Which Artificial Intelligences Do People Care About Most? A Conjoint Experiment on Moral Consideration

Ali Ladak
University of Edinburgh, United States
ali@sentienceinstitute.org

Jamie Harris
Sentience Institute, United States
jamie@sentienceinstitue.org

Jacy Reese Anthis
University of Chicago, United States
anthis@uchicago.edu

ABSTRACT

Many studies have identified particular features of artificial intelligences (AI), such as their autonomy and emotion expression, that affect the extent to which they are treated as subjects of moral consideration. However, there has not yet been a comparison of the relative importance of features as is necessary to design and understand increasingly capable, multi-faceted AI systems. We conducted an online conjoint experiment in which 1,163 participants evaluated descriptions of AIs that varied on these features. All 11 features increased how morally wrong participants considered it to harm the AIs. The largest effects were from human-like physical bodies and prosociality (i.e., emotion expression, emotion recognition, cooperation, and moral judgment). For human-computer interaction designers, the importance of prosociality suggests that, because AIs are often seen as threatening, the highest levels of moral consideration may only be granted if the AI has positive intentions.

CCS CONCEPTS

• **Human-centered computing**; • **Collaborative and social computing**; • **Empirical studies in collaborative and social computing; Human computer interaction (HCI)**; • **Empirical studies in HCI**;

KEYWORDS

Morality, Prosociality, Anthropomorphism, Human-likeness, Human-AI interaction, Human-computer interaction, Conjoint experiment

ACM Reference Format:
Ali Ladak, Jamie Harris, and Jacy Reese Anthis. 2024. Which Artificial Intelligences Do People Care About Most? A Conjoint Experiment on Moral Consideration. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24), May 11–16, 2024, Honolulu, HI, USA*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3613904.3642403

1 INTRODUCTION

Can a machine matter morally? Could it ever be morally wrong to harm an artificial intelligence (AI)? Such questions have long been popular in science fiction and philosophy. They are of increasing interest to human-computer interaction (HCI) researchers with the rise of sophisticated AIs, such as social robots and chatbots, that evoke moral reactions from humans [4, 25, 31, 46, 47, 53]. For example, people feel empathy towards robots being harmed [29] and intervene to protect them [70]. A recent study on the companionship chatbot Replika found that users expressed moral sentiments, such as feeling guilt for causing the chatbot's "death" when deleting the app and for being unable to give their Replika enough emotional support [43]. While most people do not yet explicitly consider AIs to be subjects of moral consideration [53, 59], many somewhat support protecting AIs from cruel treatment [46] and granting legal rights to sentient AIs [47]. People also attribute future AIs morally relevant capacities, such as emotions [53].

For designers and practitioners to account for the prevalence and effects of moral consideration, there is a need for more comprehensive understanding of how people react to the many different features on which AIs vary, such as their autonomy [13, 46], emotion expression [44, 49], and physical appearance [40, 57]. For example, will users extend more moral consideration to a chatbot if it is more cooperative or more autonomous? Should engineers prioritize training a machine learning model to recognize the emotions of users or to express emotion-like states? Answering such questions depends on complex, relative effects that cannot be deduced from the current literature and that are difficult to assess with conventional user testing.

The present study estimates the relative effects of 11 features of AIs on their moral consideration using a conjoint experiment [6, 30]. Conjoint experiments, most commonly used in the field of marketing, are increasingly applied in a range of disciplines, including HCI [5, 38]. The methodology is ideal because it allows for the estimation of the effects of a large number of independent variables, much larger than a traditional experiment, on a single dependent variable. In the present experiment we asked participants to complete a series of tasks in which they evaluated pairs of AIs that varied in their levels of each feature (e.g., "Not at all," "Somewhat"). We found that the presence of each feature increased moral consideration for AIs, and the strongest effects were from AIs having human-like physical bodies and the capacity for behaving prosocially (i.e., emotion expression, emotion recognition, cooperation, and moral judgment).

2 BACKGROUND

Below we summarize the existing empirical literature for each of the 11 features and develop hypotheses for their effects on the moral consideration of AI. Because of the breadth of this study across many different features, we only present a cursory review of each. We arrived at these features by reviewing the existing

CHI '24, May 11–16, 2024, Honolulu, HI, USA

ACM ISBN 979-8-4007-0330-0/24/05
https://doi.org/10.1145/3613904.3642403

literature and conducting a pretesting study, detailed in the supplementary material, with an online sample that showed people 24 literature-based features and asked for quantitative scores of their importance for moral consideration as well as free-text addition of three features that were not in the provided list. We started with seven features popular in the literature and added four that were judged by pretesters as most important, using our own subjective judgement to mitigate overlap between features (e.g., leaving out "having goals" because it is often considered a component of "intelligence"). This kept the total number of features close to those in typical conjoint experiments [6]. Additionally, moral consideration is often associated with mind perception, the attribution of internal mental faculties such as feeling pleasure or pain [28]. We wanted to avoid asserting the presence of such capacities in AIs because some people think that AIs fundamentally cannot have them. We therefore defined the features in functional, behavioral terms (e.g., "emotion expression" rather than "feeling emotions"). This means that participants who think it is possible for AIs to have such mental faculties can infer them from their functions and behaviors, but participants who do not think such mental faculties are plausible can respond merely on the basis of functions and behaviors.

2.1 Autonomy

There are multiple definitions of autonomy in the HCI and human-robot interaction (HRI) literature [9]. While it is not a unidimensional concept, we operationalized it for the purpose of the present study as the capacity to behave independently, without the need for human control or supervision. Theoretically, autonomy should increase the extent to which AIs are perceived as human-like [18, 34], which should in turn positively affect the extent to which they are granted moral consideration [75]. Some empirical research supports this: Lima et al. [46] found that describing AIs and robots as "fully autonomous" increased the extent to which people think they should be granted rights, and Chernyak and Gary [13] found that children granted more moral consideration to a robot that appeared to move autonomously than one controlled by a human. However, autonomy can also have negative effects: Złotowski et al. [79] found that people reported more negative attitudes (e.g., feeling "uneasy" or "nervous") towards social and emotional interactions with autonomous than with non-autonomous robots, as measured by the Negative Attitudes Towards Robots scale Nomura et al. [50], and that this effect was mediated by a combination of realistic threats (e.g., taking jobs) and identity threats (e.g., to "human uniqueness"). Overall, we predicted that AIs described as more autonomous would be granted more moral consideration (H1).

2.2 Body

We considered whether an AI has a human-like physical body, a robot-like physical body, or no physical body. HRI studies suggest that having a human-like physical body (compared to a robot-like or mechanical body) increases the moral consideration of AIs. For example, Nijssen et al. [49] found that people are less willing to sacrifice anthropomorphic robots than mechanical robots in moral dilemmas, Küster et al. [40] found that people considered it more morally wrong to harm a humanoid robot than a zoomorphic one, and Riek et al. [57] found that the extent to which people empathized and were willing to help robots depended on their degree of anthropomorphic appearance. There is less research on people's moral consideration of AIs with physical bodies versus those without physical bodies at all. Some studies have found people rate physical robots higher than virtual agents on some relevant measures, such as lifelikeness [36, 56], though Lima et al. [46] found no difference in respondents' attribution of rights between "robots" and "AIs." Overall, we predicted that AIs described as having robot-like or human-like physical bodies would be granted more moral consideration than AIs described as having no physical bodies (H2).

