It matters to me if you are human - Examining categorical perception in human and nonhuman agents

Eva Wiese[*] & Patrick P. Weis[*]

George Mason University, USA

[*] both authors contributed equally to the manuscript

Corresponding author:
Patrick Weis
Email: pweis@gmu.edu
Address: 4400 University Drive, Fairfax, VA 22030

The final version of this manuscript has been published in the International Journal of Human-Computer Studies at https://doi.org/10.1016/j.ijhcs.2019.08.002

ABSTRACT

Humanlike but not perfectly human agents frequently evoke feelings of eeriness, a phenomenon termed the *Uncanny Valley* (UV). The *Categorical Perception Hypothesis* proposes that effects associated with the UV are due to uncertainty as to whether to categorize agents falling into the valley as “human” or “nonhuman”. However, since UV studies have traditionally looked at agents of varying human-likeness, it remains unclear whether UV-related effects are due to categorical uncertainty in general or are specifically evoked by categorizations that require decisions regarding an agent’s human-likeness. Here, we used mouse tracking to determine whether agent spectra with (i.e., robot-human) and without (i.e., robot-animal and robot-stuffed animal) a human endpoint cause phenomena related to categorical perception to comparable extents. Specifically, we compared human and nonhuman agent spectra with respect to existence and location of a category boundary (*H1-1* and *H2-1*), as well as the magnitude of cognitive conflict around the boundary (*H1-2* and *H2-2*). The results show that human and nonhuman spectra exhibit category boundaries (*H1-1*) at which cognitive conflict is higher than for less ambiguous parts of the spectra (*H1-2*). However, in human agent spectra cognitive conflict maxima were more pronounced than for nonhuman agent spectra (*H2-1*) and category boundaries were shifted towards the human endpoint of the spectrum (*H2-2*). Overall, these results suggest a quantitatively, though not qualitatively, different categorization process for spectra containing human endpoints. Possible reasons and the impact for virtual and robotic agent design are discussed.

Keywords: Uncanny valley; Human–robot interaction; Categorical perception; Cognitive conflict

1 INTRODUCTION

Very human-like, though not perfectly human, robot entities are frequently perceived more negatively than agents that are unambiguously human or nonhuman, a phenomenon termed the *Uncanny Valley* (UV). For example, a study on human-computer interaction (HCI) has shown that embodied avatars are found to be more uncanny and to more strongly evoke negative emotions than their text-based counterparts (Ciechanowski, Przegalinska, Magnuski, & Gloor, 2018). Being exposed to “uncanny” agents that are neither clearly human nor nonhuman is also associated with a depletion of cognitive resources over time and negatively impacts cognitive performance during human-robot interaction (Wiese, Mandell, Shaw, & Smith, 2019). The aim of the current experiment is to examine to what extent phenomena previously associated with the UV, such as increased cognitive conflict processing due to categorical ambiguity, are specific to categorical uncertainty regarding an agent’s human-likeness (i.e., category A vs. human) as opposed to representing general effects associated with categorization processes (i.e., category A vs. any category B).

Although empirical evidence in support of the existence of the UV has been increasing recently (Chattopadhyay & MacDorman, 2016; MacDorman & Chattopadhyay, 2016; Mathur & Reichling, 2016), there is no clear consensus yet regarding its theoretical underpinnings (for a review, see Kätsyri, Förger, Mäkäräinen, & Takala, 2015). Two theories that receive most support in the literature are the *categorical perception hypothesis* and the *perceptual mismatch hypothesis*: the categorical perception hypothesis purports that the physical appearance of humanoid agents triggers a categorization-related cognitive conflict as to whether the agents represent human or nonhuman entities, and that this conflict may result in negative emotional evaluations due to increased cognitive processing costs needed to resolve categorical ambiguities (Cheetham, Suter, & Jäncke, 2011). The perceptual mismatch hypothesis states that negative affinity associated with “uncanny” stimuli would be caused by an inconsistency between the human-likeness levels of specific sensory signals contained in nonhuman images, such as grossly enlarged eyes displayed on an otherwise perfectly human-like face (MacDorman, Green, Ho, & Koch, 2009). The categorical perception hypothesis is in line with insights from evolutionary biology linking categorization to survival and the failure to categorize stimuli to negative emotional responses (Burleigh & Schoenherr, 2015). Psychological research has further demonstrated that category boundaries exist for identification of facial images along morphed spectra (Cheetham, Pavlovic, Jordan, Suter, & Jancke, 2013; Cheetham et al., 2011; Cheetham, Suter, & Jancke, 2014; Looser & Wheatley, 2010; Yamada, Kawabe, & Ihaya, 2013; see **Figure 1** for an illustration of the category boundary concept), and that discrimination performance reaches its peak when agent images straddle the category boundary, indicating categorical ambiguity (Cheetham et al., 2011, 2014; Looser & Wheatley, 2010). Increased categorical ambiguity has been reported to coincide with negative stimulus evaluations in some studies (Burleigh, Schoenherr, & Lacroix, 2013; Ferrey, Burleigh, & Fenske, 2015; Yamada et al., 2013) but not in others (Cheetham et al., 2014; Looser & Wheatley, 2010; MacDorman & Chattopadhyay, 2016). Studies on the perceptual mismatch hypothesis suggest that the most negative affective evaluations are elicited by images where the mismatch between a subset of realistic (e.g., human face shape) and a subset of unrealistic image features (e.g., enlarged eyes) is maximal (MacDorman et al., 2009; Mäkäräinen, Kätsyri, & Takala, 2014; Mitchell et al., 2011; Seyama & Nagayama, 2007), and that maximal negative affinity does not coincide with maximal categorical uncertainty (when manipulating human-likeness within a category from rendered to real; see MacDorman & Chattopadhyay, 2016).

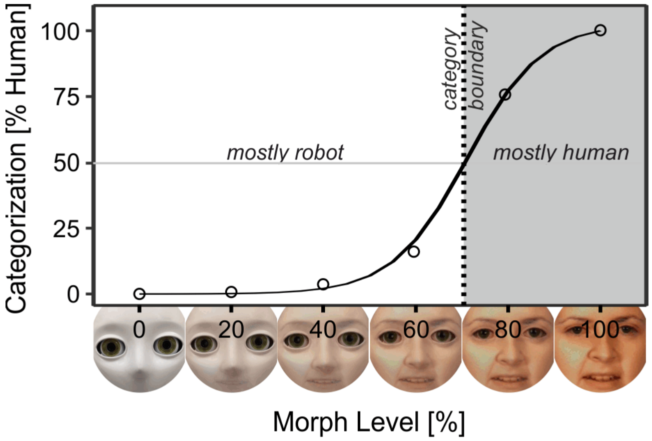

Figure 1. Visualization of the category boundary: A logistic function is fitted to categorization data (i.e., percentage of trials the respective agent was categorized as human; *black line*). The Morph Level at which categorization is most ambiguous, i.e. where the agent was categorized as human in 50% of trials and as nonhuman in 50% of trials, is called the category boundary (here at 70% Morph Level; *dotted line*).

Despite the progress that has been made in recent years in understanding the UV (with good empirical evidence for the perceptual mismatch hypothesis and some evidence for the categorical mismatch hypothesis; see Kätsyri et al., 2015), it remains unclear whether cognitive conflict processing due to categorical uncertainty is specifically related to the perception of human-likeness (i.e., using spectra with a human endpoint) or rather occurs generally for all sorts of categorically ambiguous stimuli (i.e., for spectra without human endpoint), and whether categorical uncertainty and negative affinity coincide when using cross-category spectra (i.e., robot-human, -animal, or -stuffed animal). The aim of the present experiment is to examine whether cognitive conflict processing in response to categorical ambiguity is specific to nonhuman-human judgments or occurs in a similar fashion for nonhuman-nonhuman judgments.

