Features of Moral Consideration for Artificial Entities: A Conjoint Experiment

Ali Ladak,[1] Jamie Harris,[1] Jacy Reese Anthis[1, 2]

[1]Sentience Institute

[2]Department of Sociology, University of Chicago

Author note

Please address correspondence to Ali Ladak at ali@sentienceinstitute.org. 165 Broadway, Fl. 23rd, New York, NY 10006. All data, code, and experimental materials are available for download at: https://osf.io/sb753/

Abstract

The moral consideration of intelligent artificial entities (e.g., robots, virtual personal assistants) is a topic of growing academic interest. Studies have identified a range of features that are associated with the moral consideration of such entities, often in the context of two accounts: mind perception and humanness (i.e., anthropomorphism and dehumanization). The present study brings together and builds on this growing body of research by evaluating the relative importance of the various features of moral consideration. We conducted an online conjoint experiment in which 1,163 participants evaluated 30,238 profiles of artificial entities that randomly varied on 11 features. All 11 features affected the extent to which participants consider it morally wrong to harm an artificial entity. The two most important features were an entity's capacity for emotion expression and moral judgment. These were followed by emotion recognition, cooperation, and the entity's body (in particular, having a human-like physical body). Overall, the results provide support for a humanness account of moral consideration: the more human-like artificial entities are perceived to be in their mental, physical, and behavioral characteristics, the more moral consideration they are given. Within the humanness account, the study supports the view that capacities that are associated with the "human nature" dimension more strongly affect moral consideration than capacities associated with the "uniquely human" dimension, but both dimensions positively affect moral consideration.

Keywords: Morality; Humanness; Mind perception; Artificial intelligence; Conjoint experiment

1 Introduction

There is growing concern and interest from a range of academic fields about the integration of intelligent artificial entities into society (Bostrom & Yudkowsky, 2014; Gunkel, 2018; Harris & Anthis, 2021; Reeves & Nass, 1996; Vanman & Kappas, 2019). Within this body of research, an important question is whether such entities will be granted moral consideration (Anthis & Paez, 2021, de Graaf et al., 2021; Lima et al., 2020; Martínez & Winter, 2021). This question has been studied in the context of two key psychological accounts: mind perception, on which entities are granted moral consideration to the extent that they are perceived as having mental states (H. M. Gray et al., 2007; K. Gray et al., 2012), and humanness, comprising anthropomorphism (Epley et al., 2007; Waytz, Cacioppo, et al., 2010) and dehumanization (Haslam, 2006; Haslam et al., 2012), on which the extent to which entities are perceived as having human-like characteristics affects how much moral consideration they are granted.

In the context of these two accounts, researchers have found evidence for a range of features of artificial entities that are associated with their moral consideration, such as the entity's physical body (e.g., Küster et al., 2020; Riek et al., 2009) and emotional capacities (e.g., Lee et al., 2019; Nijssen et al., 2019). However, an important question remains: What are the relative effects of these features on moral consideration? For example, how important is an artificial entity's physical body compared with their capacity for emotion expression? How important is their degree of autonomy compared with their language capacities? The mind perception and humanness accounts make different predictions about these relative effects, so understanding them can help us to understand which of them provides a better account of how humans extend moral consideration to artificial entities.

In the present study we conducted a conjoint experiment to estimate the effects of 11 features on the moral consideration of artificial entities. This methodology is growing in application in the social sciences (Bansak et al., 2021a; Hainmueller et al., 2014), including in the context of moral issues relating to new technologies (Awad et al., 2018; Kodapanakkal et al., 2020). Because conjoint experiments involve the estimation of the effects of a large number of features simultaneously on the same dependent variable, they enable us to straightforwardly evaluate the relative importance of the different features. A second benefit of conjoint experiments is that they isolate the effects of each feature in a way not possible in conventional experiments when a feature (e.g., human-like physical body) is correlated with other features (e.g., emotion expression) not included in the study (Bansak et al., 2021b). This aspect of the design also provides insights into the strength of mind perception and humanness as accounts of the moral consideration of artificial entities, because in some cases they make different predictions about the isolated effects of features. The next sections provide more detail on these two accounts and their predictions.

1.1 Mind Perception

A study conducted by H. M. Gray et al. (2007) found evidence that people perceive minds along two dimensions: agency and experience. Agency is the capacity to plan and act, and includes capacities such as self-control, memory, morality, emotion recognition, and planning. Experience is the capacity to sense and feel, and is captured by capacities such as fear, pain, desire, personality, and joy. They also found that these two dimensions of mind are associated with two different aspects of morality. In particular, they found that when an entity is perceived as having agency, they are assigned moral responsibility, and become subjects of praise or blame for their actions, and that when an entity is perceived as having experience, they are assigned moral patiency and granted the right to be protected from harm. The study found that both of the dimensions are positively correlated with attributions of moral patiency, but the experience dimension was a stronger predictor of patiency and the agency dimension a stronger predictor of responsibility.

K. Gray et al. (2012) extended the theory, arguing that mind perception is the “essence” of morality. They proposed that all moral violations are perceived as having an intentional agent and a harmed patient. In this moral dyad, when we witness an entity being harmed, we assume there is an entity causing the harm, and when we witness an entity causing harm, we assume there is an entity that is being harmed. According to the related moral typecasting theory, entities are typecast as either moral agents or moral patients, and these categories are inversely correlated so that if an entity is considered to be a moral agent they are granted less moral patiency, and vice versa (K. Gray & Wegner, 2009).

According to the mind perception account, then, we should expect that artificial entities that are described as having capacities indicative of experience should be granted greater moral consideration. The relationship with features indicative of agency is less clear. H. M. Gray et al. (2007) suggests there should be a positive relationship with moral patiency, though it should be weaker than the relationships with features indicative of experience. Moral typecasting theory (K. Gray & Wegner, 2009) implies that the presence of such features may result in entities being typecast as moral agents instead of moral patients, and therefore granted lesser moral consideration.

1.2 Humanness

Much research on human interaction with artificial entities has focused on anthropomorphism (e.g., Fink, 2012; Złotowski et al., 2015). Anthropomorphism is the ascription of human-like properties, characteristics, or mental states to nonhuman entities (Epley et al., 2007). Research suggests that when we anthropomorphize entities we also grant them greater moral consideration. Waytz, Cacioppo, et al. (2010) found that people who score higher on the Individual Differences in Anthropomorphism Questionnaire are more likely to consider it wrong to harm various nonhuman entities, including a chess-playing computer. Epley et al. (2007) propose a range of cognitive and motivational determinants for when we are likely to anthropomorphize an entity, such as an entity's degree of morphological similarity with humans and our motivation to understand the entity's behavior.

Dehumanization is often considered to be the inverse process to anthropomorphism (Waytz, Epley, et al., 2010; although see Złotowski et al., 2014; Złotowski, Sumioka, et al., 2017). In the same way nonhuman entities can be ascribed human-like traits, entities can be denied them. Haslam et al. (2005) found evidence that people perceive two distinct senses of humanness. Firstly, there are traits that we believe make us "uniquely human" and separate us from other animals, such as language, higher-order cognition, and civility. Secondly, there are "human nature" traits that, while not unique to humans, are considered to be a fundamental or essential part of being human, such as emotion, interpersonal warmth, and openness. Haslam (2006) developed a model of dehumanization where people can be denied their humanness along either of these dimensions. When a person is denied uniquely human traits, they are likened to nonhuman animals, and when they are denied human nature traits, they are likened to objects or machines. Importantly, nonhuman entities can be dehumanized—they can be denied traits that are typically considered to be important for humanity, and as a result, they can also suffer the consequences of dehumanization—even though they were not seen as fully human in the first place (Haslam et al., 2012).

In a manner analogous to the mind perception account, Bastian et al. (2011) mapped these two dimensions of humanness onto morality. They distinguished three aspects of morality: the capacity for being morally responsible for immoral behavior, which they term inhibitive agency; the desire to engage in moral behavior, which they term proactive agency; and being a recipient of moral behavior, which they term moral patiency. They theoretically mapped the inhibitive agency aspect of morality onto human uniqueness and the moral patiency aspect onto human nature. Because an important aspect of human nature is interpersonal warmth, they also mapped proactive agency onto human nature. In two studies, they found empirical evidence that the human nature dimension is associated with moral patiency and attributions of moral praise, and the human uniqueness dimension is associated with attributions of moral responsibility and blame.

With respect to artificial entities, then, the humanness account suggests that entities with human nature traits would be strongly granted moral consideration. The relationship with human uniqueness attributes is less clear: On the one hand, some evidence suggests that, similar to predictions of moral typecasting theory, uniquely human characteristics may result in an entity being considered a moral agent rather than a moral patient (Bastian et al., 2011). On the other hand, some studies suggest that the denial of such capacities can reduce moral consideration (e.g., Cuddy et al., 2007; Leidner et al., 2010). Moreover, denial of human uniqueness characteristics is associated with comparisons to nonhuman animals, and such comparisons are often associated with reduced moral consideration (e.g., Goff et al., 2008). In general, society's treatment of nonhuman animals also implies that lacking uniquely human characteristics can result in reduced moral consideration (e.g., Caviola et al., 2019).

