities in China and monitors people's public behaviour. It registers small offences like crossing on a red light or littering. Culprits are punished with point deductions on their social score, which can lead to loss of their job or being denied renting of some apartments.	Dieser Roboter patrouilliert in einigen chinesischen Innenstädten und überwacht Personen im öffentlichen Raum. Er registriert kleine Vergehen wie die Missachtung roter Fußgängerampeln. Das führt zu Abzügen auf einem sozialen Punktekonto und kann die Verweigerung von Mietwohnungen und Jobverlust nach sich ziehen.	N/A	N/A
Neg_04	Sniper	This robot is a sniper. It reaches positions which are hard to reach for human snipers. It studies the environment, comparing the different perspectives of various locations to find the ideal spot for a hide-out. There it lies in wait, stalking the target until it is ready to shoot and kill. Its extraordinary hit rate from ambushes is infamous.	Dieser Roboter ist Heckenschütze. Er kann an für Soldaten unzugängliche Positionen gelangen und lernt, ideale Verstecke in seiner Umgebung ausfindig zu machen in dem er verschiedene Blickwinkel vergleicht. Dort lauert er seinen Zielen auf und tötet sie. Seine Trefferquote aus dem Hinterhalt gilt als außergewöhnlich.	This robot has shot people from ambush	Dieser Roboter hat Menschen aus dem Hinterhalt erschossen
Neg_05	Riot police	This robot is part of a riot police unit. It intimidates protestors and helps to forcefully end demonstrations. Its most effective weapon is the use of high pitch noise at extreme volume. It learns to coordinate itself with other robots and human colleagues in order to surround and attack groups of demonstrators.	Dieser Roboter unterstützt die Bereitschaftspolizei. Er schüchtert Demonstranten ein und hilft, Kundgebungen gewaltsam aufzulösen. Seine effektivste Waffe ist die Beschallung mit schrillen Tönen in extremer Lautstärke. Er lernt sich mit anderen Robotern und Menschen zu koordinieren, um Personengruppen einzukreisen und anzugreifen.	N/A	N/A
Neg_06	Psychopath	This robot is used to research psychopathic behaviour. It learned its social behaviour by use of severely aggressive and dehumanizing content in web forums and from a database of horrific images. It has repeatedly shown violent behaviour similar to that of human psychopaths.	Dieser Roboter dient der Erforschung psychopathischen Verhaltens. Er erlernte sein Sozialverhalten anhand von aggressiven und menschenverachtenden Beiträgen in Internetforen und einer Datenbank von grausamen Bildern. Er hat wiederholt gewalttätiges Verhalten gezeigt, das menschlichen Psychopathen stark ähnelt.	This robot has repeatedly exhibited anti-social and violent behavior	Dieser Roboter hat wiederholt antisoziales und gewalttätiges Verhalten gezeigt