A prominent way to investigate categorical perception is to present morphed images, where a picture of category A (e.g., robot) is morphed into a picture of category B (e.g., human) in percent steps resulting in a sequence of stimuli gradually decreasing in A-likeness and increasing in B-likeness (see **Figure 1**) and ask participants to categorize them as belonging to category A or category B (i.e., forced choice task). Using such a procedure with nonhuman agents as category A (i.e., left end of the spectrum) and human agents as category B (i.e., right end of the spectrum), it was found that categorization follows a qualitative pattern, with substantial changes in categorization decisions only at the nonhuman-human category boundary (i.e., % physical humanness of the image that 50% of people categorize as "human": at around 60-70% physical humanness; see **Figure 1**) but relatively constant categorization decisions to the left and right of the boundary (Cheetham et al., 2011; Hackel, Looser, & Van Bavel, 2014; Looser & Wheatley, 2010; Martini, Gonzalez, & Wiese, 2016; Mathur & Reichling, 2016; Yamada et al., 2013). Although pairs of morphed stimuli straddling the nonhuman-human category boundary were easier to *discriminate* than equally similar pairs of stimuli located on the same side of the boundary (improved performance on same-different judgments; Cheetham et al., 2013), high reaction times indicate that morphed stimuli straddling the nonhuman-human category boundary are difficult to *categorize* (Yamada et al., 2013). In another study researching robotic agents, high reaction times have been associated with maximal negative affective evaluations (Mathur & Reichling, 2016). Negative evaluations of stimuli located at the category boundary have been associated with co-activation of competing categories, which requires additional cognitive resources to process (Ferrey et al., 2015; Meng & Tong, 2004; Sterzer, Kleinschmidt, & Rees, 2009), and negatively impacts performance on tasks that are sensitive to the drainage of cognitive resources over time (Mandell, Smith, & Wiese, 2017; Weis & Wiese, 2017; Wiese et al., 2019).

How strongly ambiguous stimuli co-activate different categories and induce cognitive conflict between multiple categorizations can be measured using *mouse tracking*, a method in which mouse trajectories are recorded during a forced-choice task with labels representing category A and B in the top corners of the computer screen and the to-be-evaluated stimulus at the center bottom (for details, see section 2.2). Previous studies found that the mouse movements' curvatures positively correlate with the degree of cognitive conflict the participants experience during categorization (Freeman & Ambady, 2010), and that negative affective evaluations reach their maximum where categorization is most difficult (Yamada et al., 2013), indicating that negative affective reactions to categorically ambiguous stimuli may be linked to increased cognitive processing effort and decreased cognitive fluency (Winkielman, Schwarz, Fazendeiro, & Reber, 2003).

Although these studies provide evidence that morph spectra containing "human", such as human-robot (Cheetham et al., 2011; Martini et al., 2016; Mathur & Reichling, 2016) or human-doll (Hackel et al., 2014; Looser & Wheatley, 2010) spectra, show a categorical pattern with the maximum of categorization difficulty and the minimum of positive stimulus evaluations coinciding at the category boundary (e.g., Mathur & Reichling, 2016), it is unclear whether this pattern would universally be observed for any kind of categorization or whether it is specific to categorizations that require a "human" versus "nonhuman" categorization. Whether evaluation patterns similar to those observed for nonhuman-human spectra would also be observed for spectra not containing the human category is an important question, as it informs us about whether phenomena related to the uncanny valley are specific to perceptions of human-likeness or generally related to all sorts of categorization processes. The assumption that the UV may be specific to perceptions of human-likeness is in line with several observations emphasizing the special status of human versus nonhuman stimuli in social-cognitive processing: First, being exposed to human agents activates brain areas responsible for social-cognitive processing more strongly than being exposed to nonhuman agents (Looser, Guntupalli, & Wheatley, 2013; Özdem et al., 2016; Wagner, Kelley, & Heatherton, 2011; Wheatley et al., 2011; Wiese, Buzzell, Abubshait, & Beatty, 2018; Wykowska, Wiese, Prosser, & Müller, 2014), and activation in social brain areas is known to reflect the social relevance of observed behaviors and to enable reactions to observed actions that are social in nature (different from those triggered by nonhuman agents; Waytz, Cacioppo, & Epley, 2010; Wiese, Metta, & Wykowska, 2017; for reviews). Second, social categorization is a highly specialized process with different neural networks being involved in the identification of living versus non-living (Forde & Humphreys, 2002), primate versus non-primate (Tovée & Cohen-Tovée, 1993; Young & Yamane, 1992), and human versus animal (Assal, Favre, & Anderes, 1984; McNeil & Warrington, 1993) stimuli, which can potentially affect the extent to which ambiguous stimuli co-activate multiple category representations and trigger categorization conflicts. In line with this assumption, categorizations within the "human" category (e.g., male vs. female; Yamada et al., 2013) induce weaker negative affective evaluations than human versus nonhuman categorizations (e.g., human vs. robot; Mathur & Reichling, 2016). Third, observers seem to be more sensitive to detecting changes in physical features of ingroup versus outgroup facial stimuli (Hugenberg, Young, Bernstein, & Sacco, 2010; Hugenberg, Wilson, See, & Young, 2013), with the consequence that more confirmatory perceptual evidence is needed before a stimulus is categorized as belonging to an observer's ingroup (e.g., human). As a result, the category boundaries are shifted from the center of the spectrum towards the ingroup side of the spectrum (e.g., Hackel et al., 2014; Sigala, Logothetis, & Rainer, 2011), which would match the location of the UV that is typically observed at around 70% human-likeness (for ingroup human observers).

To date, only very few studies have examined UV patterns in morph spectra not containing “human” stimuli (e.g., Campbell, Pascalis, Coleman, Wallace, & Benson, 1997; Ferrey et al., 2015; Steckenfinger & Ghazanfar, 2009; Yamada, Kawabe, & Ihaya, 2012; Yamada et al., 2013). Yamada and colleagues (2013), for instance, used human and dog stimuli varying in their degree of realism from cartoonish to stuffed to real to show that increased categorization difficulty and negative evaluations were observable at transition points from cartoonish to stuffed to real within a given category. Increased categorization difficulty was also noticeable for animal-animal and fruit-fruit morphs (Ferrey et al., 2015; Yamada et al., 2012), as well as when macaque morphs with different degrees of realism were presented to macaque monkeys (Steckenfinger & Ghazanfar, 2009). Although these studies have shown that increased categorization difficulty at the category boundary can be observed for morph spectra not containing “human”, they cannot determine whether spectra requiring human versus nonhuman categorizations (e.g., robot vs. human) differ from nonhuman versus nonhuman categorizations (e.g., robot vs. animal) in terms of the location of the category boundary and the extent of cognitive conflict that categorically ambiguous stimuli induce. To the best of our knowledge, the only study that has compared nonhuman-human and nonhuman-nonhuman spectra has morphed the same nonhuman starting point (i.e., macaque) into nonhuman (i.e., cow) or human endpoints and showed that independent of the specific endpoint, categorizations were most difficult at the category boundary at around 40-60% “category-B-ness” (Campbell et al., 1997). Although this finding suggests that the area of highest categorization difficulty is located around the category boundary, it does not precisely determine the location of spectrum-specific category boundaries and does not assess if categorization difficulty is comparable across spectra or significantly enhanced for spectra containing “human” as endpoint. We will address these questions in the current experiment.

1.1 Aim of Study

As argued in the above, the human category seems to have an exceptional status, both on the neural and the behavioral level, which raises the question to what extent typical findings associated with the UV, such as the rightward shift in the location of the category boundary towards the human end of the spectrum (i.e., 60-70% physical human-likeness) and the observation of increased categorization difficulty for stimuli located at the category boundary are specific to evaluations of a stimulus' humanness rather than a general effect of categorical processing. In the current study, we first investigate whether the assumptions of the categorical perception hypothesis hold for both human and nonhuman agent spectra, i.e. if human and nonhuman agent spectra exhibit a category boundary (*H1-1*) and if cognitive conflict is highest in proximity of that boundary (*H1-2*). Second, we examine whether categorizations of humanlike stimuli differ from categorizations of non-humanlike stimuli due to the special social status of the "human" category. Specifically, we explore whether the nature of a spectrum's endpoint (i.e., category B) affects the location of its category boundary (i.e., right shift for spectra with a human endpoint; *H2-1*), as well as the strength of cognitive conflict that is induced by categorically ambiguous stimuli at the category boundary (i.e., higher cognitive conflict for spectra with a human endpoint; *H2-2*).