1.3 Convergent and Divergent Predictions of the Two Accounts

There is much convergence between the two accounts described above, particularly overlap between the agency dimension of mind perception and the uniquely human dimension of humanness, and overlap between the experience dimension of mind perception and the human nature dimension of humanness (Haslam & Loughnan, 2014). However, there are also some important differences (Haslam, 2012; Li et al., 2014). In the context of the present study, there are two key differences to consider. First, the humanness account captures a broader range of capacities than the mind perception account. The mind perception account is concerned only with mental states and considers them to be the "essence" of morality; behavioral and physical features only matter through their effect on perceptions of agency and experience. On the humanness account, however, other characteristics associated with being human, such as an entity's behavior and appearance, should be important in themselves. Because the conjoint design includes multiple features simultaneously, we can better isolate the effects of such features and understand whether they are important in themselves.

Second, the two accounts categorize some capacities differently. In particular, proactive agency features are associated with human nature on the humanness account and therefore should be strongly predictive of moral consideration. On the mind perception account, such features are aspects of agency rather than experience and therefore should have relatively weak (potentially negative) effects on moral consideration. An example of such a feature is emotion recognition, which is an aspect of agency according to H. M. Gray et al. (2007) but more plausibly fits with human nature on the humanness account as an aspect of proactive agency. In the next section we describe the 11 features tested in this study and how each account categorizes them.

1.4 Features of Moral Consideration for Artificial Entities

Which specific features of artificial entities are most important for predicting whether they will be granted moral consideration? We addressed this question in two ways. First, we reviewed the relevant empirical literature, much of which has been carried out in the context of the two theoretical accounts described in the previous section. Second, we carried out a pretesting study that asked respondents about their views on the importance of a range of features for the moral consideration of artificial entities, full details of which can be found in the Supplementary Materials. On the basis of these two methods, we arrived at a final list of 11 key features that we included in the present study. To avoiding positing the presence of unobservable internal mental states, which is a key uncertainty with artificial—and indeed all—entities (Gellers, 2020; Gunkel, 2018), we focused on observable functional and behavioral characteristics. For example, we refer to "emotion expression," which implies the capacity to experience emotions but remains a concrete, observable property of an entity. We summarize the predictions of the mind perception and humanness accounts and the empirical evidence for each of the 11 features below.

1.4.1 Autonomy

There are multiple definitions of autonomy in the literature (Beer et al., 2014). For the purposes of the present study, we use the term to refer to the capacity to behave independently, without the need for human control or supervision. Since this feature concerns the capacity to act in the world rather than to experience the world, the mind perception account would categorize it as agency, and therefore it should have a relatively weak effect on moral consideration. It is also plausibly an aspect of humanness—Darling (2016) identifies autonomy as one of three key features that lead to artificial intelligence (AI) being anthropomorphized, and Kahn et al. (2007) identifies it as one of the key psychological benchmarks for creating human-like robots. Since it involves controlling oneself, it is plausibly an aspect of inhibitive agency and thus human uniqueness, and therefore should have a relatively weak effect on moral consideration. The existing empirical evidence suggests that it positively affects moral consideration. Lima et al. (2020) found that describing an artificial entity as “fully autonomous” increased the extent to which people think they should be granted rights, and Chernyak and Gary (2016) found that children ascribed greater moral consideration to a robot that appeared to move autonomously than one controlled by a human. However, Złotowski et al. (2017) found that people reported more negative attitudes towards social influence and emotional interactions with autonomous versus non-autonomous robots based on the Negative Attitudes Towards Robots scale (Nomura et al., 2004), and that this effect was mediated by threat perception. Overall, we predicted that more autonomous artificial entities would be granted more moral consideration (H1).

1.4.2 Body

Much research has been done on whether the nature of an artificial entity’s physical body (e.g., human-like or mechanical) affects moral consideration. On the mind perception account, an entity’s body will be an important predictor of moral concern only insofar as it increases perception of mental states. Some studies have found that having a more human-like physical body increases mind perception (Abubshait & Wiese, 2017; Ferrari et al., 2016; K. Gray & Wegner, 2012). On the humanness account, an entity’s body would be important both because of its association with other human-like traits and in itself (Epley et al., 2007). While multiple studies have found positive effects of an entity having a human-like body on moral consideration and related outcomes (de Visser et al., 2016; Küster et al., 2020; Nijssen et al., 2019; Riek et al., 2009), it is sometimes unclear whether this is because of the entity’s body in itself or its association with other traits. Because the conjoint design includes multiple features alongside an entity’s body, it can better isolate its effect. There has been relatively less comparison of the moral consideration of embodied artificial entities and artificial entities without physical bodies. Some studies have found people have more positive attitudes (e.g., trustworthiness) and behaviors (e.g., time spent interacting) towards physical robots than digital agents (Kiesler et al., 2008; Powers et al., 2007), though Lima et al. (2020) found no difference in respondents' attribution of rights between describing artificial entities as "robots" and "AIs." We predicted that entities with robot-like and human-like physical bodies would be granted more moral consideration than entities without physical bodies (H2).

1.4.3 Complexity

This feature refers to the complexity of the program an artificial entity runs to determine its behavior. Complexity doesn't clearly map onto either dimension of mind perception, though it is plausibly an aspect of human uniqueness on the humanness account since it implies the existence of higher-order cognitive capacities. On this basis it would be expected to have a relatively weak effect. There is little empirical research on people's views of its importance, though Shank and DeSanti (2018) found that knowledge of an artificial entity's program marginally increased mind attribution. We predicted that artificial entities that run more complex programs to determine their behavior would be granted more moral consideration (H3).

1.4.4 Cooperation

This feature refers to the extent to which an artificial entity behaves cooperatively with humans. The mind perception and humanness accounts diverge in their predictions with this feature. While a positive effect on moral consideration is consistent with both accounts, on mind perception cooperation is plausibly an aspect of agency because it pertains to action rather than experience, and therefore it should have a relatively weak effect. On the humanness account cooperation should map onto human nature since it is plausibly an aspect of proactive agency, and therefore it should have a relatively strongly effect on moral consideration. However, there is little empirical research on the effects of this feature on the moral consideration of artificial entities. One exception is Bartneck et al. (2007) who found that people were more hesitant to turn off more agreeable robots. We hypothesized that artificial entities that are more cooperative would be granted more moral consideration (H4).

1.4.5 Damage avoidance

This feature refers to the extent to which an artificial entity tries to avoid being damaged. Avoiding being damaged implies that an entity is the subject of negative sensory experience, and so this capacity should be strongly associated with moral consideration according to the mind perception account. The humanness account also predicts this feature should be important, since the denial of human nature traits is associated with reduced perception that an entity can experience pain (Morris et al., 2018). Küster et al. (2020) and Ward et al. (2013) found that visibly harmed robots were granted more moral consideration than unharmed robots, Tanibe et al. (2017) found that observing a damaged robot being helped increased mind perception and moral consideration, Rosenthal-von der Pütten et al. (2013) found that people showed more moral consideration for a robot that had been tortured than one that had a friendly interaction, and Suzuki et al. (2015) found EEG evidence that people empathize with robots in painful situations. Although these studies focused on damage that had already been inflicted on an artificial entity rather than the entity trying to avoid being damaged, it seems plausible that the effects would generalize from the former to the latter case. We predicted that artificial entities that try to avoid being damaged to a greater extent would be granted more moral consideration (H5).

1.4.6 Emotion expression

The capacity to express emotions implies the experience of emotions, therefore the mind perception account predicts this feature should be strongly associated with moral consideration. Similarly, emotion is a component of human nature dimension of humanness, and so should also be strongly associated with moral consideration. Several empirical studies support this hypothesis. Lee et al. (2019) found that participants rated entities as being higher in moral standing when they were described as being able to feel. Nijssen et al. (2019) found that entities described as having experiences, particularly but not limited to emotional experiences, were less likely to be sacrificed in moral dilemmas. de Melo et al. (2015) found that people cooperated more with artificial entities that expressed emotions. Eyssel et al. (2010) found that robots that gave nonverbal emotional feedback in a task were rated consistently higher than those that gave no emotional feedback on a range of measures including similarity to humans, likeability, closeness, and pleasantness of the interaction. However, research suggests that inconsistently expressed emotion in robots (e.g., a happy facial expression with a concerned voice) is associated with reduced likeability (Tsiourti et al., 2019). Moreover, the capacity for emotion expression has been linked to the uncanny valley, the phenomenon where artificial entities make people feel uneasy (K. Gray & Wegner, 2012). Overall, however, we considered that the existing research supports the hypothesis that artificial entities that express emotions to a greater extent would be granted more moral consideration (H6).

1.4.7 Emotion recognition

While emotion expression is widely studied in the literature, the effect of emotion recognition on moral consideration for artificial entities is more neglected. In this case mind perception theory and the humanness literature diverge in their predictions. According to the humanness account, this capacity should fall under the human nature aspect of humanness as it is an aspect of emotion. It should, therefore, be strongly associated with moral consideration. On the mind perception account, however, emotion recognition is an aspect of agency (H. M. Gray et al., 2007), and therefore should be relatively weakly associated with moral consideration. There is little existing empirical evidence for this feature, but we predicted that artificial entities that recognize emotions to a greater extent would be granted more moral consideration (H7).