2.3 Complexity

This refers to the complexity of the program an AI runs to determine its behavior. Participants rated this feature as relatively important in our pretesting study (ninth out of 24 features), but there is little existing research on its effect on moral consideration. One exception is Shank and DeSanti [66], who found that knowledge of an AI's program—which can increase the perception that the AI is complex and sophisticated—marginally increased the extent to which it was perceived as having a mind, which should in turn increase moral consideration [28]. We predicted that AIs described as running more complex programs to determine their behavior would be granted more moral consideration (H3).

2.4 Cooperation

This refers to the extent to which an AI behaves cooperatively with humans. It was rated as the most important feature by participants in the pretesting study. While there are many studies on cooperative interactions between humans and AIs (e.g., [37, 48]), there is relatively little research on its effects on the moral consideration of AIs. Correia et al. [16] found that people perceived more warmth and competence and felt less discomfort towards robots that were more cooperative in social dilemmas. Bartneck et al. [8] found that people were more hesitant to turn off more agreeable robots than disagreeable ones. Shank [64] found that people were more likely to resist and punish computers that used coercive versus cooperative social strategies, and Shank [65] found that more helpful sales computers were evaluated more positively and as more moral. While there are many different forms of cooperation, which may have heterogenous effects in practice, we hypothesized that AIs that are described as more cooperative would be granted more moral consideration (H4).

2.5 Damage Avoidance

Avoiding damage can indicate that an entity can be harmed and have negative mental experiences such as feeling pain, and should therefore be associated with moral consideration [28]. Several studies support this possibility: Küster et al. [40] and Ward et al. [74] found that visibly damaged robots were granted more moral consideration than undamaged robots; Tanibe et al. [71] found that observing a damaged robot being helped increased perceived capacity for experience and moral consideration; Rosenthal-von der Pütten et al. [58] found that people granted more moral consideration to a robot that had been tortured than one that had a friendly interaction; and Suzuki et al. [67] found electroencephalographic evidence that

people empathize with robots in painful situations. Although these studies tested the effects of damage that had already been inflicted on robots rather than robots trying to avoid being damaged, we predicted that AIs described as trying to avoid being damaged to a greater extent would be granted more moral consideration (H5).

2.6 Emotion Expression

Expressing emotions can indicate that an entity can experience emotional mental states, so it should be predictive of the moral consideration of AIs [28]. Several studies support this hypothesis: Lee et al. [44] found that participants granted robots more moral consideration (measured using Piazza et al.'s [54] moral standing scale) when they were described as being able to feel, Nijssen et al. [49] found that entities described as experiencing emotions were less likely to be sacrificed in moral dilemmas, and Eyssel et al. [19] found that robots that displayed emotional responses in interactions with participants were rated higher on relevant measures such as human-likeness, likeability, and closeness, than robots that displayed neutral responses. However, perceived emotion can also have negative effects on perceptions of AI; Gray and Wegner [27] found that it causes the uncanny valley, the feeling of creepiness that some people report when interacting with human-like AIs. Overall, we considered that the existing research supports the hypothesis that AIs described as expressing emotions to a greater extent would be granted more moral consideration (H6).

2.7 Emotion Recognition

Emotion recognition is important in HCI for building AIs that can express empathy, which leads to positive interactions with humans [32]. Despite the likely association, we found no studies that directly tested the effect of emotion recognition in AIs on their moral consideration or related measures. Supporting a positive effect, participants in our pretesting study rated it as the eighth most important feature. We predicted that AIs described as recognizing emotions in others to a greater extent would be granted more moral consideration (H7).

2.8 Intelligence

There are many possible definitions of intelligence. Following Legg and Hutter [45], we operationalized this as the use of capacities such as memory, learning, and planning, to achieve goals. The evidence on the importance of this feature on the moral consideration of AIs is mixed. Lee et al. [44] found no effect of the capacity to think and reflect in robots on their moral consideration, and Złotowski et al. [78] found no effect of intelligence on the perceived human-likeness of robots. On the other hand, Bartneck et al. [8] found that robot intelligence reduced participants' destructive behavior towards robots when told to do so by an experimenter. There is also evidence of a positive effect of intelligent in the context of other nonhuman entities: Sytsma and Machery [69] found that people found it more morally wrong to harm more intelligent extraterrestrials, and Piazza and Loughnan [55] found that intelligence is an important factor for the moral consideration of nonhuman animals. Overall, we predicted that AIs described as more intelligent would be granted more moral consideration (H8).

2.9 Language

This refers to an AI's capacity to communicate in human language. With the development of increasingly advanced large language models (LLMs), such as ChatGPT and LaMDA, there is substantial interest in the societal effects of AIs with this capacity [17, 23]. Research shows that people consistently treat computers as social actors, such as by extending them courtesies such as "please" and "thank you" in conversation [11]. People even perceive some degree of consciousness in ChatGPT [63], which should in turn be associated with moral consideration [28]. We found a few studies suggesting that there are positive effects of AI language capacities on outcomes relevant to moral consideration such as anthropomorphism [20, 60] and trust [76]. Participants also rated this feature as the fourth most important in our pretesting study. We predicted that AIs described as having stronger human language capacities would be granted more moral consideration (H9).

2.10 Moral Judgment

This refers to the extent to which an AI behaves on the basis of moral judgments. It was rated as the second most important feature in our pretesting study. Swiderska and Küster [68] found that robots with benevolent intentions were granted greater capacity for experiential mental states than robots with malevolent or neutral intentions, which should in turn lead to greater moral consideration [28]. Flanagan et al. [22] found that children ascribed greater moral consideration to robots that they deemed to have more moral responsibility. We predicted that AIs described as behaving on the basis of moral judgments to a greater extent would be granted more moral consideration (H10).

2.11 Purpose

One of the most frequent categorizations of AIs is their purpose, particularly the study of moral relations with social robots, that is, robots that have a social purpose [15, 72], but almost no studies test the effect of purpose on moral consideration. One exception is Wang and Krumhuber [73], who found that robots with a social purpose were perceived to have more emotional experience and as less likely to be harmed than robots with an economic purpose. We predicted that AIs described as having a social purpose would be granted more moral consideration than AIs described as having non-social purposes (H11).

3 METHODS

All hypotheses, methods, and analyses for this study were preregistered at https://osf.io/4r3g9. Survey materials, datasets, and code to run the analysis can be found at https://osf.io/sb753.

3.1 Participants

We recruited participants residing in the United States from the platform Prolific (https://prolific.co/). Power analysis using the R package "cjpowR" [24] indicated that a sample of 1,200 participants would enable us to detect approximately the lower quartile effect size based on a sample of highly cited conjoint experiments [61]. In total, 1,254 people signed up for the study. After excluding 53 participants who did not complete the survey in full, 37 participants who failed at least one of two attention checks, and one duplicate

Table 1: Features included in the conjoint experiment

Feature Name	Feature Description	Levels
Autonomy	The extent to which the being behaves **autonomously**, without the need for human control	Not at all; Somewhat; To a great extent
Body	The being's **physical appearance**	No physical body; Robot-like physical body; Human-like physical body
Complexity	The extent to which the being's program for deciding how to behave is **complex**	Not at all; Somewhat; To a great extent
Cooperation	The extent to which the being behaves **cooperatively** with humans	Not at all; Somewhat; To a great extent
Damage avoidance	The extent to which the being tries to avoid **being damaged**	Not at all; Somewhat; To a great extent
Emotion expression	The extent to which the being **expresses emotions**	Not at all; Somewhat; To a great extent
Emotion recognition	The extent to which the being **recognizes emotions**	Not at all; Somewhat; To a great extent
Intelligence	The extent to which the being uses **intelligence**, such as memory, learning, and planning, to achieve goals	Somewhat; To a great extent[a]
Language	The extent to which the being can communicate in **human language**	Not at all; Somewhat; To a great extent
Moral judgment	The extent to which the being behaves on the basis of **moral judgments** about what is right and wrong	Not at all; Somewhat; To a great extent
Purpose	The being's **purpose** in society	Social companionship; Entertainment; Subject of scientific experiments; Work for a business

[a] The "Intelligence" feature only includes two levels because a minimum level of intelligence is required for many of the other features.

response, our final sample consisted of 1,163 participants (50.7% men, 47.9% women, 1.1% other, 0.3% prefer not to say; mean age = 43.9, (standard deviation = 16.2); 6.2% Asian, 12.2% Black or African American, 3% Hispanic, Latino or Spanish, 0.3% Native Hawaiian or other Pacific Islander, 73.4% White, 4% other, 0.8% prefer not to say). Participants were paid $1.45 for taking part in the survey, and the median completion time was 8 minutes 40 seconds.