2 EXPERIMENTS

To examine these questions, we first acquired photos of human and nonhuman agents to be later on used as endpoints for a morphing procedure. Since the category of nonhuman social agents is quite heterogeneous in terms of features other than humanness, we further differentiated the nonhuman agents into agents that are alive (i.e., animals) and agents that are not alive (i.e., stuffed animals) to be able to separate effects of "humanness" from those of "aliveness" (see Gray, Gray,

& Wegner, 2007; for the importance of animacy). In order to validate how the human and non-human agents (robot, stuffed animal, animal, human) were perceived, we conducted a pilot study in which participants were asked to rate the agents in terms of ”humanness”, “aliveness”, and “similarity-to-self”. For the main experiment, an image morphing procedure was employed to create three spectra with the same starting point (i.e., robot) and three different target agents as end points (nonhuman-nonalive: stuffed animal; nonhuman-alive: animal; human-alive: human): robot-stuffed animal, robot-animal and robot-human. To avoid confounds due to perceptual features specific to single spectra, we created nine spectra for each of the three target agents (see *Stimuli* for details). In the main experiment, a mouse-tracking paradigm was used that required participants to categorize agent images along a spectrum from category A to category B as either belonging to category A (e.g., nonhuman) or category B (e.g., human) by making mouse movements towards a text box representing the respective category on a computer screen (see *Task* for details) while mouse movement curvatures and movement onsets were measured.

We hypothesized that all spectra would exhibit category boundaries (*H1-1*) and that categorization would be most difficult at the spectrum-specific category boundaries (see Hackel et al., 2014), which would be reflected in mouse curvatures being maximal when categorizing stimuli that are located around the category boundary (*H1-2*; see Yamada et al., 2013; for reaction time data). We expected locations of spectrum-specific category boundaries to be modulated by the group status of its endpoint (ingroup versus outgroup), such that the category boundary would be shifted towards that end of the spectrum that participants identify with more (*H2-1*; see Hackel et al., 2014; for rating data). We also hypothesized categorization difficulty to be influenced by the nature of the categorization task, such that categorizations requiring assignments of “own group status” to a stimulus (e.g., human) would cause more uncertainty than categorizations requiring the assessment of “other group” categories (e.g., animal) thus leading to more pronounced mouse curvatures (*H2-2;* in line with Sigala et al., 2011).

2.1 **Pilot Experiment**

The pilot experiment served the purpose of validating the stimuli used in the mouse tracking study. Specifically, the aim was to validate that stimuli depicting human agents were indeed unique in being more “human”, “alive”, and “similar-to-self” than the nonhuman stimuli. For that purpose, we presented all agent images (nine per category: robot, stuffed animal, animal and human) in an online survey and asked participants to rate them in terms of their “humanness”, “aliveness” and “similarity-to-self” on a 7-point Likert scale. The experiment was programmed and hosted on Qualtrics (www.qualtrics.com).

2.1.1 Methods & Materials

2.1.1.1 Participants

77 participants were recruited via Amazon Mechanical Turk (www.mturk.com). One participant was excluded because of an unreasonably large amount of time needed to complete the survey, resulting in a final sample size of 76 participants (42 females, mean age: 32.6, range: 21 – 76). All participants reported normal or corrected-to-normal vision and gave informed consent prior to participating. The study took about 15 minutes to complete, and participants received $ 0.20 for their participation.

2.1.1.2 Stimuli

In total, 36 photographs were presented during the experiment, nine for each of the four agent categories (i.e., robot, human, animal, stuffed animal); see **Figure 2**. Photographs were acquired using the following procedure: first, robot names were obtained from Mathur & Reichling (2016), and the photos were subsequently gathered using a Google image search. Only full-frontal photos depicting robots with human-like faces (i.e., having eyes and nose) were included in the study. Second, after selection of the robot photos, they were matched on apparent gender, head orientation, and facial features with a photo from the MUCT (Milborrow University of Cape Town) human face image database (Milborrow, Morkel, & Nicolls, 2010), with a photo from the Stanford dog database (Khosla, Jayadevaprakash, Yao, & Li, 2011), and with a photo from a Google image search with the term “stuffed animal”. All photos were cropped to a 1:1 aspect ratio and rescaled to 450 x 450 pixels. After rescaling, all backgrounds were removed.

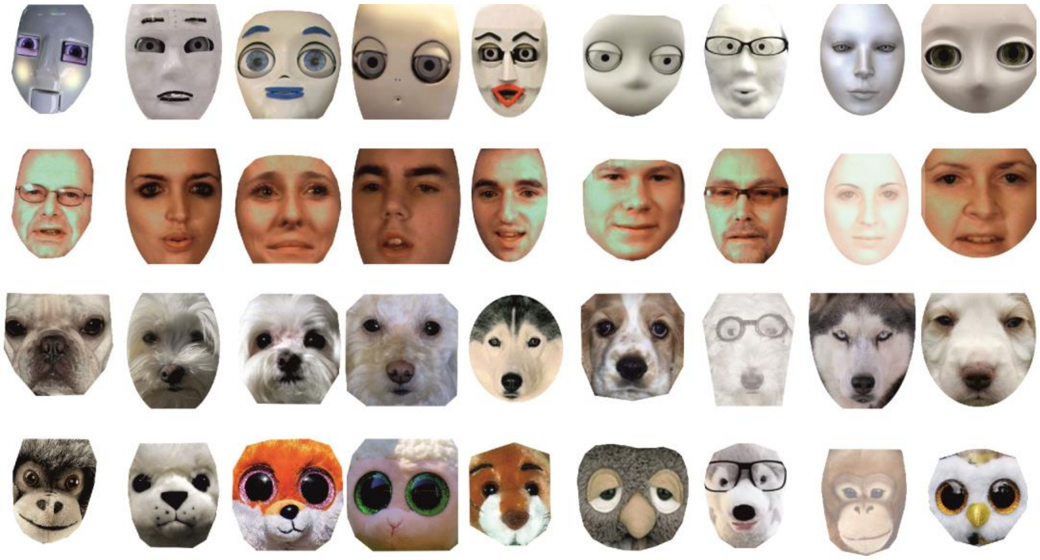

Figure 2. Source Images: These photographs were used as start- and endpoints for the morphing spectra. Each of the human, animal, and stuffed animal agents was morphed into the robot agent on top of the respective column. Spectra created with the transparent images were excluded for the final analysis of the main experiment. For details, see *Stimuli*.

2.1.1.3 Design, Procedure, and Task

The task followed a one-factorial design with the within-participants factor *Agent* (four levels: human, robot, animal, stuffed animal). After starting the experiment online, participants were to read and agree to the consent form and fill out a brief demographic survey. Subsequently, participants were shown the instructions and began the main part of the pilot study. Participants' task was to rate the different agents on humanness ("This agent is human."), aliveness ("This agent is alive."), and similarity-to-self ("This agent is similar to me."). Each trial, the image of one agent and a 7-point Likert scale (Strongly disagree, Disagree, Somewhat disagree, Neither agree nor disagree, Somewhat agree, Agree, Strongly agree) was shown. Only one dimension was tested per trial. Trials were blocked with respect to the different rating dimensions, resulting in 36 trials for each block and 108 trials in total. Block order as well as agent order within blocks was randomized.

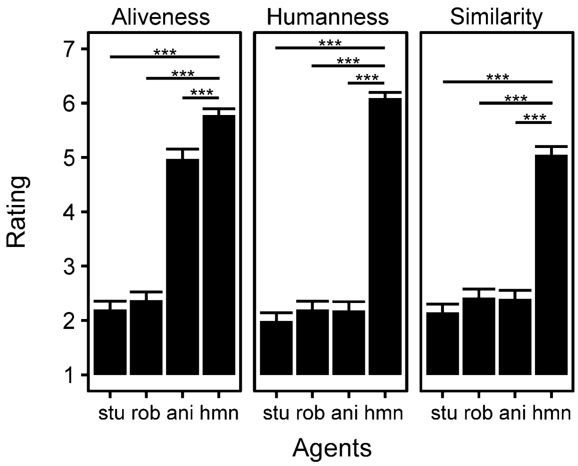

Figure 3. Aliveness, Humanness, and Similarity-to-Self Ratings of Source Images: The human images were perceived as being more alive, more human, and more similar to self than the other images (see *Results* for more details). Ratings were obtained based on a Likert scale ranging from 1 (strongly disagree) to 7 (strongly agree). Error bars depict SEM. stu: stuffed animal, rob: robot, ani: animal, hmn: human; *** : $p < .001$

2.1.2 Results & Discussion

Humanness, aliveness, and similarity-to-self were each analyzed with a one-way ANOVA with the factor Agent (stuffed, robot, animal, human) and followed up with post-hoc paired t-tests.

Aliveness, humanness, and similarity-to-self differed between Agents (all $F(3,\ 225) > 100$, all $p < .001$, all $\eta^2_G > .4$; see **Figure 3**). Human stimuli were perceived as being more alive than the animal ($t(75) = 5.62$, $p < .001$), robot ($t(75) = 15.81$, $p < .001$) and stuffed animal ($t(75) = 16.24$, $p < .001$) stimuli. The human agents were also perceived as being more human and more similar-to-self than the animal, robotic, or stuffed animal agents (all $t(75) > 14$, all $p < .001$).