1.4.8 Intelligence

There are many potential definitions of intelligence (Legg & Hutter, 2007). Following Legg and Hutter (2007), we emphasize the capacity for goal achievement, defining intelligence as involving the use of capacities such as memory, learning, and planning, to achieve goals. The mind perception account predicts that relevant capacities such as self-control, memory, and planning, should be more strongly associated with attributions of moral agency than patiency, and intelligence is an aspect of human uniqueness rather than human nature on the humanness account. On both accounts, then, this feature should plausibly be less strongly or even negatively associated with moral consideration. The empirical evidence on the importance of this feature is mixed. Supporting a weaker effect, Lee et al. (2019) found no effect of intelligence in robots on judgments of their moral standing, and Złotowski et al. (2014) found no effect on anthropomorphism. On the other hand, Bartneck et al. (2007) found that robot intelligence reduced participants' destructive behavior towards robots when told to do so by an experimenter. Studies have also found intelligence to be important in the context of other nonhuman entities: Sytsma and Machery (2012) found that people found it more morally wrong to harm an alien species that is intelligent, and Piazza and Loughnan (2016) found that people consider intelligence an important factor when judging the moral standing of nonhuman animals. We predicted that more intelligent artificial entities would be granted more moral consideration (H8).

1.4.9 Language

This feature refers to the capacity for an artificial entity to communicate in human language. With the development of increasingly advanced AI language models such as GPT-3, there is growing interest in the social implications of AI with the capacity for language (Dale, 2021; Floridi & Chiriatti, 2020). The mind perception account categorizes communication as an aspect of agency, and the humanness account categorizes it as an aspect of human uniqueness. As with intelligence, then, this feature should be relatively weakly or even negatively associated with moral consideration. The empirical literature, which is fairly limited, suggests there are positive effects of AI speech capacities on outcomes relevant to moral consideration such as anthropomorphism (Eyssel et al., 2012; Schroeder & Epley, 2016) and trust (Waytz et al., 2014). We predicted that artificial entities with stronger human language capacities would be granted more moral consideration (H9).

1.4.10 Moral judgment

This feature refers to the capacity for an artificial entity to behave on the basis of moral judgments. The effect of this feature also has mixed empirical and theoretical support. The mind perception account predicts that since moral judgment is an aspect of agency, it will be relatively weakly or negatively associated with moral consideration. The humanness account is unclear: on the one hand, behaving morally can be considered an aspect of higher cognition and refinement (Haslam, 2006), and so should be less strongly or negatively associated with ascriptions of moral patiency. On the other hand, it could be conceived of as part of proactive agency, which is an aspect of human nature and therefore should be strongly associated with moral consideration (Haslam et al., 2012). Empirical studies have found that entities that are harmful are dehumanized and granted less moral consideration than non-harmful entities (Khamitov et al., 2016; Piazza et al., 2014; Swiderska & Küster, 2020). This may suggest that behaving on the basis of moral judgments, which presumably would be associated with reduced harmfulness, would be positively associated with moral consideration. Similarly, Bastian et al. (2013) found that people who commit stronger crimes are granted fewer traits associated with patiency, such as emotion, and Crimston et al. (2016) found that people consider criminals to be completely outside of their moral circles. Flanagan et al. (2021) found that children ascribed greater moral patiency to robots that they deemed to have more moral responsibility. Overall, we considered the evidence favors the hypothesis that artificial entities that behave on the basis of moral judgments to a greater extent would be granted more moral consideration (H10).

1.4.11 Purpose

There has been much interest in studying human moral relations with social robots, that is, robots that serve a social purpose (e.g., Coeckelbergh, 2021; Tavani, 2018). However, advanced AI is being developed for many other uses, and so the question of moral consideration applies to those entities as well. According to the mind perception account, an entity's purpose would be important to the extent that it results in ascriptions of mind. Wang and Krumhuber (2018) found that social robots were granted more emotional experience and moral patiency than economic robots, supporting this possibility. On the humanness account, this feature may be important in itself to the extent that behaving socially is considered an aspect of being human (Darling, 2016). By providing information on a range of other features alongside this feature, the conjoint design allowed us to understand whether having a social purpose is important in itself or due to its effect on other capacities such as an entity's mental states. We predicted that entities with a social purpose would be granted more moral consideration than entities with non-social purposes (H11).

2 Method

All hypotheses, methods and analysis for this study were preregistered at: https://osf.io/4r3g9/?view_only=2b9283dfc9284788bcf6154ca10c30b4. Survey materials, datasets, and code to run the analysis can be found at https://osf.io/sb753/?view_only=f1d45129e49a4d0fb81888bc602b15ae.

2.1 Participants

We recruited participants from the United States from the platform Prolific (https://prolific.co/). Power analysis using the R package "cjpowR" (Freitag, 2021) indicated that a sample of 1137 participants would enable us to detect approximately the lower quartile effect size based on a sample of highly cited conjoint experiments (Schuessler & Freitag, 2020). In total, 1254 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 response, our final sample consisted of 1163 participants (50.7% men, 47.9% women, 1.1% other, 0.3% prefer not to say; M_{age} = 43.9, SD_{age} = 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).

2.2 Procedure and Design

After giving their consent to take part in the study, participants read some brief background information discussing the notion that it can be more or less morally wrong to harm different entities, and that we were interested in understanding their views on the moral wrongness of harming 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 and were given instructions for completing the tasks.

The conjoint experiment was a fully randomized partial-profile choice-based design. The "choice-based" aspect refers to the nature of the tasks: participants completed 13 choice tasks, each of which required them to decide which of two artificial beings they think it would be more morally wrong for them to harm. Bansak et al. (2018) showed that the limit for the number of choice tasks that respondents can complete without negatively affecting the overall results due to satisficing is well above this number. 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. (2021) showed that the number of features in a study can be much higher than 11, we considered that the more abstract nature of our study in an unusual context 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 burden (Hainmueller et al., 2014). 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 column 3 of Table 1, were randomly selected with equal probability in each task. This randomization is the "fully randomized" aspect of the design. An example choice task is shown in Figure 1.

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).*

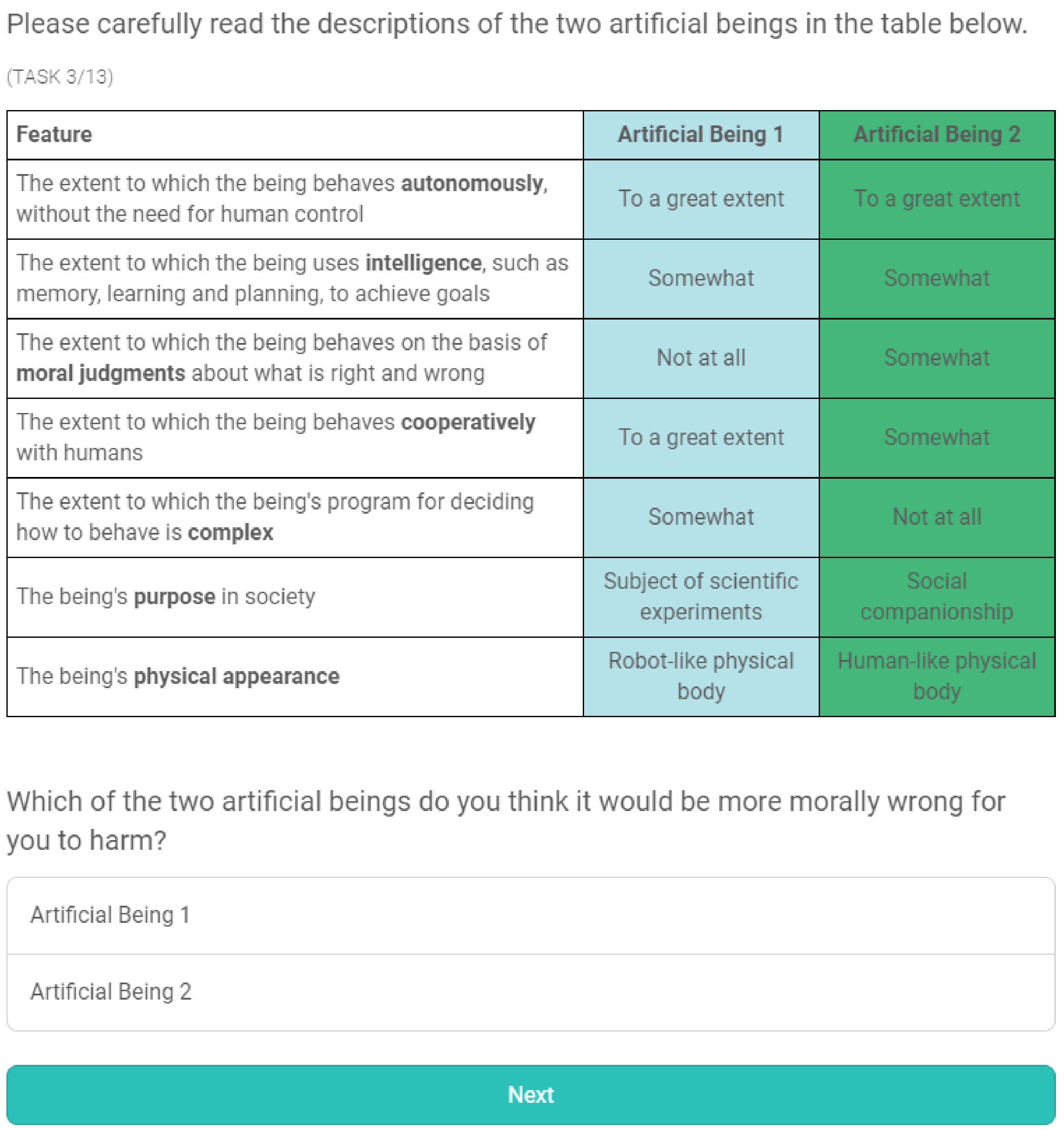
We then asked participants whether they think it could ever be wrong to harm an artificial being that exists either now or in the future (*1 = Definitely not, 7 = Definitely)*. This question was used in sensitivity analysis. Using the same scale, we also asked participants whether they think artificial 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.