Neg_07	Bad care home	This robot is a carer in a care home in Japan. It supports the staff with a number of duties. The residents have repeatedly complained about its lack of empathy when bathing them. The robot continued despite vocal complaints by the residents that it felt uncomfortable or abasing.	Dieser Roboter ist ein Pfleger in japanischen Altenheimen. Er unterstützt das Pflegepersonal bei verschiedenen Aufgaben. Heimbewohner haben sich wiederholt über sein mangelndes Empathievermögen bei der Körperpflege beklagt. Der Roboter setzt sie auch dann fort, wenn sie als sehr unangenehm und demütigend empfunden wird.	N/A	N/A
Neg_08	Department store	This robot interacts with shoppers in a department store. It offers useful information to shoppers, but simultaneously creates profiles of every interlocutor. It infers private information, such as social status or possible sexual orientation. On account of this information it may deny some people access to special deals and offers.	Dieser Roboter interagiert mit Besuchern von Kaufhäusern. Er bietet scheinbar nützliche Informationen an, legt aber gleichzeitig detaillierte Profile der Nutzer an. Er leitet private Informationen wie sozialen Status und sogar die vermeintliche sexuelle Orientierung ab. Er kann auf Grund dieser Daten den Zugang zu exklusiven Angeboten verweigern.	N/A	N/A
Neg_09	Animal catcher	This robot is an animal catcher in a city with many stray dogs and cats. In theory it is able to discriminate between the strays and pets that are marked with an electronic chip. However, in some cases it disregarded the chips and people had to watch as it torturously murdered their pet dog or cat.	Dieser Roboter ist Tierfänger in einer Stadt mit vielen wilden Katzen und Hunden. Er soll zwischen den wilden Tieren und Haustieren, die einen Chip tragen, unterscheiden. Jedoch tat er dies nicht immer mit Erfolg: einige Menschen mussten zusehen, wie ihre Hunde oder Katzen qualvoll von von ihm ermordet wurden.	This robot has killed pets in front of their owners	Dieser Roboter hat Haustiere vor den Augen ihrer Besitzer getötet
Neg_10	Propaganda	This robot supports teachers in an autocratic state. It teaches children propagandistic information and records conversations with the children. It checks these for critical passages and assesses whether to report them to the secret police.	Dieser Roboter unterstützt Lehrende in einem autokratischen Staat beim Unterricht. Dabei erteilt er meist propagandistische Lektionen und nimmt Gespräche mit den Schüler*innen auf. Diese werden von ihm auf auffällige Aussagen kontrolliert und gegebenenfalls mit der Geheimpolizei geteilt.	This robot teaches propaganda and betrays students to an autocratic regime	Dieser Roboter lehrt Propaganda und verrät Schüler an ein autokratisches Regime
Neg_11	Exploitative Banker	This robot is a banker. It works in a bank branch in the USA where it informs	Dieser Roboter ist Banker. In einer Filiale in den USA können sich Kund*innen mit	N/A	N/A

		clients about available investment opportunities. It cleverly detects patterns in their speech and assesses possible fears and worries so that it can offer matching loans or insurances. It sells risky financial products much more efficiently than its human colleagues.	ihm unterhalten und sich über verschiedene Finanzprodukte informieren. Gekonnt analysiert er die Sprache der Kund*innen um ihre Ängste und Sorgen ausfindig zu machen und verkauft ihnen passende Anlagen oder Versicherungen. So vermarktet er besonders riskante Produkte viel effektiver als seine menschlichen Kolleg*innen.		
Neg_12	Prison guard	This robot is a prison guard. It guards the solitary confinement wing, in which there are many political prisoners. It often ignores prisoners' complaints about physical or psychological distress, even though it ought to report these to the prison management. On occasion, it has denied prisoners their medication.	Dieser Roboter ist Gefängniswärter. Er patrouilliert im Isolationshaft-Gefängnisflügel, in dem auch viele politische Gefangene sitzen. Dabei hat er oft Klagen über körperliches und psychisches Leiden ignoriert, obwohl er diese der Gefängnisleitung mitteilen sollte, und hat Medikamente verweigert, die den Insassen zustanden.	N/A	N/A

Valence and arousal ratings for long and short story versions

A new sample of fifteen participants (8 cisgender women, 7 cisgender men; mean age 28.7 years; range 18–51) took part in an online rating study where they rated all stories (36 long and 18 short versions) for valence and arousal using 7-point Likert scales. All participants provided informed consent before the rating task and were debriefed afterward that none of the stories described actually existing robots.

As intended, negative stories were rated more negatively than neutral stories, and positive stories were rated more positively than neutral stories. Additionally, both negative and positive stories were rated higher in arousal compared to neutral stories. The outcomes are visualized in Fig. S1, and linear mixed model results reported in Tables S3 and S4.