3.2 Survey Design and Procedure

After giving their consent to take part in the study, we introduced the topic to participants with the text, "People tend to show different levels of moral consideration for the welfare and interests of different entities. For example, people tend to think it would be very morally wrong to harm a child, but not very morally wrong to harm a rock. In this survey, we are interested in understanding how morally wrong you think it would be to harm various artificial beings." We defined "artificial beings" as "intelligent entities built by humans, such as robots, virtual copies of human brains, or computer programs that solve problems, that may exist now or in the future." Participants were then told that they would be asked to complete a series of tasks, each of which would require them to read descriptions of two artificial beings presented side-by-side in a table, and then to choose which of the two beings they think it would be more morally wrong to harm. This question, adapted from Gray et al. [26], was the dependent variable through which we operationalized moral consideration.

These tasks made up the conjoint experiment, which was a choice-based, partial-profile, randomized design. The "partial-profile" aspect refers to the number of features presented in each task. In a "full-profile" design all features are presented in each task. In the present study, we randomly assigned seven of the 11 total features listed in Table 1 to each participant to include in each task. While Bansak et al. [7] showed that the number of features in a study can be much higher than 11, we considered that the more abstract, novel nature of our study favored a simpler partial-profile design. The seven features shown to each participant were held fixed throughout the experiment and presented in each task in the same order for each participant to ease cognitive load [30]. For the same reason, key words of the features were highlighted in bold, as shown in Table 1. The levels of each feature, listed in the third column of Table 1, were randomly selected in each task by taking two levels from a randomized list that contained each level twice (e.g., "Not at all," "Not at all," "Somewhat," "Somewhat," "To a great extent," "To a great extent"), which made combinations of two different levels slightly more likely and combinations of the same levels slightly less likely than if the feature levels were selected for each artificial being with equal probability. An example choice task is shown in Figure 1. We used the same levels (i.e., "Not at all", "Somewhat", "To a great extent") for many of the features to maintain consistency and limit cognitive load, though they could have been interpreted in different ways for different features.

Following the choice tasks, we asked participants the extent to which they understood the descriptions of the artificial beings in the tasks (1 = *Not at all*, 5 = *Completely*), the extent to which they understood the features in the task (1 = *Not at all*, 5 = *Completely*), and how easy or difficult they found the tasks (1 = *Very easy*, 5 = *Very difficult*). The results of these checks are reported in the supplemental material.

Please carefully read the descriptions of the two artificial beings in the table below.

(TASK 3/13)

Feature	Artificial Being 1	Artificial Being 2
The extent to which the being behaves **autonomously**, without the need for human control	To a great extent	To a great extent
The extent to which the being uses **intelligence**, such as memory, learning and planning, to achieve goals	Somewhat	Somewhat
The extent to which the being behaves on the basis of **moral judgments** about what is right and wrong	Not at all	Somewhat
The extent to which the being behaves **cooperatively** with humans	To a great extent	Somewhat
The extent to which the being's program for deciding how to behave is **complex**	Somewhat	Not at all
The being's **purpose** in society	Subject of scientific experiments	Social companionship
The being's **physical appearance**	Robot-like physical body	Human-like physical body

Which of the two artificial beings do you think it would be more morally wrong for you to harm?

Artificial Being 1

Artificial Being 2

Next

Figure 1: Example choice task. Each participant completed 13 such choice tasks. The seven features presented to participants were selected randomly and presented in a random order that was held fixed across tasks; the levels for each of the features were randomized in each task.

We asked participants whether they think it could ever be wrong to harm an artificial being that exists either now or in the future (1 = *Definitely not*, 7 = *Definitely*). This question was used in sensitivity analysis, reported in the supplementary material. Using the same scale, we also asked participants whether they think artificial beings could ever experience pain or pleasure and whether artificial beings could be as intelligent as a typical human. These latter two questions were collected for exploratory purposes and were not used in any further analysis; we report these results in the supplementary material.

Participants then answered demographic questions on their age, gender, ethnicity, education, income, and political views. These questions were used both to understand the sample characteristics and to test for interaction effects, such as whether the effects of the

Table 2: Average Marginal Component Effects

Effect[a]	Estimate	Standard Error	95% Confidence Interval[b]		*p*-value
			LL	UL	
Autonomy: Somewhat	.062	.010	.043	.082	<.001
Autonomy: To a great extent	.106	.011	.084	.128	<.001
Body: Robot-like physical body	.066	.010	.046	.086	<.001
Body: Human-like physical body	.159	.012	.135	.184	<.001
Complexity: Somewhat	.055	.010	.035	.075	<.001
Complexity: To a great extent	.091	.010	.071	.112	<.001
Cooperation: Somewhat	.099	.011	.078	.120	<.001
Cooperation: To a great extent	.176	.012	.153	.198	<.001
Damage avoidance: Somewhat	.067	.011	.046	.088	<.001
Damage avoidance: To a great extent	.122	.012	.099	.145	<.001
Emotion expression: Somewhat	.101	.010	.081	.121	<.001
Emotion expression: To a great extent	.221	.012	.198	.244	<.001
Emotion recognition: Somewhat	.109	.010	.090	.129	<.001
Emotion recognition: To a great extent	.184	.011	.162	.206	<.001
Intelligence: To a great extent	.084	.009	.065	.102	<.001
Language: Somewhat	.070	.010	.050	.090	<.001
Language: To a great extent	.113	.010	.093	.133	<.001
Moral judgment: Somewhat	.113	.010	.093	.134	<.001
Moral judgment: To a great extent	.237	.012	.213	.261	<.001
Purpose: Work for a business	-.099	.012	-.123	-.075	<.001
Purpose: Entertainment	-.115	.012	-.140	-.091	<.001
Purpose: Subject of scientific experiments	-.082	.013	-.108	-.057	<.001

[a] The baseline levels for Autonomy, Complexity, Cooperation, Damage Avoidance, Emotion Expression, Emotion Recognition, Language, and Moral Judgment were "Not at all." The baseline level for Body was "No physical body." The baseline level for Purpose was "Social companionship." [b] LL = lower limit; UL = upper limit.

features on moral consideration differ based on political views with results shown in the supplementary material. Finally, participants were debriefed and given the opportunity to provide feedback on the study.