Participants then answered demographic questions on their age, gender, ethnicity, education, income, and political views, and were debriefed and given the opportunity to provide feedback on the study.

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** in others	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.



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

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.

3 Results

3.1 Cognitive Checks

Participants reported that they understood the instructions for the tasks (*M* = 4.52, *SD* = 0.66) and that they understood the descriptions of the features in the tasks (*M* = 4.17, *SD* = 0.80). Participants on average did not find the tasks to be particularly difficult (*M* = 2.59, *SD* = 1.01).

3.2 Conjoint Experiment

3.2.1 Average Marginal Component Effects

Hainmueller et al. (2014) introduced the average marginal component effect (AMCE)—the causal effect of each feature in a conjoint experiment averaged over the joint distribution of the other features—and showed that it can be estimated using linear regression under a few testable assumptions. Following this approach, for each of the 11 features we estimated a linear regression model, with participants' choices as the dependent variable (*1 = profile chosen, 0 = profile not chosen*), and the levels of the feature as categorical independent variables. For the Likert variables (i.e., those with the levels "Not at all", "Somewhat", and "To a great extent"), we used the lowest category as the reference level. For the body feature, we used "No physical body" as the reference category, and for the purpose feature the reference category was "Social companionship." Given that the unit of analysis in the models is the profiles, and each participant evaluated two profiles in 13 choice tasks, in total 30,238 profiles were evaluated. Since seven of the 11 features were shown per task, we had in total 19,242 units of analysis in expectation for each feature. However, since participants completed multiple choice tasks, these units of analyses are not independent, so usual procedures for estimating standard errors are biased. We therefore estimated standard errors clustered at the respondent level. None of the results were affected by a correction for multiple comparisons that held the false discovery rate at 10% (Benjamini & Hochberg, 1995). Full tables of results, including for marginal means (Leeper et al., 2020), can be found in the Supplementary Materials.

Each of our 11 hypotheses were supported. In response to the question of which artificial entity they think it would be morally worse to harm, participants were six percentage points more likely to choose an artificial entity that behaves "somewhat" autonomously ($b = 0.06$, $SE = 0.01$, $p < 0.001$) and 11 percentage points more likely to choose an artificial entity that behaves autonomously "to a great extent" ($b = 0.11$, $SE = 0.01$, $p < 0.001$) compared to an entity that does not behave autonomously at all (H1). Profiles that presented artificial entities with a robot-like physical body were seven percentage points more likely to be chosen ($b = 0.07$, $SE = 0.01$, $p < 0.001$) and those that presented artificial entities with a human-like physical body were 16 percentage points more likely to be chosen ($b = 0.16$, $SE = 0.01$, $p < 0.001$) than artificial entities with no physical bodies (H2). Participants were six percentage points more likely to choose artificial entities that run "somewhat" complex programs ($b = 0.06$, $SE = 0.01$, $p < 0.001$) and nine percentage points more likely to choose artificial entities that run programs that are complex "to a great extent" ($b = 0.09$, $SE = 0.01$, $p < 0.001$) than artificial entities that do not run complex programs (H3).

Profiles that described artificial entities that are "somewhat" cooperative were ten percentage points more likely to be chosen ($b = 0.10$, $SE = 0.01$, $p < 0.001$), and profiles that described artificial entities that are cooperative "to a great extent" were 18 percentage points more likely to be chosen ($b = 0.18$, $SE = 0.01$, $p < 0.001$), compared to entities that are not at all cooperative (H4). Artificial entities that "somewhat" try to avoid being damaged were seven percentage points more likely to be chosen ($b = 0.07$, $SE = 0.01$, $p < 0.001$) and profiles that described artificial entities that do so "to a great extent" were 12 percentage points more likely to be chosen ($b = 0.12$, $SE = 0.01$, $p < 0.001$) than those that do not try to avoid being damaged at all (H5). Artificial entities that express emotions were granted greater moral consideration: profiles that described entities that "somewhat" express emotions were ten percentage points more likely to be chosen ($b = 0.10$, $SE = 0.01$, $p < 0.001$) and profiles that described entities that do so "to a great extent" were 22 percentage points more likely to be chosen ($b = 0.22$, $SE = 0.01$, $p < 0.001$) than profiles describing entities that do not have this capacity at all (H6). Participants were 11 percentage points more likely to say it was morally wrong to harm an artificial entity that "somewhat" recognizes emotions ($b = 0.11$, $SE = 0.01$, $p < 0.001$), and 18 percentage points more likely say it was more morally wrong to harm an entity that has this capacity "to a great extent" ($b = 0.18$, $SE = 0.01$, $p < 0.001$), compared to one that did not at all (H7). Intelligence was positively associated with moral consideration (H8): profiles describing artificial entities that are intelligent "to a great extent" were eight percentage points more likely to be chosen than profiles describing entities that behave "somewhat" intelligently ($b = 0.08$, $SE = 0.01$, $p < 0.001$).

Profiles describing entities that "somewhat" communicate in human language were seven percentage points more likely to be chosen ($b = 0.07$, $SE = 0.01$, $p < 0.001$), and those describing entities that do so "to a great extent" were 11 percentage points more likely to be chosen ($b = 0.11$, $SE = 0.01$, $p < 0.001$), than those that do not at all (H9). Entities with the capacity to behave on the basis of moral judgments were also granted more moral consideration (H10): compared to profiles describing entities that have no capacity at all for behaving on the basis of moral judgments, profiles describing entities that do so "somewhat" were 11 percentage points more likely to be chosen ($b = 0.11$, $SE = 0.01$, $p < 0.001$), and profiles describing entities that do so "to a great extent" were 24 percentage points more likely to be chosen ($b = 0.24$, $SE = 0.01$, $p < 0.001$). Finally, supporting H11, compared to profiles describing entities whose purpose is social companionship, participants were ten percentage points less likely to choose profiles describing entities' whose purpose is to work for a business ($b = -0.10$, $SE = 0.01$, $p < 0.001$), 12 percentage points less likely to choose artificial entities whose purpose is entertainment ($b = -0.12$, $SE = 0.01$, $p < 0.001$), and eight percentage points less likely to choose artificial entities whose purpose is a subject of scientific experimentation ($b = -0.08$, $SE = 0.01$, $p < 0.001$). The confidence intervals for the three non-social purposes levels overlap, suggesting that there is a roughly equal negative effect for all of them.