Results suggest that the images chosen as start and end points for the to-be constructed morph spectra show the desired differentiation of aliveness, humanness, and similarity-to-self ratings between "human" and "nonhuman" agents, and can therefore be used to create the morph spectra for the main experiment. The associated R analysis script and data files can be freely accessed online through the Open Science Framework at https://osf.io/w76eq/.

2.2 Mouse Tracking Experiment

The goal of the mouse tracking experiment was to examine Hypotheses 1 and 2. In particular, the experiment investigated whether the different spectra, irrespective of the target agent, exhibited signs of a categorical boundary (*H1-1*), with maximal cognitive conflict processing around said boundary (*H1-2)* and whether the location of the categorical boundary (*H2-1*) and the magnitude of the cognitive conflict (*H2-2*) around the boundary were altered for the spectra with the human target agent.

2.2.1 Methods & Materials

2.2.1.1 Participants

165 undergraduate students participated in this experiment, and were randomly assigned to one of the three different experimental conditions: robot-human spectrum, robot-animal spectrum or robot-stuffed animal spectrum. Two participants were excluded because their categorization be-

havior could not be fitted with a sigmoid function, resulting in a final sample size of 163 participants (robot-human: 40 females, mean age: 21.2, range: 18 – 29; 47 right handed; robot-animal: 40 females, mean age: 19.7, range: 18 – 39; 51 right handed; robot-stuffed animal: 35 females, mean age: 19.8, range: 18 – 35; 45 right handed). All participants reported normal or corrected to normal vision, had not been diagnosed with a psychological or neurological disorder, and were not taking any medications affecting the central nervous system at the time of data collection. The Ethics Committee at George Mason University approved the experiment, and participants provided informed consent prior to participation.

2.2.1.2 Apparatus

Stimuli were presented at a distance of about 57 cm on an ASUS VB198T-P 19-inch monitor set to a resolution of 1280 × 1024 pixels and a refresh rate of 65 Hz using the Mouse Tracker software (Freeman & Ambady, 2010). Mouse clicks and trajectories from an USB-connected optical mouse were recorded.

2.2.1.3 Stimuli

Pictures along nine different morphing spectra for each target morph condition (human, animal, stuffed animal) were created using the morphing software FantaMorph 5.4.8 (Abrosoft). More than one spectrum for each target agent condition was chosen in order to increase external validity and to minimize artifacts originating from specific source photographs. Along each spectrum, the produced morph images were set apart by 5% morphing steps, resulting in 21 stimuli for each spectrum (see **Figure 4*;*** for examples). Since each target condition consisted of nine spectra, 189 stimuli were created for each target agent condition, resulting in 567 stimuli for the whole study with three target conditions. Each spectrum was based on one photograph of a unique face of the respective target category (human, animal, stuffed animal) and one portrait of a unique robot (see

Figure 2 and *Pilot Experiment* for details about image selection). To ensure comparable fidelity among morph images, high priority was given to smoothly morph eyes, noses, eyebrows, and head shape (requiring at least eight reference points for each feature). All images had a resolution of 450 x 450 pixels. All image backgrounds were removed after morphing.

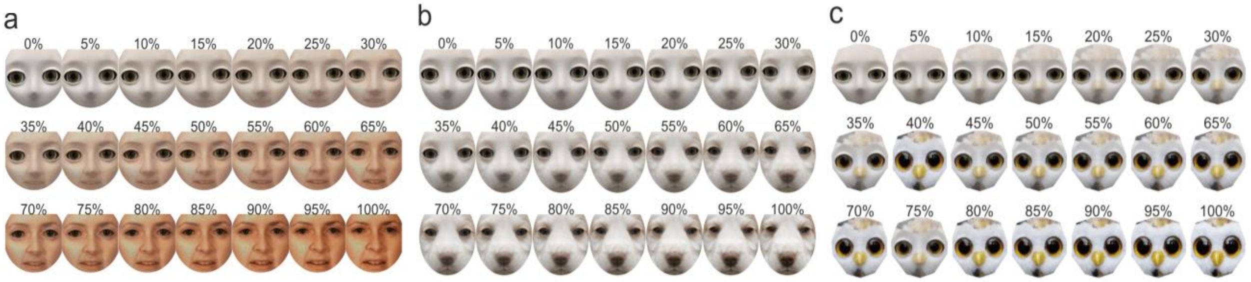

Figure 4. Example Spectra for each Target Agent: A robot image (top left) was morphed into a target image (bottom right) in steps of 5%, resulting in a set of 21 stimuli per spectrum. Target images belonged to one of three Target Agent categories: human (a), animal (b), or stuffed animal (c).

2.2.1.4 Task

Participants were asked to categorize the morph images as belonging either to a given agent category (e.g., human) or not (e.g., non-human). Specifically, participants in the human, animal, and stuffed animal conditions were asked to categorize the images as “human” or “non-human”, “animal” or “non-animal”, or “stuffed animal” or “non-stuffed-animal”, respectively. Morphed images were presented one at a time in the bottom center of the computer screen and the order of their presentation was randomized throughout the experiment. At the beginning of each trial, participants had to click a start button located in the bottom center of the screen to make the image appear. Afterwards, participants were asked to move the mouse cursor from the bottom center of the screen (where the image was placed) to one of two response boxes positioned in the left and right top corners of the screen (depicting the two different categories) to indicate whether the image belonged to a given agent category or not (e.g., “human” versus. “non-human”). During this decision-making process, mouse movement onset times and curvatures were measured; see **Figure 5.** Clicking one of the two response boxes concluded the trial. Between trials, a blank screen was presented for 1000 ms (i.e., inter trial interval, ITI). Response boxes in the top corners of the screen were always shown right from the beginning of the trial; agent images only appeared after the start button was pressed.

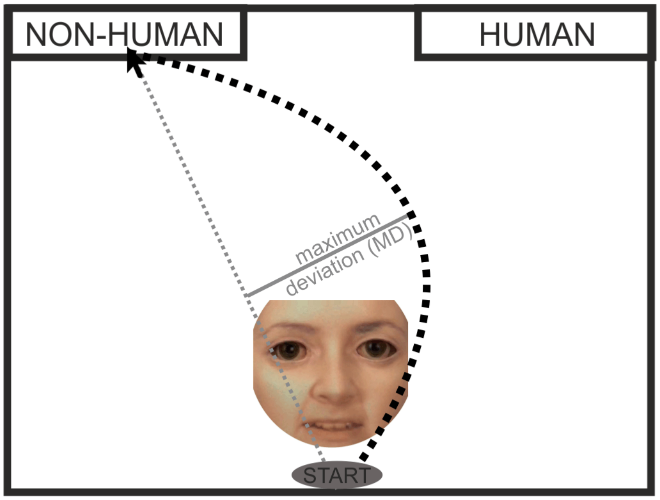

Figure 5. ***Example Trial*****:** After pressing the start button, an agent image appeared on the screen (center, bottom) and participants were to categorize the image as either belonging to the target category (here, human) or not belonging to the target category by moving the mouse cursor to one of the two answer boxes (top left and right, respectively). The dotted black line shows an example mouse trajectory. The dotted gray line represents an ideal trajectory with no measurable cognitive conflict. The solid gray line represents maximum deviation (MD), a measure of cognitive conflict for the black trajectory. Note that for MD calculations, the trajectory is first standardized with respect to time (for details, see Freeman & Ambady, 2010). Between trials, a blank screen was presented for 1000ms.

2.2.1.5 Cognitive Conflict Measurement

Analyzing mouse movements supposedly captures cognitive conflict and co-activation of categories more precisely than reaction times, and can be obtained using the Mouse Tracking software developed by Freeman and Ambady (2010). The software allows for obtaining time-standardized mouse trajectories of individual trials and computing each trajectory's maximum deviation (MD) from a straight line towards the answer box (for an illustration, see **Figure 5**), which is an established measure of cognitive conflict processing in mouse-tracking studies (Freeman & Ambady, 2010). A similar Mouse-Tracking-based indicator of cognitive conflict is area under the curve (AUC; Freeman & Johnson, 2016). For the current paper, since both measures capture the same process, we decided for the MD and against the AUC measure, as it is slightly easier to explain conceptually. To account for the fact that cognitive conflict could also be reflected in participants' hesitation to move the mouse immediately after stimulus presentation (i.e., participants pause to figure out what they are looking at) time-to-first mouse movement after stimulus presentation is also recorded.