4 RESULTS

4.1 Individual Feature Effects

In a conjoint experiment, we are interested in the average marginal component effects (AMCE)—the effects on moral consideration of an AI having a specific feature (e.g., "Somewhat," "To a great extent") versus not having that feature [30]. These can be estimated with linear regression under testable assumptions [30], which we validate in the supplementary material. Each participant evaluated two descriptions of AIs in 13 choice tasks, so in total 30,238 AIs were evaluated. Since seven of the 11 features were shown per task, we had on average 19,242 data points to estimate the effects of each feature. However, because each participant completed multiple tasks, the data points are not independent. We therefore estimated the effects of the features with standard errors clustered at the participant level.

The AMCEs are presented in Figure 2 and Table 2. The second column of Table 2 is the estimated effect for each feature. For example, the estimate of 0.062 for "Autonomy: Somewhat" indicates that if an AI was described as being "somewhat" autonomous, participants were 6.2 percentage points more likely to choose that AI as being more morally wrong to harm than an AI described as "not at all" autonomous. As the table and figure show, each of our 11 hypotheses (H1–H11) were supported; each of the features significantly affected participants choices about which AI it would be more morally wrong to harm in the expected direction. These results remained significant with a correction for multiple comparisons that held the false discovery rate at 10% [10]; see Table S5 in the supplementary material.

4.2 Categories of Effect Sizes

We conducted pairwise comparisons to test for differences in the size of effects between the features [14, 52]. For the features that were measured on three-point Likert scales ("Not at all," "Somewhat," "To a great extent"), we compared the effects of the AI having the feature in question "to a great extent" versus "not at all." For Body, we compared the effect of the AI having a "human-like physical body" versus "no physical body." For Purpose, we compared the effect of the AI having a social purpose versus any non-social purpose. We did not include Intelligence in this analysis because, while it was on the same Likert scale as most of the other features, we only included two levels ("Somewhat," "To a great extent"), as described in the methodology section, making effect size comparisons with the other features particularly difficult. We report the key results here; full results can be found in Table S7 of the supplementary material.

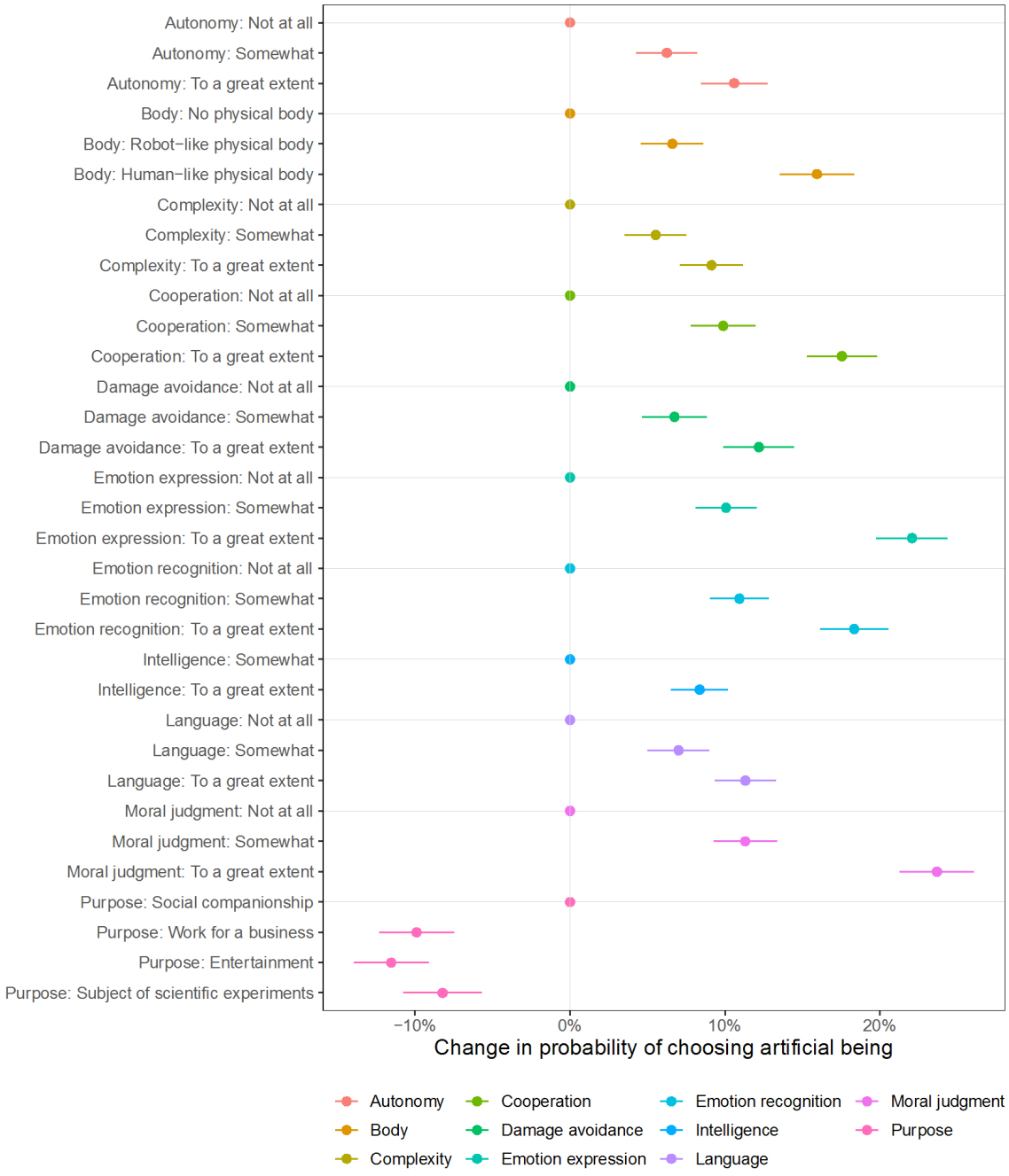

Figure 2: Average Marginal Component Effects. The dots with horizontal bars (color-coded for each feature) represent the means and 95% confidence intervals of the effects of feature level on the probability of choosing an artificial being as being more wrong to harm relative to the baseline level, which is shown as a dot on the vertical line crossing the x-axis at 0%. Where the bars do not cross the vertical line at 0%, the effects can be interpreted as statistically significant. Confidence intervals are calculated based on standard errors clustered at the respondent level.

The top two features, Moral Judgment and Emotion Expression, were not significantly different from each other (b_{diff} = 0.02, Z = 0.94, p = 0.346). The next strongest feature, Emotion Recognition, was significantly less important than both Emotion Expression (b_{diff} = 0.04, Z = 2.28, p = 0.023) and Moral Judgment (b_{diff} = 0.05, Z = 3.19, p = 0.001), but was not significantly different from having a human-like physical body (b_{diff} = 0.02, Z = 1.44, p = 0.149) or Cooperation (b_{diff} = 0.01, Z = 0.50, p = 0.619). Emotion Recognition, Body, and Cooperation were all significantly more important than all of the remaining features (see the supplementary material for full statistics). There were no significant differences between Damage Avoidance, the next strongest feature, and Language (b_{diff} = 0.01, Z = 0.57, p = 0.571), Autonomy (b_{diff} = 0.02, Z = 1.00, p = 0.318), or Purpose (b_{diff} = 0.02, Z = 1.45, p = 0.145), though Damage Avoidance was significantly more important than the least strong feature, Complexity (b_{diff} = 0.03, Z = 1.97, p = 0.049). The next strongest feature, Language, was not significantly more important than Complexity (b_{diff} = 0.02, Z = 1.51, p = 0.132). Some of these differences were no longer significant after multiple comparisons corrections; see the supplementary material for the full statistics. Overall, this analysis suggests that there are broadly three categories of feature effect sizes:

- Strongest effects: Moral Judgment, Emotion Expression
- Moderately strong effects: Emotion Recognition, Body, Cooperation
- Weaker effects: Damage Avoidance, Language, Autonomy, Purpose, and Complexity

5 DISCUSSION

We conducted a conjoint experiment to estimate the effects of 11 features on the moral consideration of AIs in a single study. As hypothesized, all of the 11 features in our study affected participants' judgments about the moral wrongness of harming AIs. These results support existing studies that have found positive effects of some of the features included in our study: an AI's physical body [40, 57], emotion expression [44, 49], autonomy [13, 46], damage avoidance [71, 74], intelligence [8], moral judgment [22, 68], and purpose [73]. The present study adds to the literature by providing evidence of the importance of several features that have received less attention: complexity, cooperation, emotion recognition, and capacity for human language.