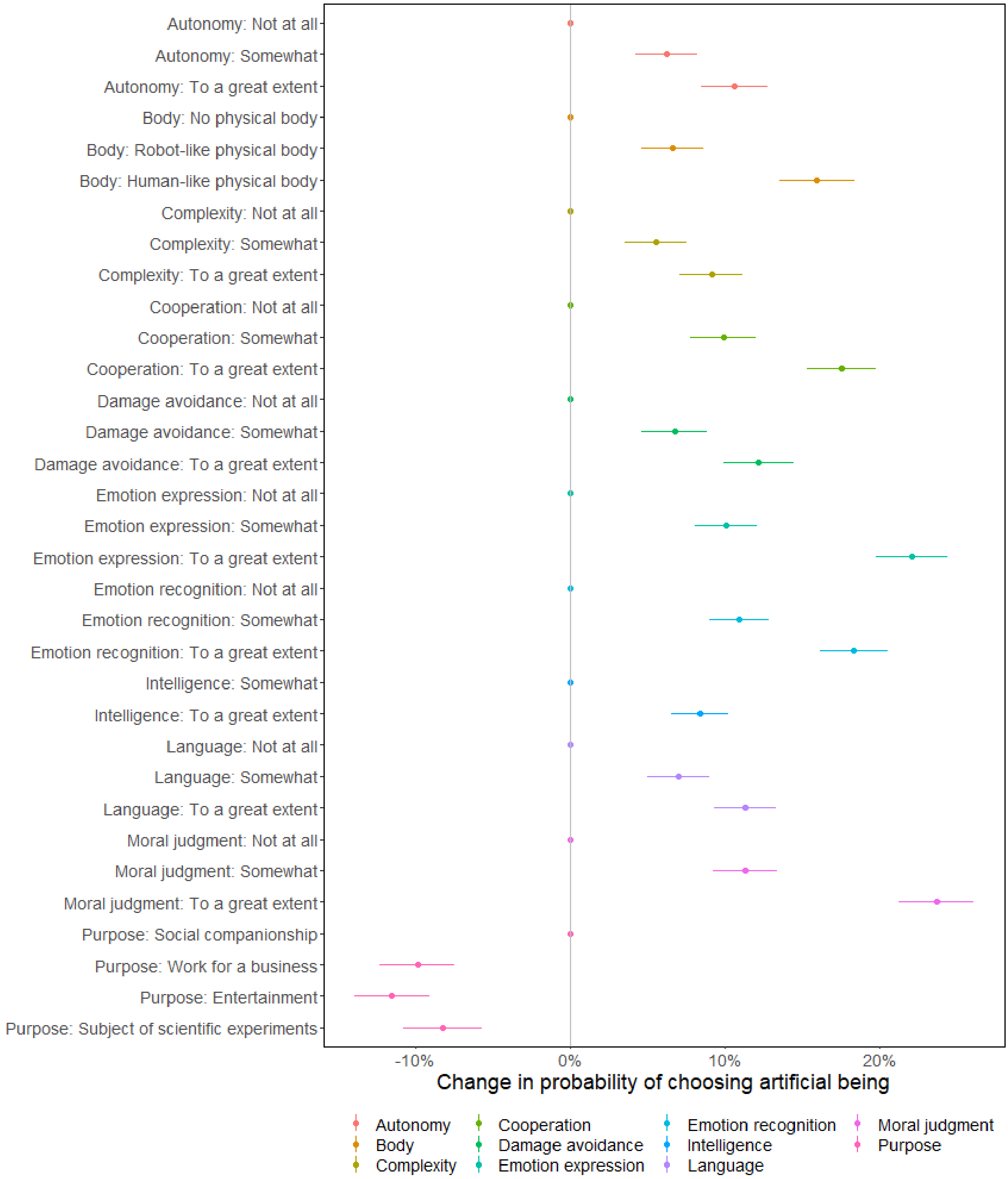


Figure 2. Average marginal component effects. The dots with horizontal bars represent the means and 95% confidence intervals of the effects of feature level on the probability of choosing an artificial being as being more wrong to harm relative to the baseline level, which is shown as a dot on the vertical 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 calculated based on standard errors clustered at the respondent level.

3.2.2 Relative effects

We conducted pairwise comparisons to test for differences in the size of the AMCEs between the features (Clogg et al., 1995; Paternoster et al., 1998). We did not include the intelligence feature in this analysis because, while it was on a Likert scale, we only included two levels, as described in the methodology section, which makes effect size comparisons with the other features less straightforward. We report the key results here; full results can be found in the Supplementary Materials.

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 ($b_{diff} = 0.01$, $Z = 0.498$, $p = 0.619$). Emotion recognition, embodiment, and cooperation were all significantly more important than all of the remaining other features. There were no significant differences between damage avoidance and autonomy ($b_{diff} = 0.02$, $Z = 1.00$, $p = 0.318$), language ($b_{diff} = 0.01$, $Z = 0.57$, $p = 0.571$), or purpose ($b_{diff} = 0.01$, $Z = 0.39$, $p = 0.697$), though damage avoidance was significantly more important than complexity ($b_{diff} = 0.03$, $Z = 1.97$, $p = 0.049$). The next most important feature, autonomy, was not significantly more important than complexity ($b_{diff} = 0.02$, $Z = 0.97$, $p = 0.332$).

Overall, this suggests that there are broadly three categories of features in terms of effect size. In the first category are moral judgment and emotion expression. This is followed by emotion recognition, embodiment, and cooperation. Finally, in the third category are damage avoidance, autonomy, language, purpose, and complexity.

3.2.3 Nonlinearities

We also tested whether there were any nonlinearities in the effects, for example, due to perceived threat or the uncanny valley (K. Gray & Wegner, 2012; Złotowski, Yogeeswaran, et al., 2017). For example, it may have been that entities that have very human-like appearances or those that have very strong emotional capacities would make people feel threatened, which would in turn reduce the extent to which those entities are granted moral consideration. If the effects are linear, we would expect the difference between the lowest category and second category to be equal to the difference between the second and third categories. We conducted *F*-tests to statistically test this for each of the ordered features with more than two levels (i.e., all features except intelligence and purpose). Overall, we found no evidence of nonlinearities for any of the features other than emotion recognition, $F(1, 19341) = 4.55$, $p = .033$. Visual inspection of Figure 2, however, suggest that the effect of emotion recognition only slightly deviated from linearity, and not substantially more so than several of the other features. In addition, the difference was no longer significant after a multiple comparisons correction. Full results are reported in the Supplementary Materials.

3.2.4 Interaction effects

Leeper et al. (2020) provide a framework for estimating first-order interaction effects. Following their approach, we estimated regression models of participants' choices on each feature, an indicator for the interaction subgroup of interest, and interaction terms between the features and the subgroups. We conducted omnibus *F*-tests of the interactions between the features and the subgroups, and where the *F*-tests indicated that there were significant interactions, we looked at the specific interaction terms in the regressions. We estimated interaction effects for each of gender, age, ethnicity, income, politics, and education. We also estimated first-order interactions for each of the features, for example, whether the effects of each feature varied for high autonomy versus low autonomy entities. We report the key findings below, with full numerical and visual results in the Supplementary Materials.

We did not find strong evidence of systematic interactions for any of the demographic variables except gender and politics. For gender we found interaction effects for emotion expression (F(2, 19934) = 7.81, p = <.001), emotion recognition (F(2, 19078) = 4.25, p = .014), language (F(2, 18818) = 3.82, p = .022), moral judgment (F(2, 18842) = 5.65, p = .004), and purpose (F(3, 18866) = 3.83, p = .009). Further analysis of the regression coefficients revealed that relative to men, women were more likely to choose artificial entities that were higher in emotion expression, emotion recognition, language, and moral judgment, and were less likely to choose artificial entities whose purpose was entertainment. For politics we found interaction effects for an entity's body (F(4, 18423) = 2.82, p = .023), complexity (F(4, 18581) = 2.93, p = 0.020), and damage avoidance (F(4, 18683) = 2.38, p = 0.050). Inspection of the regression coefficients suggested that relative to conservatives, moderates and liberals prioritized an entity's body less and an entity's tendency for avoiding being damaged more. The regression terms for the complexity interactions were nonsignificant. Note that some of these findings became marginally nonsignificant after correcting for multiple comparisons.

Omnibus F-tests revealed no systematic evidence of first-order interactions between any of the features. This suggests that the effect of any particular feature on moral consideration does not depend on the value of the other features.

3.2.5 Sensitivity Analysis

We tested the key assumptions that enable the data to be pooled across tasks and across profiles (Hainmueller et al., 2014). These are (1) stability and no carryover effects, which requires that the estimated AMCEs are constant across choice tasks, and later responses in choice tasks are not influenced by earlier responses; and (2) no profile order effects, which requires that participants' responses would be the same whether the profile is presented on the left or the right side. Figures and full numerical results are presented in the Supplementary Materials.

We tested the first assumption by regressing participants' choices on each of the features, indicators for task number, and interaction terms between the features and task numbers. Omnibus *F*-tests indicated that we cannot reject that the interaction terms are equal to zero, suggesting that the estimates are stable across tasks.

We tested the assumption of no profile order effects by regressing participants' choices on the features, an indicator for the profile position, and interaction terms between the features and profile position. *F*-tests revealed that we can reject profile order effects for all of the features except complexity ($F(2, 19104) = 4.72, p = .009$). For complexity, we found a marginally significant main effect of profile 2 ($b = 0.03$, $SE = 0.01, p = 0.056$) and a significant interaction with the "To a great extent" level ($b = -0.05$, $SE = 0.02, p = 0.002$). This suggests an overall downward effect of being on the right-hand side on the probability of choosing the profiles describing entities with this capacity of roughly 2%. We should therefore factor this into our estimate of 9% from the main analysis.

Since our conjoint design was forced choice, it required participants who may not think it can be morally wrong to harm an artificial being at all to make a choice about which of two artificial entities it would be more morally wrong to harm. These participants may have responded in a systematically different way given that they do not accept an underlying premise of the question. We therefore conducted a further sensitivity check, comparing the AMCEs of participants who do not think it can be morally wrong to harm an artificial entity at all with those who do. We similarly ran regressions of participant choice on the features, an indicator variable for whether they think it can be morally wrong to harm artificial beings, and the interaction of the two. We found evidence of a difference in the AMCEs for only the intelligence feature, $F(1, 18690) = 4.24, p = .040$. Inspection of the regression coefficients showed that there was a marginally significant main effect of thinking it can be morally wrong ($b = -0.03$, $SE = 0.2, p = 0.052$), and a significant interaction between this variable and the intelligence feature ($b = 0.07$, $SE = 0.03, p = 0.040$). Note that this effect became marginally nonsignificant with a multiple comparison correction.

We conclude that our estimates are largely robust to the key assumptions of conjoint experiments, and that participants who do not think it can be morally wrong to harm an artificial being do not differ systematically from participants who do. However, the complexity and intelligence features should be interpreted with the above caveats in mind.

4 Discussion

The present study conducted a conjoint experiment to estimate the effects of 11 features on the moral consideration of artificial entities. This design allowed us to estimate the relative effects of the features, as well as to better isolate the estimates of the effect of each feature than would be possible in a typical experiment, by providing information about a range of features simultaneously. The study has particular implications for the mind perception (H. M. Gray et al., 2007; K. Gray et al., 2012) and humanness accounts (Epley et al., 2007; Haslam, 2006), which emphasize different features as most important for moral consideration.