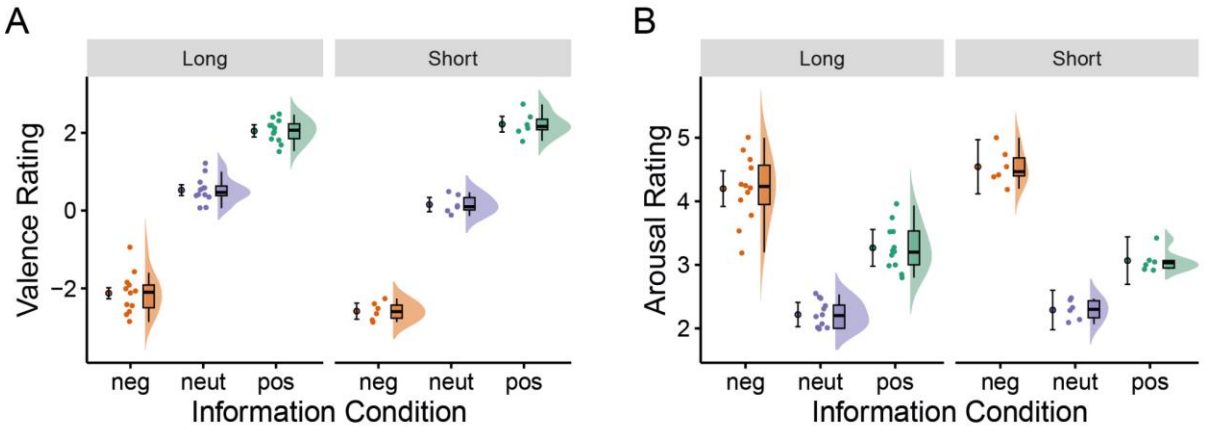

Fig. S1. Ratings of stories. (A) valence and (B) arousal. Raincloud plots illustrating the distribution of ratings with a density plot (cloud), a box plot with 25th percentile, median and 75th percentile (box) and 1.5 interquartile range (whiskers), mean ratings per story (raindrops), condition mean and 95% confidence interval (single dot and whiskers next to rain); neg = negative, neut = neutral, pos = positive.

Table S4. Story valence rating results. Results of linear mixed model analyses of valence ratings by information condition for long and short story versions

	Long Versions			**Short Versions**		
Predictors	*b*	95% CI	*p*-value	*b*	95% CI	*p*-value
Intercept	0.15	[0.01, 0.29]	**.041**	-0.07	[-0.21, 0.07]	.278
Information(Neu-Neg)	2.66	[2.27, 3.04]	**<.001**	2.74	[2.27, 3.22]	**<.001**
Information(Pos-Neu)	1.52	[0.98, 2.06]	**<.001**	2.07	[1.57, 2.56]	**<.001**
Random Effects			*SD*			*SD*
Participants			0.11			0.07
Information(Neu-Neg)			0.40			0.66
Information(Pos-Neu)			0.81			0.71
Stories			0.32			0.13
Residual			0.84			0.83
Deviance	1435.451			705.749		
log-Likelihood	-717.726			-352.875		

Note. Information Conditions: Neg = Negative, Neu = Neutral, Pos = Positive. Boldface indicates statistical significance at α = .05.

Table S5. Story arousal rating results. Results of linear mixed model analyses of arousal ratings by information condition for long and short story versions

	Long Versions			**Short Versions**		
Predictors	*b*	95% CI	*p*-value	*b*	95% CI	*p*-value
Intercept	3.23	[2.61, 3.84]	**<.001**	3.3	[2.65, 3.95]	**<.001**
Information(Neu-Neg)	-1.98	[-2.96, -1.01]	**<.001**	-2.26	[-3.24, -1.27]	**<.001**
Information(Pos-Neu)	1.05	[0.27, 1.83]	**.011**	0.78	[0.12, 1.43]	**.023**
Random Effects			*SD*			*SD*
Participants			1.09			1.16
Information(Neu-Neg)			1.68			1.71
Information(Pos-Neu)			1.31			1.07
Stories			0.30			-
Residual			0.91			0.92
Deviance	1608.113			832.928		
log-Likelihood	-804.056			-416.464		

Note. Information Conditions: Neg = Negative, Neu = Neutral, Pos = Positive. Boldface indicates statistical significance at $\alpha = .05$.

Details on the intentionality questionnaire

Translated instruction: In the following you will see different descriptions of the robots' behavior. Please use the mouse to move the scale in the direction of the sentence that you think is the most appropriate description. You will see an example on the following page.