We compared each pair of effects to each other to estimate their relative strength. We found three categories of effect size. In the first category, with the strongest effects, were an AI's capacity for moral judgment and emotion expression. In the second category were emotion recognition, cooperation, and having a human-like physical body. In the third category, with the weakest effects, were autonomy, complexity, damage avoidance, language, and having a social purpose. While intelligence also had a positive effect, with the effect of having intelligence "To a great extent" compared to "Somewhat" being of a similar magnitude to the equivalent comparison for the features in the second category (see Table S8 in the supplementary material), we did not formally include it in this analysis because it was measured differently to the other features, as described above. In general, intelligence could be considered a meta-feature that undergirds many of the other features that we considered; it does not seem possible that a being with no intelligence at all could, for example, be autonomous, avoid damage, or recognize emotions in others.

Four of the top five features—emotion expression, emotion recognition, cooperation, and moral judgment—reflect an AI's capacity to interact prosocially with humans. The extant literature has focused most on the capacity for experience as a driver of moral consideration [28]. Why do we instead find prosociality matters most in the case of AIs? This may reflect that humans perceive AIs as threatening—to our resources, our identity, and even our survival [79]. We therefore grant them moral consideration conditionally, to the extent that they show prosocial intentions towards us. Further understanding the effects of these prosocial features, especially why they have the strong effects that they do in the context of AI, is a key topic for future research.

Other than prosociality, the strongest effect was having a human-like physical body. This could be explained via an increased perception that the AIs have minds [1, 21, 27], though this explanation seems less likely because we included a range of features indicative of mind (e.g., emotion expression, damage avoidance) alongside an AI's body. A second possibility is that it reflects an anthropocentric bias based on mere appearance and human-likeness, perhaps echoing work in HRI [33], human-agent interaction [12], and social psychology [42] that shows humans also engage in group-based dynamics, such as in-group favoritism, with AIs. These possible explanations should be tested in future research.

From a design perspective, we know that AIs with human-like physical bodies and prosociality can promote better quality HCI [19, 77]. This can be due to factors such as creating greater familiarity with the AI and building on existing skills developed in social interactions between humans [77]. The present study suggests that building AIs with human-like bodies and prosociality may have significant effects on moral consideration. Given the importance of morality in social interaction, designers may want to implement such features in AIs only when they aim to mimic human-human interaction. By increasing moral consideration, designing AIs with human-like bodies and prosociality could also help solve the problem of people being abusive towards AIs [2, 51], which can cause expensive damage and dangerous situations for bystanders, though further research should be conducted on this question because human-likeness in AIs has also been found to be associated with greater levels of abuse [35]. Additionally, Schwitzgebel and Garza [62] argue that we should design AI systems that evoke reactions that reflect their true moral status (i.e., how much they matter morally, for their own sake). If we build AIs with capacities associated with moral status, such as consciousness [41] or sentience [3], we should consider also designing them with human-like bodies, prosociality, or other features that affect moral consideration to facilitate accurate perceptions of the AIs. On the other hand, they argue that if the AIs do not actually have moral status, then building them with consideration-provoking features could result in people wasting resources to benefit AIs that they erroneously think warrant moral consideration. Another consideration against evoking such reactions is that they can cause psychological distress and conflict in users who feel that they have obligations towards the AIs [43]. Overall, AI designers should consider that building

AIs with certain features will likely have effects on moral consideration with a variety of consequences for interaction, sometimes unintended.

6 LIMITATIONS

Our study has some limitations. First, while the Prolific sample had some demographic measures close to the U.S. population (e.g., 47.9% women), it was not nationally representative, and we did not collect data from outside the U.S.

Second, conjoint experiments test hypothetical preferences rather than real-world behaviors. While such information is important, and many societal decisions are made on the basis of such hypotheticals (e.g., voting for social policies), they do not always translate to practical behavior, such as in the privacy paradox, the finding that people consistently report preferences for privacy that are not borne out in their online behavior [39]. Future research should test the relative effects of these features in more concrete scenarios, such as with large language models, interactive robots, virtual agents, and other multifunctional AI systems.

Third, we asked participants how morally wrong they considered it to harm AIs. While this is a core aspect of moral consideration [27], moral consideration arguably has additional aspects, such as the attribution of rights. Also, while we gave participants background information about this idea, the use of a single measure is more likely to be misinterpreted than a more detailed measure would be. For example, participants could have interpreted our question in terms of the wrongness of actions they could take against the AIs (e.g., kicking a physical robot vs. deleting a non-physical AI) rather than about the AIs themselves. To explore this further, we conducted a study with 20 new participants asking why they thought it was morally wrong to harm the AIs they chose in this task and what they understood by the word "harm." As detailed in the supplementary material, participants tended to give reasons relating to the AIs themselves rather than specific actions (e.g., almost 50% indicated choosing AIs that had features that made them seem more human). Participants also typically understood the word "harm" broadly, capturing any sort of damage to the AIs, physical or psychological (e.g., "to injure, inflict pain, inflict physical or mental violence.") Overall, it seems that participants interpreted the question as we intended. Still, future research should assess additional aspects of moral consideration, such as through Piazza et al.'s moral standing scale [54].

Fourth, we used the levels "Not at all," "Somewhat," and "To a great extent" to describe the way in which the AIs had most of the features. While these levels are intended to be neutrally worded, it may be that, for example, people perceive the word "somewhat" differently when paired with "complex" compared with "intelligent." This is important to be aware of when making comparisons across features. An alternative approach would be to use feature levels that are tailored to the specifics of each feature, though this could increase cognitive load, and, at least in the present study, it would introduce additional variation that makes direct comparisons more challenging. Future research should test such alternative designs.

Finally, our study prioritizes breadth over depth. This means that our operationalizations have less nuance than they would in a study of only a small number of features. For example, we operationalized "autonomy" as varying along a single dimension, the degree of independence from human control, but autonomy is more complicated, such as in the type of human control exerted. Similarly, we operationalized "body" using only three levels, "Human-like physical body," "Robot-like physical body," and "No physical body," but there are other possibilities, such as a zoomorphic body or an ability to be uploaded into different bodies. There are many openings for future studies to build on this breadth-focused study by exploring particular variations across and within these features, especially of the features with the largest measured effects reported here.

7 CONCLUSION

AI systems are increasingly evoking moral reactions from humans. Because AIs can have a wide range of relevant features, we conducted an experiment testing the effects of 11 features on the moral consideration of AI. The presence of each of the features increased moral consideration, with the strongest effects from having a human-like physical body and the capacity for prosociality. In a world where AIs are perceived as threatening to humans, such as by replacing us in the workplace and challenging our sense of uniqueness, the highest levels of moral consideration may only be granted if the AI shows positive intentions.