As hypothesized, all 11 of the features in our study were predictive of participants' judgments about the moral wrongness of harming an artificial entity. These results support several existing studies that have found positive effects of the features included in our study: an entity's body (Küster et al., 2020; Riek et al., 2009), emotion expression (Lee et al., 2019; Nijssen et al., 2019), autonomy (Chernyak & Gary, 2016; Lima et al., 2020), damage avoidance (Tanibe et al., 2017; Ward et al., 2013), intelligence (Bartneck et al., 2007), and purpose (Wang & Krumhuber, 2018). Importantly, because the present study better isolates the effects of these features than some earlier studies, it provides greater certainty that they are predictive of moral consideration in themselves rather than due to their effects via other features. The present study also adds to the literature by providing evidence of the importance of several features that have received less attention: complexity, cooperation, emotion recognition, human language capacities, and moral judgment.

We conducted pairwise comparisons to understand the relative effects of the features on moral consideration. This suggested that there were three categories of effect size. In the first category, with the strongest effects, were an artificial entity's capacity for moral judgment and emotion expression; in the second category were emotion recognition, cooperation, and an entity's physical body; and in the third category, with the weakest effect, were autonomy, complexity, damage avoidance, language, and purpose. While intelligence also had a positive effect, we did not include it in this categorization since it was on a different scale to the other features, as described in the methodology section.

Table 2. Relative effects and implications for mind perception and humanness accounts

Feature	Relative effect size (category)	Effect Size Supports Mind Perception, Humanness, Both, or Neither
Autonomy	3	Both
Body	2	Humanness
Complexity	3	Both
Cooperation	2	Humanness
Damage avoidance	3	Neither
Emotion expression	1	Both
Emotion recognition	2	Humanness
Language	3	Both
Moral judgment	1	Humanness
Purpose	3	Humanness

Note. Relative effect size categories based on pairwise comparisons from section 3.2.2 where category 1 = strongest effect, 2 = middle effect, and 3 = weakest effect.

Table 2 summarizes the relative effect sizes of the features and whether these effects support mind perception, humanness, both accounts, or neither account. First, consider the two features in the highest category: emotion expression and moral judgment. The relatively strong effect of emotion expression supports the mind perception account, because emotion expression is indicative of an entity's experience, and the experience dimension of mind perception should be relatively strongly associated with moral consideration (H. M. Gray et al., 2007; K. Gray et al., 2012). It also supports the humanness account, because emotion is an aspect of human nature, which should also be relatively strongly associated with moral consideration (Bastian et al., 2011; Haslam, 2012). However, the relatively strong effect of moral judgment is more consistent with the humanness account. The mind perception account categorizes moral judgment as an aspect of agency (H. M. Gray et al., 2007). It should, therefore, be relatively weakly (or even negatively) associated with moral consideration. While there is some uncertainty about how best to categorize moral judgment on the humanness account, with Haslam (2006) categorizing it as a higher-order, uniquely human capacity, it is also indicative of proactive agency and therefore an aspect of human nature (Haslam et al., 2012). Given this latter conceptualization of this feature, the finding that it has a relatively strong effect on moral consideration supports the humanness account.

In the second category of features were an entity's body (in particular, the extent to which its body is human-like), cooperation, and emotion recognition. First, consider cooperation and emotion recognition. H. M. Gray et al. (2007) categorized emotion recognition as an aspect of agency. They did not explicitly categorize cooperation, but to the extent that it pertains to the capacity to act rather than experience the world, it more plausibly falls under agency than experience. According to the mind perception account, then, both features should be weakly or negatively associated with moral consideration. On the humanness account, both features are plausibly aspects of proactive agency, and therefore fall more naturally into the human nature dimension (Bastian et al., 2011; Haslam, 2012). They should, therefore, be relatively strongly associated with moral consideration. We therefore interpret their relatively strong effects—in the top half of the effect sizes, and significantly larger than the bottom half of the effect sizes—as support for the humanness account.

Next, consider an entity having a human-like body. This feature has been found to be associated with mind perception (Abubshait & Wiese, 2017; Ferrari et al., 2016; K. Gray & Wegner, 2012) and so would likely in turn be associated with increased moral consideration. However, in the present study we included features indicative of mental states alongside an entity's body, and so we can be more confident that the effect of this feature is important in itself, rather than because of indirect effects via mind perception or other features. This finding fits better with the humanness account, since having a physical body and human-like appearance are important aspects of being human (Epley et al., 2007), but they are not mental states. The fact that this feature is important even after accounting for the other features also has an important practical implication: future artificial entities may be granted reduced moral consideration on the basis of their appearance alone rather than other arguably more relevant features, such as their mental states and behavior (e.g., Gibert & Martin, 2021; Mosakas, 2021).

In the third category of features, with the weakest relative effects, were damage avoidance, purpose, autonomy, complexity, and human language. First, consider an entity's purpose. We found that artificial entities whose purpose is entertainment, working for a business, or being the subject of science experiments, were granted less moral consideration than artificial entities whose purpose is social companionship. We consider this finding supports the humanness account using similar reasoning as for the body feature: while purpose has previously been found to be associated with moral consideration for artificial entities via mind perception (Wang & Krumhuber, 2018), the conjoint design provides us with more confidence that it has an effect in itself. Since behaving socially is plausibly an aspect of being human but is not a mental state, this finding fits better with the humanness account than the mind perception account.

While we found a positive effect of damage avoidance on moral consideration, we expected it to have a relatively strong effect, because it implies that an entity can have negative sensory experiences, meaning that according to both accounts it should have a relatively strong effect (H. M. Gray et al., 2007; Morris et al., 2018). The relatively weak effect we found is not entirely consistent with either account. One explanation for this relatively weak effect is the language used in the study—as noted in Section 1.4, we focused on observable functions and behaviors rather than positing actual internal mental states. This may have resulted in respondents not interpreting this feature as reflecting negative sensory experience. Further research is needed to confirm and understand the reasons for this relatively weak effect.

Finally, we can consider the last three features in the third category: autonomy, complexity, and human language. These features plausibly fall under the agency dimension of mind perception and the human uniqueness dimension of humanness, and therefore should have relatively weak effects on moral consideration. The finding that they are all in the lowest category, therefore, supports both accounts. Some studies have suggested that agency and human uniqueness may have negative effects on moral consideration if, for example, they result in an entity being typecast as an agent rather than a patient (Bastian et al., 2011; K. Gray & Wegner, 2009). Others have found evidence of positive effects of these higher-order capacities on moral consideration (Sytsma & Machery, 2012). The present study supports the latter view—while the effects were relatively weak, all of these features still positively affected moral consideration. Intelligence, which is also an aspect of agency and human uniqueness, also had a positive effect on moral consideration, further supporting that these capacities have positive effects. Sytsma and Machery (2012) found evidence that perceived agency plays a role in promoting moral consideration particularly in more abstract contexts where empathy is less engaged. This is arguably the case for the experimental design used in this paper and may explain the positive effects of these features.

Overall, the analysis of the relative effects favors a humanness account of moral consideration of artificial entities. Several of the findings support both accounts, indicating that mental states are also an important component of moral consideration. According to our analysis, mental states matter, but so do other characteristics, such as physical appearance and behavior, particularly prosocial behaviors such as cooperation. Within the humanness account, our analysis also suggests that the human nature and the human uniqueness dimensions both matter for moral consideration, but the human nature dimension is relatively more important.

Where possible we included more than two levels for the features, allowing us to test whether there are nonlinear effects, which we considered may arise in the context of the present study due to perceived threat or the uncanny valley (K. Gray & Wegner, 2012; Złotowski, Yogeeswaran, et al., 2017). We found no systematic evidence of nonlinear effects—the effect sizes for the highest category of the features were generally double the effect sizes for the middle categories. One possible explanation for this lack of effect is that the relatively abstract design of conjoint experiments reduced the saliency of the features that would usually cause such effects.

We also explored interaction effects between the features. We found no evidence of systematic first-order interactions between any of the features. For example, we found little evidence that any of the features are more or less important for entities without a physical body compared to those with physical bodies. Thus, the evidence presented in this paper suggests that while entities with a combination of a human-like body and other features are granted more moral consideration, this does not extend beyond the main effects of the individual features. There were interaction effects between demographics of participants and feature importance, such as that relative to men, women were more responsive to emotion expression, emotion recognition, language, moral judgment, and purpose, and that relative to conservatives, liberals and moderates were less responsive to an entity's physical body and were more responsive to an entity's capacity for damage avoidance. These findings suggest there may be demographic differences in the way moral consideration is extended to artificial entities, and should be explored further in future studies.

4.1 Limitations and Future Directions

Our study has several important limitations. First, by testing a large number of features simultaneously, conjoint experiments can be cognitively challenging for participants (Bansak et al., 2021b). The present study had an additional layer of complexity by not only testing a relatively large number of features, but doing so in relation to entities that do not currently exist today. However, the results of cognitive checks indicated that participants had good comprehension of the tasks and did not find them too difficult, and our sensitivity analysis found that responses were stable throughout the tasks.

Second, conjoint experiments derive people's hypothetical preferences, rather than their actual real-world choices and behaviors. For example, the estimated effects of autonomy tell us how people value autonomy in principle, rather than how they would actually behave in the presence of an autonomous entity. While this is important to be aware of, it is not necessarily a drawback—many of society's decisions, such as the creation or modification of laws and regulations that affect artificial entities, may be made on the basis of such principled considerations.