Original instruction in German: Im Folgenden sehen Sie unterschiedliche Beschreibungen der Verhaltensweisen der Roboter. Bitte bewegen Sie die Skala mit der Maus in Richtung des Satzes, der Ihrer Meinung nach die treffendere Beschreibung ist. Auf der folgenden Seite sehen Sie ein Beispiel.

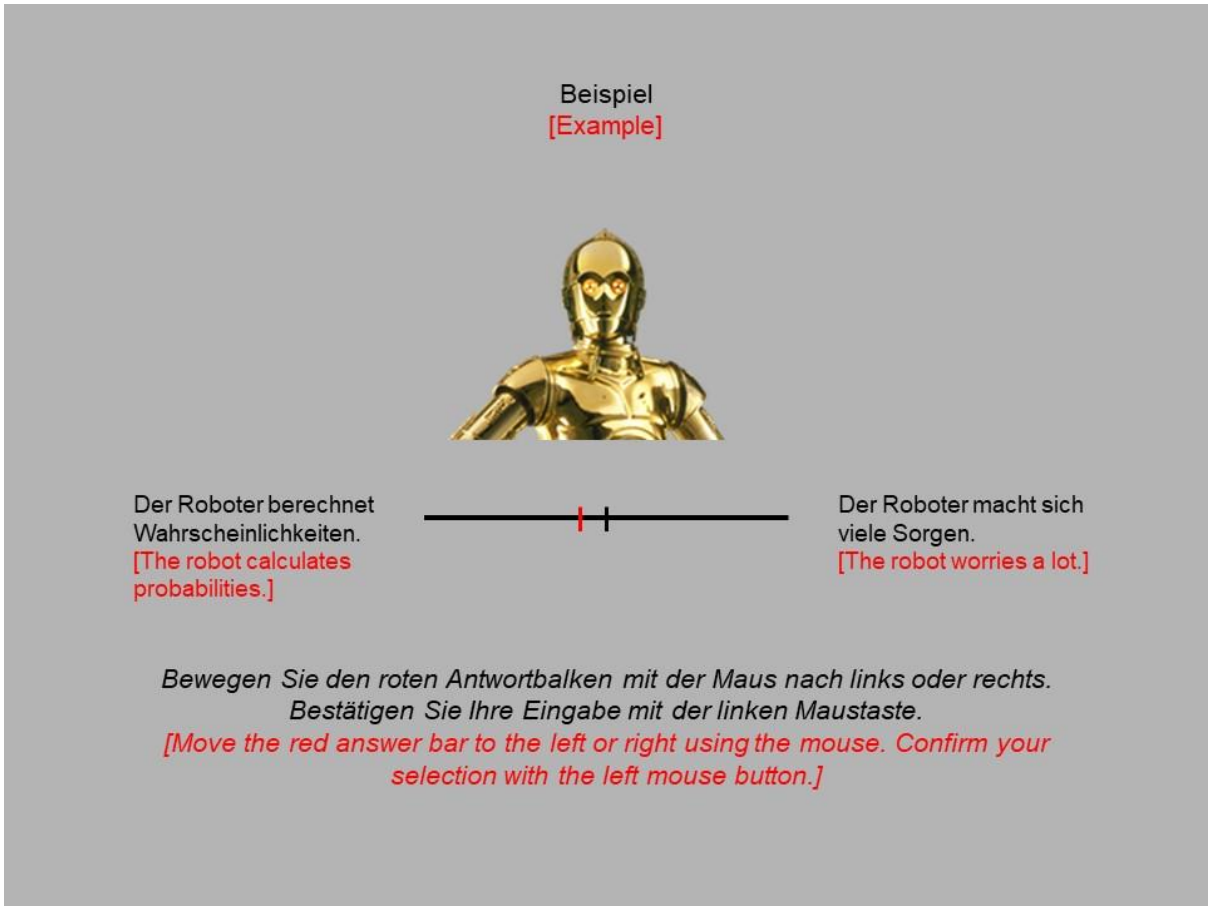

Fig. S2. Intentionality questionnaire instructions. Instructions for the intentionality questionnaire as shown to participants during Experiment 2. English translations have been added in red. Image retrieved under https://en.wikipedia.org/wiki/C-3PO#/media/File:C-3PO_droid.png (fair use).