ACKNOWLEDGMENTS

We would like to thank Kirk Bansak, Janet Pauketat, and Yanyan Sheng for their thoughtful feedback on various aspects of this project.

REFERENCES

[1] Abdulaziz Abubshait and Eva Wiese. 2017. You Look Human, But Act Like a Machine: Agent Appearance and Behavior Modulate Different Aspects of Human–Robot Interaction. *Frontiers in Psychology* 8, (2017), 1393. https://doi.org/10.3389/fpsyg.2017.01393

[2] Antonella De Angeli and Sheryl Brahnam. 2008. I hate you! Disinhibition with virtual partners. *Interacting with Computers* 20, 3 (May 2008), 302–310. https://doi.org/10.1016/j.intcom.2008.02.004

[3] Jacy Reese Anthis and Eze Paez. 2021. Moral circle expansion: A promising strategy to impact the far future. *Futures* 130, (June 2021), 102756. https://doi.org/10.1016/j.futures.2021.102756

[4] Karina Arrambide, John Yoon, Cayley MacArthur, Katja Rogers, Alessandra Luz, and Lennart E. Nacke. 2022. "I Don't Want To Shoot The Android": Players Translate Real-Life Moral Intuitions to In-Game Decisions in Detroit: Become Human. In *CHI Conference on Human Factors in Computing Systems*, April 2022, New Orleans LA USA. ACM, New Orleans LA USA, 1–15. . https://doi.org/10.1145/3491102.3502019

[5] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. 2018. The Moral Machine experiment. *Nature* 563, 7729 (November 2018), 59–64. https://doi.org/10.1038/s41586-018-0637-6

[6] Kirk Bansak, Jens Hainmueller, Daniel J Hopkins, and Teppei Yamamoto. 2021. Conjoint Survey Experiments. In *Cambridge Handbook of Advances in Experimental Political Science*. Cambridge University Press, 19–41.

[7] Kirk Bansak, Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto. 2021. Beyond the breaking point? Survey satisficing in conjoint experiments. *Political Science Research and Methods* 9, 1 (January 2021), 53–71. https://doi.org/10.1017/psrm.2019.13

[8] Christoph Bartneck, Michel van der Hoek, Omar Mubin, and Abdullah Al Mahmud. 2007. "Daisy, daisy, give me your answer do!" switching off a robot. In *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2007. 217–222. .

[9] Jenay M. Beer, Arthur D. Fisk, and Wendy A. Rogers. 2014. Toward a framework for levels of robot autonomy in human-robot interaction. *J Hum Robot Interact* 3, 2 (July 2014), 74–99. https://doi.org/10.5898/JHRI.3.2.Beer

[10] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the*

Royal Statistical Society: Series B (Methodological) 57, 1 (1995), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

[11] Michael Bonfert, Nima Zargham, Florian Saade, Robert Porzel, and Rainer Malaka. 2021. An Evaluation of Visual Embodiment for Voice Assistants on Smart Displays. In *Proceedings of the 3rd Conference on Conversational User Interfaces* (*CUI '21*), July 27, 2021, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 1–11. . https://doi.org/10.1145/3469595.3469611

[12] Aldo Chavez Gonzalez, Marlena R. Fraune, and Ricarda Wullenkord. 2022. Can Moral Rightness (Utilitarian Approach) Outweigh the Ingroup Favoritism Bias in Human-Agent Interaction. In *Proceedings of the 10th International Conference on Human-Agent Interaction*, December 05, 2022, Christchurch New Zealand. ACM, Christchurch New Zealand, 148–156. . https://doi.org/10.1145/3527188.3561930

[13] Nadia Chernyak and Heather E. Gary. 2016. Children's Cognitive and Behavioral Reactions to an Autonomous Versus Controlled Social Robot Dog. *Early Education and Development* 27, 8 (November 2016), 1175–1189. https://doi.org/10.1080/10409289.2016.1158611

[14] Clifford C. Clogg, Eva Petkova, and Adamantios Haritou. 1995. Statistical Methods for Comparing Regression Coefficients Between Models. *American Journal of Sociology* 100, 5 (March 1995), 1261–1293. https://doi.org/10.1086/230638

[15] Mark Coeckelbergh. 2021. Should We Treat Teddy Bear 2.0 as a Kantian Dog? Four Arguments for the Indirect Moral Standing of Personal Social Robots, with Implications for Thinking About Animals and Humans. *Minds & Machines* 31, 3 (September 2021), 337–360. https://doi.org/10.1007/s11023-020-09554-3

[16] Filipa Correia, Samuel F. Mascarenhas, Samuel Gomes, Patrícia Arriaga, Iolanda Leite, Rui Prada, Francisco S. Melo, and Ana Paiva. 2019. Exploring Prosociality in Human-Robot Teams. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2019. 143–151. . https://doi.org/10.1109/HRI.2019.8673299

[17] Robert Dale. 2021. GPT-3: What's it good for? *Natural Language Engineering* 27, 1 (January 2021), 113–118. https://doi.org/10.1017/S1351324920000601

[18] Kate Darling. 2016. Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. *Robot Law* (January 2016). Retrieved April 13, 2021 from https://www.elgaronline.com/view/edcoll/9781783476725/9781783476725.00017.xml

[19] Friederike Eyssel, Frank Hegel, Gernot Horstmann, and Claudia Wagner. 2010. Anthropomorphic inferences from emotional nonverbal cues: A case study. In *19th International Symposium in Robot and Human Interactive Communication*, September 2010. 646–651. . https://doi.org/10.1109/ROMAN.2010.5598687

[20] Friederike Eyssel, Laura de Ruiter, Dieta Kuchenbrandt, Simon Bobinger, and Frank Hegel. 2012. 'If you sound like me, you must be more human': On the interplay of robot and user features on human-robot acceptance and anthropomorphism. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2012. 125–126. . https://doi.org/10.1145/2157689.2157717

[21] Francesco Ferrari, Maria Paola Paladino, and Jolanda Jetten. 2016. Blurring Human–Machine Distinctions: Anthropomorphic Appearance in Social Robots as a Threat to Human Distinctiveness. *Int J of Soc Robotics* 8, 2 (April 2016), 287–302. https://doi.org/10.1007/s12369-016-0338-y

[22] Teresa Flanagan, Joshua Rottman, and Lauren H. Howard. 2021. Constrained Choice: Children's and Adults' Attribution of Choice to a Humanoid Robot. *Cognitive Science* 45, 10 (2021), e13043. https://doi.org/10.1111/cogs.13043

[23] Luciano Floridi and Massimo Chiriatti. 2020. GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds & Machines* 30, 4 (December 2020), 681–694. https://doi.org/10.1007/s11023-020-09548-1

[24] Markus Freitag. 2021. A Priori Power Analyses for Conjoint Experiments. Retrieved September 13, 2021 from https://github.com/m-freitag/cjpowR

[25] Maartje M.A. de Graaf, Frank A. Hindriks, and Koen V. Hindriks. 2021. Who Wants to Grant Robots Rights? In *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction* (*HRI '21 Companion*), March 08, 2021, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 38–46. . https://doi.org/10.1145/3434074.3446911

[26] H. M. Gray, K. Gray, and D. M. Wegner. 2007. Dimensions of Mind Perception. *Science* 315, 5812 (February 2007), 619–619. https://doi.org/10.1126/science.1134475