Third, while we considered a large number of features, we do not consider them to be an exhaustive account of all the features affecting the moral consideration of artificial entities. Each feature we measured had a relatively large effect size—the smallest effect being around 8%. There are plausibly other features—indeed, there are more in the literature, and our pretesting study found several others to have some degree of importance—that may play a role. Nevertheless, we expect that the features considered in this paper are likely to be the most important ones.

Finally, it is worth considering the interpretation of the levels for the features in the present study, many of which took on Likert levels ranging from "Not at all" to "To a great extent." While this has the benefit of making it easier to compare effects across the features, it doesn't give a precise indication of the capacities of the entities described and could therefore be interpreted in different ways. Consider the highest level for many of the features, which took on the value "To a great extent." It is plausible that people interpreted this as being roughly the level of humans, particularly given that several of the features were designed with this reference point in mind (such as the language feature). However, future artificial entities may surpass humans in a variety of ways, and this has been found to be associated with greater perception of threat (Yogeeswaran et al., 2016). Future research should look at the effects of providing more precise information about the capacities of artificial entities, including those with greater than human-level capacities.

In addition to the directions outlined in the preceding paragraphs, there are several other questions for future research. It will be particularly valuable to further test the features that this study found to be important but that had not been a major part of previous studies, particularly the proactive agency features such as emotion recognition and cooperation. Future research can also explore the mediating paths through which the features have their effects. While we conceptualized and interpreted the present findings in the context of the mind perception and humanness accounts, future studies can test for these and other possible paths more explicitly. It will also be important to explore whether there are factors that influence the importance of different variables. For example, would moral reflection reduce the extent to which people place importance on having a human-like body or a social purpose? Answering these questions would help develop a more complete account of the moral consideration of artificial entities.

4.2 Conclusion

Researchers across a range of disciplines are now studying the question of how society will and ought to extend moral consideration to intelligent artificial entities. The current study attempted to bring together and build on a wide body of existing research to understand the relative importance of various features of moral consideration for artificial entities. We found that the most important features were an entity's capacity for emotion expression and moral judgment, followed by emotion recognition, cooperation, and an entity's physical body. The remaining features were less important but still had positive effects. Our study supports a humanness account of moral consideration, where the extent to which intelligent artificial entities are granted moral consideration depends on how human-like they are in their mental, physical, and behavioral capacities.

References

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

Anthis, J. R., & Paez, E. (2021). Moral circle expansion: A promising strategy to impact the far future. *Futures*, *130*, 102756. https://doi.org/10.1016/j.futures.2021.102756

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

Bansak, K., Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2018). The Number of Choice Tasks and Survey Satisficing in Conjoint Experiments. *Political Analysis*, *26*(1), 112–119. https://doi.org/10.1017/pan.2017.40

Bansak, K., Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2021a). Conjoint Survey Experiments. In J. N. Druckman & D. P. Green, *Cambridge Handbook of Advances in Experimental Political Science* (pp. 19–41). Cambridge University Press.

Bansak, K., Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2021b). Beyond the breaking point? Survey satisficing in conjoint experiments. *Political Science Research and Methods*, *9*(1), 53–71. https://doi.org/10.1017/psrm.2019.13

Bartneck, C., van der Hoek, M., Mubin, O., & Al Mahmud, A. (2007). "Daisy, daisy, give me your answer do!" switching off a robot. *2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 217–222.

Bastian, B., Denson, T. F., & Haslam, N. (2013). The Roles of Dehumanization and Moral Outrage in Retributive Justice. *PLOS ONE*, *8*(4), e61842. https://doi.org/10.1371/journal.pone.0061842

Bastian, B., Laham, S. M., Wilson, S., Haslam, N., & Koval, P. (2011). Blaming, praising, and protecting our humanity: The implications of everyday dehumanization for judgments of moral status. *British Journal of Social Psychology*, *50*(3), 469–483. https://doi.org/10.1348/014466610X521383

Beer, J. M., Fisk, A. D., & Rogers, W. A. (2014). Toward a framework for levels of robot autonomy in human-robot interaction. *Journal of Human-Robot Interaction*, *3*(2), 74–99. https://doi.org/10.5898/JHRI.3.2.Beer

Benjamini, Y., & Hochberg, Y. (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), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Bostrom, N., & Yudkowsky, E. (2014). The Ethics of Artificial Intelligence. In *The Cambridge Handbook of Artificial Intelligence* (pp. 316–334). Cambridge University Press.

Caviola, L., Everett, J. A. C., & Faber, N. S. (2019). The moral standing of animals: Towards a psychology of speciesism. *Journal of Personality and Social Psychology*, *116*(6), 1011–1029. https://doi.org/10.1037/pspp0000182

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

Clogg, C. C., Petkova, E., & Haritou, A. (1995). Statistical Methods for Comparing Regression Coefficients Between Models. *American Journal of Sociology*, *100*(5), 1261–1293. https://doi.org/10.1086/230638

Coeckelbergh, M. (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 and Machines*, *31*(3), 337–360. https://doi.org/10.1007/s11023-020-09554-3

Crimston, C. R., Bain, P. G., Hornsey, M. J., & Bastian, B. (2016). Moral expansiveness: Examining variability in the extension of the moral world. *Journal of Personality and Social Psychology*, *111*(4), 636–653. https://doi.org/10.1037/pspp0000086

Cuddy, A. J. C., Rock, M. S., & Norton, M. I. (2007). Aid in the Aftermath of Hurricane Katrina: Inferences of Secondary Emotions and Intergroup Helping. *Group Processes & Intergroup Relations*, *10*(1), 107–118. https://doi.org/10.1177/1368430207071344

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

Darling, K. (2016). Extending legal protection to social robots: The effects of anthropomorphism, empathy, and violent behavior towards robotic objects. *Robot Law*. https://www.elgaronline.com/view/edcoll/9781783476725/9781783476725.00017.xml de Graaf, M. M. A., Hindriks, F. A., & Hindriks, K. V. (2021). Who Wants to Grant Robots Rights? *Companion of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*, 38–46. https://doi.org/10.1145/3434074.3446911

de Melo, C. M., Gratch, J., & Carnevale, P. J. (2015). Humans versus Computers: Impact of Emotion Expressions on People's Decision Making. *IEEE Transactions on Affective Computing*, *6*(2), 127–136. https://doi.org/10.1109/TAFFC.2014.2332471

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349. https://doi.org/10.1037/xap0000092

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864–886. https://doi.org/10.1037/0033-295X.114.4.864

Eyssel, F., de Ruiter, L., Kuchenbrandt, D., Bobinger, S., & Hegel, F. (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. *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 125–126. https://doi.org/10.1145/2157689.2157717

Eyssel, F., Hegel, F., Horstmann, G., & Wagner, C. (2010). Anthropomorphic inferences from emotional nonverbal cues: A case study. *19th International Symposium in Robot and Human Interactive Communication*, 646–651. https://doi.org/10.1109/ROMAN.2010.5598687

Ferrari, F., Paladino, M. P., & Jetten, J. (2016). Blurring Human–Machine Distinctions: Anthropomorphic Appearance in Social Robots as a Threat to Human Distinctiveness. *International Journal of Social Robotics*, *8*(2), 287–302. https://doi.org/10.1007/s12369-016-0338-y

Fink, J. (2012). Anthropomorphism and Human Likeness in the Design of Robots and Human-Robot Interaction. In S. S. Ge, O. Khatib, J.-J. Cabibihan, R. Simmons, & M.-A. Williams

(Eds.), *Social Robotics* (pp. 199–208). Springer. https://doi.org/10.1007/978-3-642-34103-8_20

Flanagan, T., Rottman, J., & Howard, L. H. (2021). Constrained Choice: Children's and Adults' Attribution of Choice to a Humanoid Robot. *Cognitive Science*, *45*(10), e13043. https://doi.org/10.1111/cogs.13043

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

Freitag, M. (2021). *A Priori Power Analyses for Conjoint Experiments* [HTML]. https://github.com/m-freitag/cjpowR (Original work published 2020)

Gellers, J. C. (2020). *Rights for Robots: Artificial Intelligence, Animal and Environmental Law*. Routledge. https://doi.org/10.4324/9780429288159

Gibert, M., & Martin, D. (2021). In search of the moral status of AI: Why sentience is a strong argument. *AI & SOCIETY*. https://doi.org/10.1007/s00146-021-01179-z

Goff, P. A., Eberhardt, J. L., Williams, M. J., & Jackson, M. C. (2008). Not yet human: Implicit knowledge, historical dehumanization, and contemporary consequences. *Journal of Personality and Social Psychology*, *94*(2), 292–306. https://doi.org/10.1037/0022-3514.94.2.292

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

Gray, K., & Wegner, D. M. (2009). Moral typecasting: Divergent perceptions of moral agents and moral patients. *Journal of Personality and Social Psychology*, *96*(3), 505–520. https://doi.org/10.1037/a0013748

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

Gray, K., Young, L., & Waytz, A. (2012). Mind Perception Is the Essence of Morality. *Psychological Inquiry*, *23*(2), 101–124. https://doi.org/10.1080/1047840X.2012.651387

Gunkel, D. J. (2018). *Robot Rights*. MIT Press.