2.2.1.6 Design & Procedure

The experiment followed a two-factorial design with the within-participants factor *Morph Level* (ranging from 0% to 100% category B-ness in steps of 5%; e.g., 0 to 100% human) and the between-participants factor *Target Agent* (Human, Animal, Stuffed Animal). All target agents were morphed into the same robot images, resulting in three different morphing spectra (robot-human, robot-animal, robot-stuffed animal) with 21 morphing levels for each spectrum.

At the beginning of the experiment, participants were seated in front of a computer and signed the informed consent form. Participants were then given instructions for the main task and asked to always answer as quickly as possible. This was done to maximize the chance that participants started with the mouse movement immediately after the stimulus was presented (time-to-first mouse movement was measured to control for mouse movement onset time). After participants read the instructions, they were asked to perform three practice trials to familiarize themselves with the mouse-tracking procedure. The stimuli used for the practice were created separately and not drawn from any of the experimental morph spectra. Upon completion of the practice trials, the main experiment began, during which participants categorized 189 agents (9 spectra per target agent group, with 21 morphing levels each). Each image was presented once per participant with the order of the images being randomized across the experiment. The main task took about 15 minutes to complete. After having completed the questionnaire, participants were informed about the purpose of the experiment and received course credit before the session concluded.

2.2.2 Results & Discussion

Trials with extreme categorization times deviating more than 2.5 standard deviations from the individual mean were excluded from analysis, leading to an exclusion of 2.3% of all trials. Also, one spectrum in each condition was excluded because one of the base stimuli was perceived as categorically ambiguous. A spectrum was excluded when, in the grand average, either the 0% morph was categorized as animal, stuffed animal, or human, respectively, in more than 10% of trials, or the 100% morph was categorized as either animal, stuffed animal, or human, respectively, in less than 90% of trials (please see transparent images in **Figure 2**). Effect sizes are reported as generalized eta squared (η_G^2), enabling comparison between between-participants and within-participants designs (Bakeman, 2005). The associated R analysis script and data files can be freely accessed online through the Open Science Framework at https://osf.io/w76eq/.

*2.2.2.1 **Hypothesis 1-1**: All spectra exhibit spectrum-specific category boundaries*

We expected all spectra, irrespective of target agent, to possess category boundaries. To investigate the existence of categorical boundaries, a three-parameter logistic function (see **Equation 1**) was fitted to each participant's individual data (predictor variable: *Morph Level*; response variable: *Proportion* of *category B categorizations*). Parameter L defines the upper asymptote, parameter k the growth rate, and parameter x_0 the predictor level at which the growth rate is the highest. For example, for values of x between 1 and 100, an L of 1, an x_0 of 50, and a k of 0.05, the function returns y values from below 0.1 (for low x) that rise in a non-linear s-shape to values above 0.9 (for high x). A one-sample t-test on growth parameters (i.e., parameter k in **Equation 1**) was used separately for the three target agent conditions (i.e., t-tests for human, animal, and stuffed animal target agents) to test deviation from linearity (see Cheetham et al., 2011; for a comparable procedure). Growth parameters above zero[1] indicate a nonlinear relationship (see Cheetham et al., 2011).

$$f(x) = \frac{\mathrm{L}}{1 + e^{-\mathrm{k}*(\mathrm{x}-\mathrm{x0})}} \tag{1}$$

[1] As a more conservative analysis, we conducted one-sample t-tests testing the sample growth rate against a growth rate of 0.05. We chose this value because the example function with parameters L = 1, k = 0.05, and $x_0 = 50$ already shows a decent deviation from linearity. A value of 0.01 on the other hand still shows a highly linear pattern. All t-tests remained significant with $t > 8.5$ and $p < 0.001$. We also confirmed the results with a bootstrapping method that allows capturing the uncertainty of the *k* estimates, which is occluded in the previous approach. We drew 163 participants with replacement from the participant pool and within each participant eight spectra with replacement from the eight spectra available. For this bootstrapped sample, we fitted the sigmoid function as in the original analytic approach and repeated the procedure 1000 times. The resulting 95% confidence intervals for the k parameter do neither include the liberal criterion of 0 nor the more conservative criterion of 0.05 (CI_{animal}= [.16 .53], CI_{human}= [.24 .84], $\mathrm{CI}_{stuffed\ animal}$= [.12 .35]). Note that the estimated k values are higher when using the bootstrapping in comparison to using the original method because the sampling with replacement frequently leads to the exclusion of one or more spectra which leads to less “smearing” due to averaging across spectra with different PSEs. We thank an anonymous reviewer for suggesting this method.

In general, the logistic function fitted the individual data very well. R^2 for individual fits ranged from 0.704 to 0.997. Mean R^2 values were comparably high for all Target Agent conditions (R^2_{animal}= 0.958, R^2_{human}= 0.966, $R^2_{stuffed\ animal}$= 0.933).

Participants exhibited step-like, in contrast to linear, functions when categorizing stimuli along the robot to human ($t(53) = 14.05$, $p < .001$, $M = 0.21$), robot to animal ($t(54) = 16.88$, $p < .001$, $M = 0.14$), and robot to stuffed animal ($t(53) = 15.66$, $p < .001$, $M = 0.11$) dimensions; see **Figure 6a**. Thus, for all *Target Agents* (human, animal, stuffed animal), the respective spectrum exhibited regions with low categorical uncertainty and, around the category boundary, regions with high categorical uncertainty. As a next step, we averaged across participants within target agent conditions and extracted spectrum-specific category boundaries by predicting the morph level at which 50% of the stimuli are categorized as category A and 50% as category B (i.e., Point of Subjective Equality; PSE): 71% physical human-likeness for the robot-human spectrum, 63% of physical animal-likeness for the robot-animal spectrum, and 37% of physical stuffed animal-likeness for the robot-stuffed animal spectrum (**Figure 6a**).