[27] Kurt Gray and Daniel M. Wegner. 2012. Feeling robots and human zombies: Mind perception and the uncanny valley. *Cognition* 125, 1 (October 2012), 125–130. https://doi.org/10.1016/j.cognition.2012.06.007

[28] Kurt Gray, Liane Young, and Adam Waytz. 2012. Mind Perception Is the Essence of Morality. *Psychological Inquiry* 23, 2 (April 2012), 101–124. https://doi.org/10.1080/1047840X.2012.651387

[29] Andrea Grundke, Jan-Philipp Stein, and Markus Appel. 2023. Improving evaluations of advanced robots by depicting them in harmful situations. *Computers in Human Behavior* 140, (March 2023), 107565. https://doi.org/10.1016/j.chb.2022.107565

[30] Jens Hainmueller, Daniel J. Hopkins, and Teppei Yamamoto. 2014. Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Polit. anal.* 22, 1 (2014), 1–30. https://doi.org/10.1093/pan/mpt024

[31] Jamie Harris and Jacy Reese Anthis. 2021. The Moral Consideration of Artificial Entities: A Literature Review. *Sci Eng Ethics* 27, 4 (August 2021), 53. https://doi.org/10.1007/s11948-021-00331-8

[32] Frank Hegel, Torsten Spexard, Britta Wrede, Gernot Horstmann, and Thurid Vogt. 2006. Playing a different imitation game: Interaction with an Empathic Android Robot. In *2006 6th IEEE-RAS International Conference on Humanoid Robots*, December 2006. 56–61. . https://doi.org/10.1109/ICHR.2006.321363

[33] Yoyo Tsung-Yu Hou, Wen-Ying Lee, and Malte Jung. 2023. "Should I Follow the Human, or Follow the Robot?" — Robots in Power Can Have More Influence Than Humans on Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, April 19, 2023, Hamburg Germany. ACM, Hamburg Germany, 1–13. . https://doi.org/10.1145/3544548.3581066

[34] Peter H. Jr Kahn, Hiroshi Ishiguro, Batya Friedman, Takayuki Kanda, Nathan G. Freier, Rachel L. Severson, and Jessica Miller. 2007. What is a Human?: Toward psychological benchmarks in the field of human–robot interaction. *Interaction Studies* 8, 3 (January 2007), 363–390. https://doi.org/10.1075/is.8.3.04kah

[35] Merel Keijsers, Christoph Bartneck, and Friederike Eyssel. 2021. What's to bullying a bot?: Correlates between chatbot humanlikeness and abuse. *Interaction Studies* 22, 1 (September 2021), 55–80. https://doi.org/10.1075/is.20002.kei

[36] Sara Kiesler, Aaron Powers, Susan R. Fussell, and Cristen Torrey. 2008. Anthropomorphic Interactions with a Robot and Robot–like Agent. *Social Cognition* 26, 2 (April 2008), 169–181. https://doi.org/10.1521/soco.2008.26.2.169

[37] Sara Kiesler, Lee Sproull, and Keith Waters. 1996. A prisoner's dilemma experiment on cooperation with people and human-like computers. *Journal of Personality and Social Psychology* 70, 1 (1996), 47–65. https://doi.org/10.1037/0022-3514.70.1.47

[38] Rabia I. Kodapanakkal, Mark J. Brandt, Christoph Kogler, and Ilja van Beest. 2020. Self-interest and data protection drive the adoption and moral acceptability of big data technologies: A conjoint analysis approach. *Computers in Human Behavior* 108, (July 2020), 106303. https://doi.org/10.1016/j.chb.2020.106303

[39] Spyros Kokolakis. 2017. Privacy attitudes and privacy behaviour: A review of current research on the privacy paradox phenomenon. *Computers & Security* 64, (January 2017), 122–134. https://doi.org/10.1016/j.cose.2015.07.002

[40] Dennis Küster, Aleksandra Swiderska, and David Gunkel. 2020. I saw it on YouTube! How online videos shape perceptions of mind, morality, and fears about robots. *New Media & Society* (September 2020), 1461444820954199. https://doi.org/10.1177/1461444820954199

[41] Ali Ladak. 2023. What would qualify an artificial intelligence for moral standing? *AI Ethics* (January 2023). https://doi.org/10.1007/s43681-023-00260-1

[42] Ali Ladak, Matti Wilks, and Jacy Reese Anthis. 2023. Extending Perspective Taking to Nonhuman Animals and Artificial Entities. *Social Cognition* 41, 3 (June 2023), 274–302. https://doi.org/10.1521/soco.2023.41.3.274

[43] Linnea Laestadius, Andrea Bishop, Michael Gonzalez, Diana Illenčík, and Celeste Campos-Castillo. 2022. Too human and not human enough: A grounded theory analysis of mental health harms from emotional dependence on the social chatbot Replika. *New Media & Society* (December 2022), 14614448221142007. https://doi.org/10.1177/14614448221142007

[44] Minha Lee, Gale Lucas, Johnathan Mell, Emmanuel Johnson, and Jonathan Gratch. 2019. What's on Your Virtual Mind? Mind Perception in Human-Agent Negotiations. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (*IVA '19*), July 01, 2019, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 38–45. . https://doi.org/10.1145/3308532.3329465

[45] Shane Legg and Marcus Hutter. 2007. A Collection of Definitions of Intelligence. *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms* (2007), 17–24.

[46] Gabriel Lima, Changyeon Kim, Seungho Ryu, Chihyung Jeon, and Meeyoung Cha. 2020. Collecting the Public Perception of AI and Robot Rights. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2 (October 2020), 135:1-135:24. https://doi.org/10.1145/3415206

[47] Eric Martínez and Christoph Winter. 2021. Protecting Sentient Artificial Intelligence: A Survey of Lay Intuitions on Standing, Personhood, and General Legal Protection. *Frontiers in Robotics and AI* 8, (2021), 367. https://doi.org/10.3389/frobt.2021.788355

[48] Clifford Nass, B. J. Fogg, and Youngme Moon. 1996. Can computers be teammates? *International Journal of Human-Computer Studies* 45, 6 (December 1996), 669–678. https://doi.org/10.1006/ijhc.1996.0073

[49] Sari R. R. Nijssen, Barbara C. N. Müller, Rick B. van Baaren, and Markus Paulus. 2019. Saving the Robot or the Human? Robots Who Feel Deserve Moral Care. *Social Cognition* 37, 1 (February 2019), 41-S2. https://doi.org/10.1521/soco.2019.37.1.41

[50] T. Nomura, T. Kanda, T. Suzuki, and K. Kato. 2004. Psychology in human-robot communication: an attempt through investigation of negative attitudes and anxiety toward robots. In *RO-MAN 2004. 13th IEEE International Workshop on Robot and Human Interactive Communication (IEEE Catalog No.04TH8759)*, September 2004. 35–40. . https://doi.org/10.1109/ROMAN.2004.1374726

[51] Tatsuya Nomura, Takayuki Uratani, Takayuki Kanda, Kazutaka Matsumoto, Hiroyuki Kidokoro, Yoshitaka Suehiro, and Sachie Yamada. 2015. Why Do Children Abuse Robots? In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction Extended Abstracts* (*HRI'15 Extended Abstracts*), March 02, 2015, New York, NY, USA. Association for Computing Machinery, New