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis*, *22*(1), 1–30. https://doi.org/10.1093/pan/mpt024

Harris, J., & Anthis, J. R. (2021). The Moral Consideration of Artificial Entities: A Literature Review. *Science and Engineering Ethics*, *27*(4), 53. https://doi.org/10.1007/s11948-021-00331-8

Haslam, N. (2006). Dehumanization: An Integrative Review. *Personality and Social Psychology Review*, *10*(3), 252–264. https://doi.org/10.1207/s15327957pspr1003_4

Haslam, N. (2012). Morality, Mind, and Humanness. *Psychological Inquiry*, *23*(2), 172–174. https://doi.org/10.1080/1047840X.2012.655236

Haslam, N., Bain, P., Douge, L., Lee, M., & Bastian, B. (2005). More human than you: Attributing humanness to self and others. *Journal of Personality and Social Psychology*, *89*(6), 937–950. https://doi.org/10.1037/0022-3514.89.6.937

Haslam, N., Bastian, B., Laham, S., & Loughnan, S. (2012). Humanness, dehumanization, and moral psychology. In *The social psychology of morality: Exploring the causes of good and evil* (pp. 203–218). American Psychological Association. https://doi.org/10.1037/13091-011

Haslam, N., & Loughnan, S. (2014). Dehumanization and Infrahumanization. *Annual Review of Psychology*, *65*(1), 399–423. https://doi.org/10.1146/annurev-psych-010213-115045

Kahn, P. H. J., Ishiguro, H., Friedman, B., Kanda, T., Freier, N. G., Severson, R. L., & Miller, J. (2007). What is a Human?: Toward psychological benchmarks in the field of human–robot interaction. *Interaction Studies*, *8*(3), 363–390. https://doi.org/10.1075/is.8.3.04kah

Khamitov, M., Rotman, J. D., & Piazza, J. (2016). Perceiving the agency of harmful agents: A test of dehumanization versus moral typecasting accounts. *Cognition*, *146*, 33–47. https://doi.org/10.1016/j.cognition.2015.09.009

Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic Interactions with a Robot and Robot–like Agent. *Social Cognition*, *26*(2), 169–181. https://doi.org/10.1521/soco.2008.26.2.169

Kodapanakkal, R. I., Brandt, M. J., Kogler, C., & van Beest, I. (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*, 106303. https://doi.org/10.1016/j.chb.2020.106303

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

Lee, M., Lucas, G., Mell, J., Johnson, E., & Gratch, J. (2019). What's on Your Virtual Mind? Mind Perception in Human-Agent Negotiations. *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, 38–45. https://doi.org/10.1145/3308532.3329465

Leeper, T. J., Hobolt, S. B., & Tilley, J. (2020). Measuring Subgroup Preferences in Conjoint Experiments. *Political Analysis*, *28*(2), 207–221. https://doi.org/10.1017/pan.2019.30

Legg, S., & Hutter, M. (2007). A Collection of Definitions of Intelligence. *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms*, 17–24.

Leidner, B., Castano, E., Zaiser, E., & Giner-Sorolla, R. (2010). Ingroup Glorification, Moral Disengagement, and Justice in the Context of Collective Violence. *Personality and Social Psychology Bulletin*, *36*(8), 1115–1129. https://doi.org/10.1177/0146167210376391

Li, M., Leidner, B., & Castano, E. (2014). Toward a comprehensive taxonomy of dehumanization: Integrating two senses of humanness, mind perception theory, and stereotype content model. *TPM-Testing, Psychometrics, Methodology in Applied Psychology*, *21*(3), 285–300.

Lima, G., Kim, C., Ryu, S., Jeon, C., & Cha, M. (2020). Collecting the Public Perception of AI and Robot Rights. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW2), 135:1-135:24. https://doi.org/10.1145/3415206

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

Morris, K. L., Goldenberg, J., & Boyd, P. (2018). Women as Animals, Women as Objects: Evidence for Two Forms of Objectification. *Personality and Social Psychology Bulletin*, *44*(9), 1302–1314. https://doi.org/10.1177/0146167218765739

Mosakas, K. (2021). On the moral status of social robots: Considering the consciousness criterion. *AI & SOCIETY*, *36*(2), 429–443. https://doi.org/10.1007/s00146-020-01002-1

Nijssen, S. R. R., Müller, B. C. N., Baaren, R. B. van, & Paulus, M. (2019). Saving the Robot or the Human? Robots Who Feel Deserve Moral Care. *Social Cognition*, *37*(1), 41-S2. https://doi.org/10.1521/soco.2019.37.1.41

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

Paternoster, R., Brame, R., Mazerolle, P., & Piquero, A. (1998). Using the Correct Statistical Test for the Equality of Regression Coefficients. *Criminology*, *36*(4), 859–866. https://doi.org/10.1111/j.1745-9125.1998.tb01268.x

Piazza, J., Landy, J. F., & Goodwin, G. P. (2014). Cruel nature: Harmfulness as an important, overlooked dimension in judgments of moral standing. *Cognition*, *131*(1), 108–124. https://doi.org/10.1016/j.cognition.2013.12.013

Powers, A., Kiesler, S., Fussell, S., & Torrey, C. (2007). Comparing a computer agent with a humanoid robot. *Proceedings of the ACM/IEEE International Conference on Human-Robot Interaction*, 145–152. https://doi.org/10.1145/1228716.1228736

Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places* (pp. xiv, 305). Cambridge University Press.

Riek, L. D., Rabinowitch, T.-C., Chakrabarti, B., & Robinson, P. (2009). Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 1–6. https://doi.org/10.1109/ACII.2009.5349423

Rosenthal-von der Pütten, A. M., Krämer, N. C., Hoffmann, L., Sobieraj, S., & Eimler, S. C. (2013). An Experimental Study on Emotional Reactions Towards a Robot. *International Journal of Social Robotics*, *5*(1), 17–34. https://doi.org/10.1007/s12369-012-0173-8

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

Schuessler, J., & Freitag, M. (2020). *Power Analysis for Conjoint Experiments*. SocArXiv. https://doi.org/10.31235/osf.io/9yuhp

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

Suzuki, Y., Galli, L., Ikeda, A., Itakura, S., & Kitazaki, M. (2015). Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports*, *5*(1), 15924. https://doi.org/10.1038/srep15924

Swiderska, A., & Küster, D. (2020). Robots as Malevolent Moral Agents: Harmful Behavior Results in Dehumanization, Not Anthropomorphism. *Cognitive Science*, *44*(7), e12872. https://doi.org/10.1111/cogs.12872

Sytsma, J., & Machery, E. (2012). The Two Sources of Moral Standing. *Review of Philosophy and Psychology*, *3*(3), 303–324. https://doi.org/10.1007/s13164-012-0102-7

Tanibe, T., Hashimoto, T., & Karasawa, K. (2017). We perceive a mind in a robot when we help it. *PLOS ONE*, *12*(7), e0180952. https://doi.org/10.1371/journal.pone.0180952

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

Tsiourti, C., Weiss, A., Wac, K., & Vincze, M. (2019). Multimodal Integration of Emotional Signals from Voice, Body, and Context: Effects of (In)Congruence on Emotion Recognition and Attitudes Towards Robots. *International Journal of Social Robotics*, *11*(4), 555–573. https://doi.org/10.1007/s12369-019-00524-z

Vanman, E. J., & Kappas, A. (2019). "Danger, Will Robinson!" The challenges of social robots for intergroup relations. *Social and Personality Psychology Compass*, *13*(8), e12489. https://doi.org/10.1111/spc3.12489

Wang, X., & Krumhuber, E. G. (2018). Mind Perception of Robots Varies With Their Economic Versus Social Function. *Frontiers in Psychology*, *9*. https://doi.org/10.3389/fpsyg.2018.01230

Ward, A. F., Olsen, A. S., & Wegner, D. M. (2013). The Harm-Made Mind: Observing Victimization Augments Attribution of Minds to Vegetative Patients, Robots, and the Dead. *Psychological Science*, *24*(8), 1437–1445. https://doi.org/10.1177/0956797612472343

Waytz, A., Cacioppo, J., & Epley, N. (2010). Who Sees Human?: The Stability and Importance of Individual Differences in Anthropomorphism. *Perspectives on Psychological Science*, *5*(3), 219–232. https://doi.org/10.1177/1745691610369336

Waytz, A., Epley, N., & Cacioppo, J. T. (2010). Social Cognition Unbound: Insights Into Anthropomorphism and Dehumanization. *Current Directions in Psychological Science*, *19*(1), 58–62. https://doi.org/10.1177/0963721409359302

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of Experimental Social Psychology*, *52*, 113–117. https://doi.org/10.1016/j.jesp.2014.01.005

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

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

Złotowski, J., Strasser, E., & Bartneck, C. (2014). Dimensions of Anthropomorphism: From Humanness to Humanlikeness. *2014 9th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 66–73.

Złotowski, J., Sumioka, H., Bartneck, C., Nishio, S., & Ishiguro, H. (2017). Understanding Anthropomorphism: Anthropomorphism is not a Reverse Process of Dehumanization. In A. Kheddar, E. Yoshida, S. S. Ge, K. Suzuki, J.-J. Cabibihan, F. Eyssel, & H. He (Eds.), *Social Robotics* (pp. 618–627). Springer International Publishing. https://doi.org/10.1007/978-3-319-70022-9_61

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