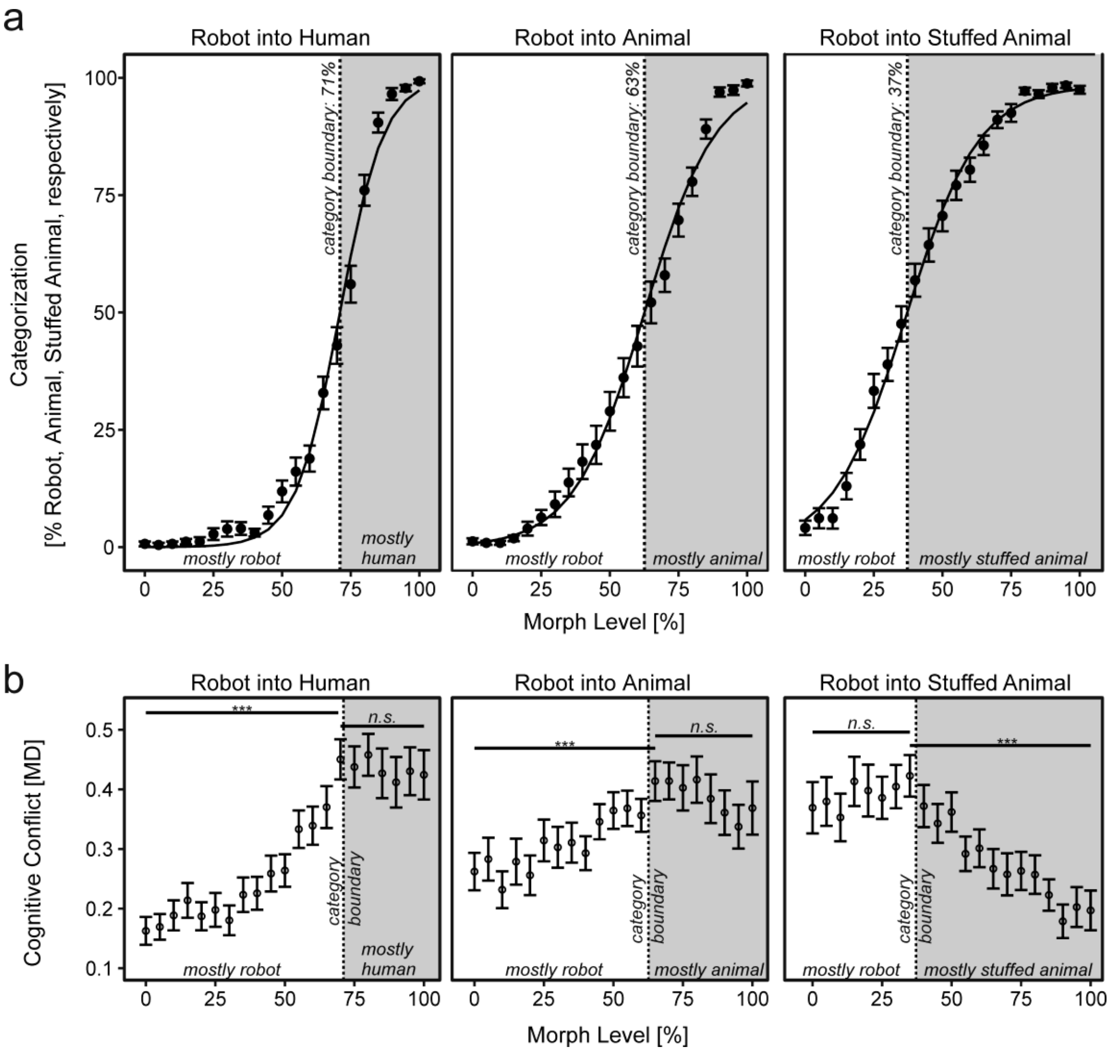

Figure 6. Cognitive Conflict at Category Boundary: (a) For all three Target Agents, the spectra exhibited a categorical boundary (see *Results: Hypothesis 1* for details). (b) Cognitive conflict varies with physical distance from the robot and, on a descriptive level, peaks around the category boundary. Error bars depict SEM. MD: Maximum Deviation (see *Methods*; for details). *** : $p < .001$, n.s. : $p > .05$

2.2.2.2 ***Hypothesis 1-2****: Cognitive conflict is maximal at spectrum-specific category boundaries*

We expected cognitive conflict processing to peak at the spectrum-specific category boundaries reported above. Cognitive conflict was measured using maximal deviation (MD), a measure derived from mouse trajectories (**Figure 5**). To investigate whether cognitive conflict processing peaked at the category boundary between the robotic and the target agents, a two-step procedure was employed. First, a mixed ANOVA with the within-factor *Morph Level* (0% to 100% category B-ness), the between-factor *Target Agent* (human, animal, stuffed animal) and MD as dependent variable was conducted as an omnibus test. A significant interaction would indicate cognitive conflict to be distributed differentially along the morph levels for the different target agents, which is what we expect since the three different target agents are associated with different category boundaries (71%, 63%, and 37%, respectively; see *H1-1*). Second, linear regression analyses were employed to investigate whether the location of the individual category boundaries (in % category B-ness) and the location with maximal cognitive conflict (in % morph level, which equals % category B-ness) co-varied.

The omnibus test indicated that cognitive conflict was altered as a function of *Morph Level* (*F*(20, 3200) = 4.18, $p < .001$, $\eta_G^2 = .03$), but not *Target Agent* (*F*(2, 160) = .79, $p = .455$, $\eta_G^2 = .01$). The interaction between *Morph Level* and *Target Agent* was significant (*F*(40, 3200) = 10.90, $p < .001$, $\eta_G^2 = 0.12$), confirming that the variation of MD along the morph levels differed between target agent types (**Figure 6b)**. Post-hoc tests that were conducted to further investigate the interaction are reported in the supplemental materials.

The linear regression analyses provided evidence that maximal cognitive conflict processing and the location of the category boundary also coincide on the individual subject level: the category boundaries (i.e., morph level at which a given participant would categorize the stimulus as belonging to the target category, e.g., “human”, in 50% of trials) were extracted from individually fitted logistic functions. For each morph spectrum (robot-human, robot-animal; robot-stuffed animal), the individual categorical boundaries were then used to predict the location of individual MD maxima (i.e. morph level at which the participant’s cognitive conflict was highest) to determine whether the location of the category boundary and the location of the MD maxima were likely to co-occur. Results show that individual category boundary locations were able to predict the location of individual cognitive conflict maxima for human target agents ($F(1, 52) = 21.75$, $p < 0.001$, $R_{adj}^2 = 0.28$)), animal target agents ($F(1, 53) = 73.19$, $p < 0.001$, $R_{adj}^2 = 0.57$), and stuffed animal target agents ($F(1, 52) = 21.83$, $p < 0.001$, $R_{adj}^2 = 0.28$); see **Figure 7**[2]. To validate that category boundary location not only *predicted* cognitive conflict maxima location but that both variables indeed *co-occurred* at the same location, we also report whether the intercept of the linear regressions differed from 0 and whether the slope differed from 1. Neither the intercept (human target agent: $t(52) = 0.53$, $p = 0.60$; animal target agent: $t(53) = 1.81$, $p = 0.08$; stuffed animal target agent: $t(52) = 0.06$, $p = 0.95$) nor the slope (human target agent: $t(52) = 0.77$, $p = 0.44$; animal target agent: $t(53) = 1.81$, $p = 0.08$; stuffed animal target agent: $t(52) = 0.05$, $p = 0.96$) were significantly different from 0 and 1, respectively. Taken together, the preceding analyses suggest that maximal cognitive conflict and the location of the category boundary tend to coincide, irrespective of whether the categorization included human or nonhuman target agents, thereby supporting *H1-2*.

[2] To validate the findings with MD as measure for cognitive conflict, we conducted the same analyses with AUC as measure for cognitive conflict. Results were highly similar for human target agents ($F(1, 52) = 24.90$, $p < 0.001$, $R_{adj}^2 = 0.31$)), animal target agents ($F(1, 53) = 53.11$, $p < 0.001$, $R_{adj}^2 = 0.49$), and stuffed animal target agents ($F(1, 52) = 31.73$, $p < 0.001$, $R_{adj}^2 = 0.37$).

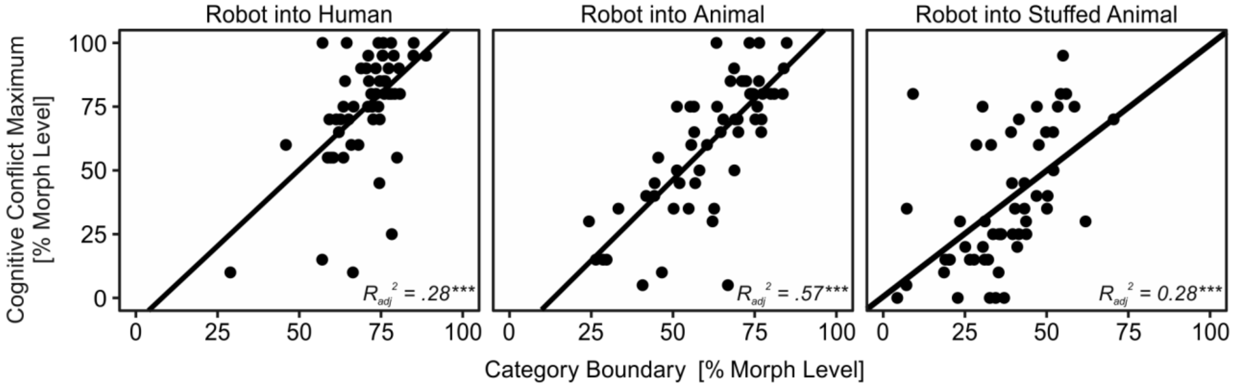

Figure 7. Relationship between Individual Cognitive Conflict Maxima and Individual Category Boundaries: For all three dimensions, irrespective of Target Agent, the location of the category boundary can be used to predict the location of the cognitive conflict maximum as measured by maximum deviation of the mouse curvatures used during categorization. *** : *p* < .001.

2.2.2.3 ***Hypothesis 2-1****: Category boundary shift towards ingroup, rendering the ingroup more exclusive*

In line with previous studies, it was hypothesized that the spectrum-specific category boundaries are shifted towards the end of the spectrum that was most representative of "own" group status (i.e., human). A one-factorial ANOVA with *Target Agent* (human, animal, stuffed animal) as between-participants factor and individual category boundaries as dependent variable was employed as an omnibus test and followed up with independent t-tests. The procedure for computing the location of individual category boundaries is analogue to the procedures employed for *H1-1*.

Results of the omnibus showed a significant effect of *Target Agent* on the location of the category boundary (*F*(2, 160) = 87.95, *p* < .001, η^2_G = .52) with the category boundary for the human target agent being located at 70.2%, for the animal target agent at 61.6%, and for the stuffed animal target agent at 36.9% of category B-ness (i.e., human, animal or stuffed animal). Three independent post-hoc t-tests confirmed significantly different category boundary locations between all target agent categories (all *t* > 3.4, all p < .001) with "human" as the most exclusive category. Please note that the slight differences between the category boundary locations reported here and in **Figure 6a** stem from the fact that in the current analysis, logistic functions were fitted to individual rather than grand average data. Also note that these results imply that there is not only a shift in the location of the category boundary between spectra with human and nonhuman target agents but also between spectra with nonhuman animal and nonhuman stuffed animal target agents.