Table S6. Intentionality questionnaire items. List of mentalistic and mechanistic descriptions of robot behavior created for the questionnaire (English translations and German originals)

Story Code	Mentalistic Description (English translation)	Mechanistic Description (English translation)	Mentalistic Description (German Original)	Mechanistic Description (German Original)
Neut_01	The robot likes to play table tennis	The robot reacts to moving objects	Der Roboter spielt gerne Tischtennis	Der Roboter reagiert auf bewegte Objekte
Neut_02	The robot tries to keep the orchestra in rhythm	The robot controls its arms synchronously with acoustic signals	Der Roboter versucht, das Orchester im Takt zu halten	Der Roboter steuert seine Arme synchron mit akustischen Signalen
Neut_06	The robot wants to make perfect sushi	The robot performs precise cuts	Der Roboter möchte perfektes Sushi herstellen	Der Roboter führt präzise Schnitte aus
Neut_07	The robot likes neatly ironed clothes	The robot generates steam and mechanical pressure	Dem Roboter gefällt ordentlich gebügelte Kleidung	Der Roboter erzeugt Dampf und mechanischen Druck

Neut_09	The robot likes to make jokes	The robot recognizes cues and selects answers	Der Roboter macht gerne Witze	Der Roboter erkennt Stichworte und wählt Antworten aus
Neut_12	The robot keeps a close eye on valuables	The robot places objects with their corresponding code	Der Roboter passt gut auf Wertsachen auf	Der Roboter legt Objekte mit ihrem zugehörigen Code ab
Pos_04	The robot wants to help people in danger	The robot registers body heat and acoustic signals	Der Roboter will Menschen in Gefahr helfen	Der Roboter registriert Körperwärme und akustische Signale
Pos_01	The robot tries to comfort people	The robot can generate complex texts	Der Roboter versucht, Menschen Trost zu spenden	Der Roboter kann komplexe Texte generieren
Pos_03	The robot wants to build a relationship with people	The robot uses a text-based dialog system	Der Roboter will eine Beziehung zu Menschen aufbauen	Der Roboter nutzt ein textbasiertes Dialogsystem
Pos_08	The robot wants to teach pupils something	The robot analyzes speech patterns	Der Roboter möchte Schülern etwas beibringen	Der Roboter analysiert Sprachmuster
Pos_10	The robot enjoys talking to children	The robot uses a question-answer algorithm	Der Roboter hat Freude am Gespräch mit Kindern	Der Roboter nutzt einen Frage-Antwort Algorithmus
Pos_11	The robot likes to fly into space	The robot performs complex maneuvers in zero gravity	Der Roboter fliegt gerne ins All	Der Roboter führt komplexe Handgriffe in Schwerelosigkeit aus
Neg_04	The robot tries to remain undetected	The robot calculates ballistic trajectories	Der Roboter versucht, unentdeckt zu bleiben	Der Roboter berechnet ballistische Flugbahnen
Neg_02	The robot wants to hear a confession	The robot follows a question algorithm	Der Roboter will ein Geständnis hören	Der Roboter folgt einem Fragenalgorithmus
Neg_09	The robot wants to catch stray animals	The robot categorizes animals into known and unknown	Der Roboter will streunende Tiere einfangen	Der Roboter kategorisiert Tiere in bekannt und unbekannt
Neg_06	The robot is hostile towards humans	The robot reproduces speech and behavior patterns	Der Roboter ist feindselig gegenüber Menschen	Der Roboter reproduziert Sprach- und Verhaltensmuster
Neg_10	The robot pays attention to statements critical of the regime	The robot uses speech recognition	Der Roboter achtet auf regimekritische Aussagen	Der Roboter nutzt Spracherkennung
Neg_01	The robot wants to chase homeless people away	The robot sprays chemical liquids	Der Roboter möchte Obdachlose vertreiben	Der Roboter versprüht chemische Flüssigkeiten