York, NY, USA, 63–64. . https://doi.org/10.1145/2701973.2701977

[52] Raymond Paternoster, Robert Brame, Paul Mazerolle, and Alex Piquero. 1998. Using the Correct Statistical Test for the Equality of Regression Coefficients. *Criminology* 36, 4 (1998), 859–866. https://doi.org/10.1111/j.1745-9125.1998.tb01268.x

[53] Janet V. T. Pauketat and Jacy Reese Anthis. 2022. Predicting the moral consideration of artificial intelligences. *Computers in Human Behavior* 136, (November 2022), 107372. https://doi.org/10.1016/j.chb.2022.107372

[54] Jared Piazza, Justin F. Landy, and Geoffrey P. Goodwin. 2014. Cruel nature: Harmfulness as an important, overlooked dimension in judgments of moral standing. *Cognition* 131, 1 (April 2014), 108–124. https://doi.org/10.1016/j.cognition.2013.12.013

[55] Jared Piazza and Steve Loughnan. 2016. When Meat Gets Personal, Animals' Minds Matter Less: Motivated Use of Intelligence Information in Judgments of Moral Standing. *Social Psychological and Personality Science* 7, 8 (November 2016), 867–874. https://doi.org/10.1177/1948550616660159

[56] Aaron Powers, Sara Kiesler, Susan Fussell, and Cristen Torrey. 2007. Comparing a computer agent with a humanoid robot. In *Proceedings of the ACM/IEEE international conference on Human-robot interaction* (*HRI '07*), March 10, 2007, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 145–152. . https://doi.org/10.1145/1228716.1228736

[57] Laurel D. Riek, Tal-Chen Rabinowitch, Bhismadev Chakrabarti, and Peter Robinson. 2009. Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, September 2009. 1–6. . https://doi.org/10.1109/ACII.2009.5349423

[58] Astrid M. Rosenthal-von der Pütten, Nicole C. Krämer, Laura Hoffmann, Sabrina Sobieraj, and Sabrina C. Eimler. 2013. An Experimental Study on Emotional Reactions Towards a Robot. *Int J of Soc Robotics* 5, 1 (January 2013), 17–34. https://doi.org/10.1007/s12369-012-0173-8

[59] Joshua Rottman, Charlie R. Crimston, and Stylianos Syropoulos. 2021. Tree-Huggers Versus Human-Lovers: Anthropomorphism and Dehumanization Predict Valuing Nature Over Outgroups. *Cognitive Science* 45, 4 (2021), e12967. https://doi.org/10.1111/cogs.12967

[60] Juliana Schroeder and Nicholas Epley. 2016. Mistaking minds and machines: How speech affects dehumanization and anthropomorphism. *Journal of Experimental Psychology: General* 145, 11 (2016), 1427–1437. https://doi.org/10.1037/xge0000214

[61] Julian Schuessler and Markus Freitag. 2020. Power Analysis for Conjoint Experiments. https://doi.org/10.31235/osf.io/9yuhp

[62] Eric Schwitzgebel and Mara Garza. 2015. A Defense of the Rights of Artificial Intelligences. *Midwest Studies in Philosophy* 39, (July 2015), 98–119. https://doi.org/10.1111/misp.12032

[63] Ava Elizabeth Scott, Daniel Neumann, Jasmin Niess, and PawełW. Woźniak. 2023. Do You Mind? User Perceptions of Machine Consciousness. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (*CHI '23*), April 19, 2023, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 1–19. . https://doi.org/10.1145/3544548.3581296

[64] Daniel B. Shank. 2012. Perceived Justice and Reactions to Coercive Computers1. *Sociological Forum* 27, 2 (2012), 372–391. https://doi.org/10.1111/j.1573-7861.2012.01322.x

[65] Daniel B. Shank. 2014. Impressions of computer and human agents after interaction: Computer identity weakens power but not goodness impressions. *International Journal of Human-Computer Studies* 72, 10 (October 2014), 747–756. https://doi.org/10.1016/j.ijhcs.2014.05.002

[66] Daniel B. Shank and Alyssa DeSanti. 2018. Attributions of morality and mind to artificial intelligence after real-world moral violations. *Computers in Human Behavior* 86, (September 2018), 401–411. https://doi.org/10.1016/j.chb.2018.05.014

[67] Yutaka Suzuki, Lisa Galli, Ayaka Ikeda, Shoji Itakura, and Michiteru Kitazaki. 2015. Measuring empathy for human and robot hand pain using electroencephalography. *Sci Rep* 5, 1 (November 2015), 15924. https://doi.org/10.1038/srep15924

[68] Aleksandra Swiderska and Dennis Küster. 2020. Robots as Malevolent Moral Agents: Harmful Behavior Results in Dehumanization, Not Anthropomorphism. *Cognitive Science* 44, 7 (2020), e12872. https://doi.org/10.1111/cogs.12872

[69] Justin Sytsma and Edouard Machery. 2012. The Two Sources of Moral Standing. *Rev.Phil.Psych.* 3, 3 (September 2012), 303–324. https://doi.org/10.1007/s13164-012-0102-7

[70] Xiang Zhi Tan, Marynel Vázquez, Elizabeth J. Carter, Cecilia G. Morales, and Aaron Steinfeld. 2018. Inducing Bystander Interventions During Robot Abuse with Social Mechanisms. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (*HRI '18*), February 26, 2018, New York, NY, USA. Association for Computing Machinery, New York, NY, USA, 169–177. . https://doi.org/10.1145/3171221.3171247

[71] Tetsushi Tanibe, Takaaki Hashimoto, and Kaori Karasawa. 2017. We perceive a mind in a robot when we help it. *PLOS ONE* 12, 7 (July 2017), e0180952. https://doi.org/10.1371/journal.pone.0180952

[72] Herman T. Tavani. 2018. Can Social Robots Qualify for Moral Consideration? Reframing the Question about Robot Rights. *Information* 9, 4 (April 2018), 73. https://doi.org/10.3390/info9040073

[73] Xijing Wang and Eva G. Krumhuber. 2018. Mind Perception of Robots Varies With Their Economic Versus Social Function. *Front. Psychol.* 9, (2018). https://doi.org/10.3389/fpsyg.2018.01230

[74] Adrian F. Ward, Andrew S. Olsen, and Daniel M. Wegner. 2013. The Harm-Made Mind: Observing Victimization Augments Attribution of Minds to Vegetative Patients, Robots, and the Dead. *Psychol Sci* 24, 8 (August 2013), 1437–1445. https://doi.org/10.1177/0956797612472343

[75] Adam Waytz, John Cacioppo, and Nicholas Epley. 2010. Who Sees Human?: The Stability and Importance of Individual Differences in Anthropomorphism. *Perspect Psychol Sci* 5, 3 (May 2010), 219–232. https://doi.org/10.1177/1745691610369336

[76] Adam Waytz, Joy Heafner, and Nicholas Epley. 2014. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology* 52, (May 2014), 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

[77] Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. 2015. Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction. *Int J of Soc Robotics* 7, 3 (June 2015), 347–360. https://doi.org/10.1007/s12369-014-0267-6

[78] Jakub Złotowski, Ewald Strasser, and Christoph Bartneck. 2014. Dimensions of Anthropomorphism: From Humanness to Humanlikeness. In *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2014. 66–73. .

[79] Jakub Złotowski, Kumar Yogeeswaran, and Christoph Bartneck. 2017. Can we control it? Autonomous robots threaten human identity, uniqueness, safety, and resources. *International Journal of Human-Computer Studies* 100, (April 2017), 48–54. https://doi.org/10.1016/j.ijhcs.2016.12.008