*2.2.2.4 **Hypothesis 2-2**: Stronger conflict for outgroup-ingroup than outgroup-outgroup categorizations*

We hypothesized that due to the higher social relevance, as well as the deeper neural processing of human stimuli, decisions requiring human-nonhuman categorizations would be associated with a higher magnitude of cognitive conflict than nonhuman-nonhuman categorizations. To test this hypothesis, we took each participant's cognitive conflict for the morph levels left and right of the average spectrum-specific category boundary (e.g., MD at 70% and 75% humanness for the robot-human spectrum with the average category boundary at 71% morph level), averaged across both values, and used these average scores to compare the extent of cognitive conflict processing between human target agents (i.e., "own") and nonhuman target agents (i.e., "other": animal and stuffed animal) using an ANOVA (DV: MD at categorical boundary; IV: Target Agent) as omnibus test and independent one sided t-tests as follow-up analyses.

When not accounting for mouse movement onsets, cognitive conflict measures at spectrum-specific category boundaries did not differ between target agents, that is: high categorical uncertainty was associated with comparable cognitive conflict irrespective of whether the categorization involved human target agents ($F(2, 157) = .55$, $p = .577$, $\eta^2_G < .01$ M(human) = .44, M(animal) = .39, M(stuffed animal) = .40). However, when only looking at trials where participants started the mouse movement immediately after stimulus presentation (i.e., 300 ms after stimulus presentation or less), which is required for a meaningful interpretation of mouse tracking data[3], cognitive conflict measures around the category boundary were impacted by target agent ($F(2, 157) = 5.74$, $p = .004$, $\eta^2_G = .07$). Follow-up one-sided independent t-tests revealed higher cognitive conflict processing for categorizations involving "human" versus "nonhuman" stimuli (human vs. animal: $t(104) = 3.37$, $p < .001$, M(human) = .57, M(animal) = .38; human vs. stuffed animal: $t(104) = 1.98$, $p = .025$; **Figure 8a**). A follow-up two-sided t-test revealed no differences in cognitive conflict processing for categorizations involving two types of "nonhuman" stimuli (M(stuffed animal) = .45; animal vs. stuffed animal: $t(106) = 1.33$, $p = .187$; **Figure 8a**).

The 300 ms threshold was determined post-hoc through visual inspection of the relationship between mouse movement onset and cognitive conflict measures across all trials (**Figure 8b**). Below threshold, cognitive conflict was stable at around 0.4 and steadily declined thereafter (i.e., MD declined from around 0.4 at threshold to around 0 for a 1000 ms onset delay), indicating that in trials above threshold, participants might have partially resolved the cognitive conflict before starting to move the mouse. Mouse movement onset was above threshold in 45.9% of trials. Five participants had to be excluded from threshold-related analyses because their mouse

[3] The quality of cognitive conflict measures in mouse-tracking studies increases if participants have to start moving the mouse immediately after stimulus onset (Scherbaum & Kieslich, 2017). If participants start moving the mouse only after they resolved the conflict, MD as well as other measures relying on the mouse trajectory are not able to capture conflict processing. Here, 'the first mouse movement' was defined by the time at which participants moved the cursor more than 20 pixels either horizontally or vertically. Note that we expect cognitive conflict to be also present in trials with a movement onset later than 300ms. However, in these trials, the conflict is supposedly not captured by the mouse movement data because it had already been resolved before movement onset. Therefore, we exclude trials with late mouse movement onset only for this specific analysis.

movement onset was above threshold for all trials in proximity to the category boundary. For exploratory purposes, we also provide a graph comparing cognitive conflict processing above and below threshold for all morph levels in the *Supplemental Materials* (**Figure S1**).

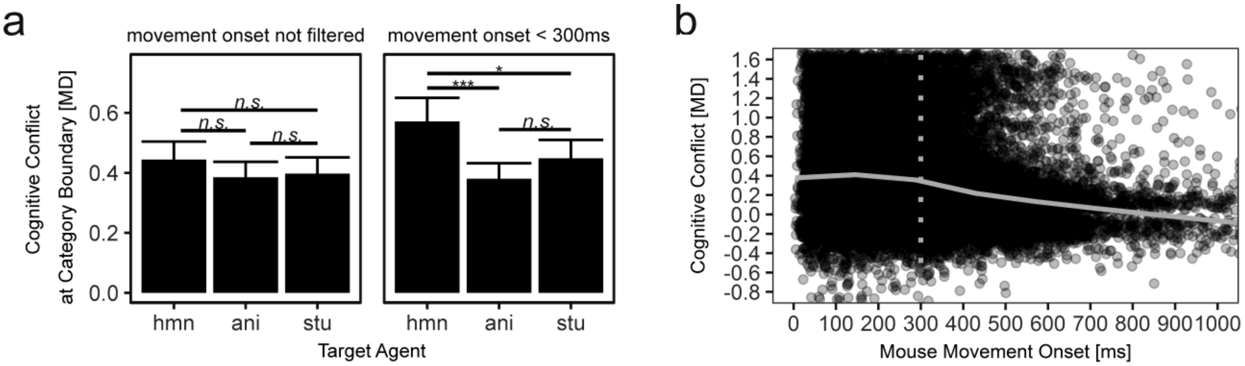

Figure 8. ***Cognitive Conflict Around Category Boundary***: (a) If trials with delayed mouse movement onset are excluded, human target agents inflict higher cognitive conflict around the category boundary than the other target agents. Error bars depict SEM. (b) Cognitive conflict processing declines with increasing mouse movement onset times. The gray line represents the loess curve (used for smoothing), which is obtained by locally weighted polynomial regressions for each point (e.g., Cleveland, Grosse, & Shyu, 1992) and was computed with the standard parameters of R’s (R Core Team, 2013) loess function. The loess curve was fitted using the whole dataset whereas the plot is zoomed in (minimal and maximal MD and maximal Mouse Movement Onset Values not depicted) and thus represents the majority but not the entirety of data. hmn: human, ani: animal; stu: stuffed animal; *** : *p* < .001, * : *p* < .05, *n.s.* : *p* > .05.

3 GENERAL DISCUSSION

It was explored to what extent previously reported observations associated with categorical perception of social entities, such as an increased categorization difficulty and shifts in the location of the category boundary, are general phenomena observed for categorically ambiguous stimuli or specific phenomena related to stimuli that are ambiguous in terms of their humanness (i.e., “human” versus “nonhuman”; e.g., Cheetham et al., 2011; Looser & Wheatley, 2010; Weis & Wiese, 2017). Using mouse tracking, it was shown that cognitive conflict processing indicative of categorical ambiguity peaks around the spectrum-specific category boundaries for all agent spectra independent of whether they contained a human endpoint or not. However, both the extent of cognitive conflict processing and the location of the spectrum-specific category boundaries were affected by the specific categorization that needed to be made, that is: stimuli located at a nonhuman-human category boundary induced stronger cognitive conflict processing than stimuli located at a nonhuman-nonhuman category boundary with no difference in the extent of cognitive conflict processing between nonalive-alive (i.e., robot-animal) and nonalive-nonalive (i.e., robot-stuffed animal) categorizations within the nonhuman spectra.

The observation that cognitive conflict is increased for all stimuli located at spectrum-specific category boundaries and not only for stimuli of ambiguous human-likeness suggests that increased processing costs for ambiguous stimuli are not specific to nonhuman-human categorizations but can be found independently of the spectrum's nature and the location of its category boundary. The results are in line with previous studies linking categorical perception to increasing cognitive processing costs (Weis & Wiese, 2017; Yamada et al., 2013), and reduced cognitive performance (Mandell et al., 2017; Wiese et al., 2019), with costs being highest and performance being lowest for categorically ambiguous stimuli. The universal observation of cognitive conflict processing for all spectra is also in line with certain claims of the *inhibitory-devaluation hypothesis* (Ferrey et al., 2015), stating that phenomena related to categorically ambiguous stimuli (i.e., falling on the mid point of the spectrum) are not directly related to human-likeness *per se*, but instead reflect a more general form of stimulus devaluation that occurs when inhibition is triggered to resolve conflict between competing stimulus-related representations. Please note that although no affective measures were obtained in the current study, the results indicate that conflict between competing categorical representations is observable for all examined spectra and not dependent on considerations regarding a stimulus' human-likeness. Since increase in cognitive processing costs and decrease in cognitive fluency has been linked to negative emotional reaction in previous studies, it is conceivable that conflict processing related to categorical ambiguity may cause negative affective reactions to uncanny stimuli; this hypothesis, however, would have to be tested empirically in future experiments.

Nevertheless, although signs of categorical processing were observed for all examined spectra, the current findings do indicate that both the *extent* to which categorically ambiguous stimuli induce a cognitive conflict (i.e., nonhuman-human categorical transitions induce more pronounced mouse curvatures than nonhuman-nonhuman categorical transitions), and the *location* of the category boundary (i.e., category boundary is biased towards “alive” stimuli and even more so towards “human” stimuli) are modulated by whether the categorization required decisions regarding a stimulus’ human-likeness, which indicates that certain aspects of categorical processing are enhanced during the perception of humanness. The observation that nonhuman-human categorizations exhibit pronounced cognitive conflict suggests that conflict resolution might be easier for spectra that do not contain the human category (e.g., robot-animal) than for spectra that contain human and nonhuman categories (e.g. robot-human). There are several possible explanations for this: First, as detailed in the introduction, “human” may be a privileged category for human observers and increase the motivation to perceive an ambiguous stimulus as “human” even though it possesses some physical features that suggest otherwise (e.g., exaggerated eyes or disproportionate eyes-nose-mouth relations). This is even more conceivable given that humans should have more perceptual expertise in processing human faces than nonhuman “faces” given the steady exposure to human faces. It is possible that increased cognitive conflict for stimuli of ambiguous physical humanness is the consequence of an ongoing competition between top-down mechanisms that lead to the expectation of a “human” stimulus and bottom-up mechanisms triggered by physical agent features implicating a “nonhuman” classification. This interpretation would be in line with previous studies showing that being in need for social connection lets individuals accept nonhuman stimuli as human despite the presence of contradicting perceptual information (Hackel et al., 2014). Similarly, it is possible that the presence of human features activate the “human” category, which is then repeatedly suppressed by knowledge that the entity is in fact not human (Misselhorn, 2009). It cannot be excluded, however, that increase in cognitive conflict processing for the robot-human morphs compared to the robot-nonhuman morphs is simply due to stronger reactions to morphed images containing “human” than “nonhuman” information (see Kätsyri et al., 2015; for a criticism of morphed images to study uncanny valley effects). Relatedly, it can also not be excluded that increase in cognitive conflict processing is related to changes in certain perceptual features (e.g., participants may have high perceptual thresholds for accepting skin color as human-like but not for nose shape, for instance) as opposed to categorical ambiguity (e.g., high thresholds for the holistic perception of a stimulus as “human” versus “nonhuman”). In other words, it is possible that perceptual ambiguity may triggered by one (or a subset of) facial feature(s) rather than the face as a holistic stimulus (in line with Moore, 2012; also see MacDorman & Chattopadhyay, 2016). Second, stimuli that possess human-like physical features or show human-like motion patterns (Castelli, Happé, Frith, & Frith, 2000) trigger anthropomorphic perceptions in a bottom-up manner within a few hundred milliseconds, and are thus harder to suppress due to their reflexive nature than nonhuman stimuli (Desimone & Duncan, 1995), which may contribute to the increased cognitive conflict. Alternatively, it is plausible that due to human preferences for “anthropomorphic” interpretations (Epley, Waytz, & Cacioppo, 2007), the activation of “nonhuman” interpretations of observed stimuli might be delayed (McMains & Kastner, 2011) and in turn may delay conflict resolution (Chattopadhyay & MacDorman, 2016; MacDorman & Chattopadhyay, 2016; Saygin, Chaminade, Ishiguro, Driver, & Frith, 2012). It cannot be excluded, however, that increased cognitive conflict processing for robot-human versus robot-nonhuman spectra is simply due to the fact that human and robot faces are more similar to each other than human and animal or human and stuffed animal faces. Third, it is possible that stimuli that are ambiguous regarding their human-likeness are more arousing due to negative affective reactions than stimuli that purport categorical ambiguities unrelated to human-likeness and thus have a stronger impact on categorical decision making during mouse tracking. Although increased arousal due to negative affective reactions has been linked to the UV (Kätsyri et al., 2015), as well as to decreased cognitive performance (Eysenck & Calvo, 1992), this interpretation seems unlikely given that Ferrey and colleagues (2015) have shown that negative affective reactions for uncanny stimuli seem to be independent of their human-likeness.

In terms of the location of the category boundaries, the current results show that although maximal cognitive conflict occurred around each spectrum's category boundary, the location of this boundary varied as a function of target agent (i.e., human vs. animal vs. stuffed animal) such that it was shifted towards the end of the spectrum that contained stimuli that were alive, human, or similar to the participant. Please note that this shift of the location of maximal cognitive conflict processing could be caused by one (or multiple) separate facial feature(s) (i.e., feature-based / quantitative explanation; compatible with the perceptual mismatch hypothesis) as opposed to the face as a whole (i.e., category-based / qualitative explanation; compatible with the categorical perception hypothesis). The observation that this rightward bias is most pronounced for alive, human, or generally "similar to self" stimuli is in line with behavioral data from previous studies using human-nonhuman spectra (Cheetham et al., 2011; Looser & Wheatley, 2010b; Martini, Buzzell, & Wiese, 2015), as well as neurophysiological data from primate studies (Sigala et al., 2011) showing preferential processing of “ingroup” stimuli. According to Sigala and colleagues (2011), this shift may reflect visual expertise for members of one’s own species and be a signature of greater brain resources assigned to the processing of privileged categories (i.e., can serve as sensitive indicators of encoding strength for categories of interest). This interpretation would be in line with numerous studies on the “other race effect” that have shown greater perceptual sensitivity for face stimuli belonging to “own” versus “other” racial groups (Hugenberg et al., 2010; for a review), as well as studies on mind perception that have shown ingroup-outgroup manipulations to affect categorical perception, such that category boundaries are shifted more strongly towards the “ingroup” end of the spectrum (e.g., same university or fan of the same sports team; Hackel et al., 2014). The current study adds to these findings by showing that shifts in category boundaries are not specific to spectra of human-likeness, but also occur for spectra of varying “nonhuman-likeness”. However, the current data cannot exclude that this shift is simply due to higher perceptual expertise of human observers for human stimuli versus animal and robot stimuli (in line with a perceptual expertise interpretation; see Sigala et al., 2011; Hugenberg et al., 2010). Future studies are needed to elucidate the impact of perceptual and motivational variables on the categorical perception of uncanny stimuli.

From a more applied point of view, our results suggest that robotic or virtual agents should be designed in the least ambiguous way possible. Interacting with unambiguous agents evokes the least cognitive conflict, thus drains the least cognitive resources, and consequently should be more pleasurable and efficient than interacting with ambiguous agents. In line with this suggestion, semi-realistic animated film characters were shown to be perceived as eerier and less likable than characters impersonated by real actors (Kätsyri, Mäkäräinen, & Takala, 2017) and interacting with ambiguous agents has been linked to decreased performance (Wiese et al., 2019). The current results indicate that such undesirable effects, though less pronounced, should not only occur when interacting specifically with ambiguous human-like but when interacting with ambiguous agents in general. The current findings however also indicate that “human” is quite an exclusive category, with the category boundary shifted far to the right side of a robot-human spectrum, making it challenging to design unambiguous humanlike agents. Thus, whenever specifically humanlike properties are not absolutely necessary, designing for nonhuman but unambiguous agents might lead to more desirable interaction outcomes than for humanlike but ambiguous agents.

4 CONCLUSION

The current experiment used mouse tracking to examine the effect of stimuli’s human-likeness on categorical perception and cognitive conflict processing. Results indicate that cognitive conflict processing is universally observed at category boundaries across morph spectra with- *and* without involvement of human agents. However, the *extent* of cognitive conflict processing and the location of category boundaries are affected by the specific nature of the spectrum. Cognitive conflict was higher for spectra containing versus not containing human agents, and the location of the category boundary was shifted towards the end of the spectrum that was more “alive”, “similar to self”, and “human”. While the current study empirically showed how human-likeness affects categorical perception, the mechanisms underlying the described modulations of cognitive conflict processing and category boundary locations remain, for the most part, unexplored. Future studies need to address this gap in the literature by exploring whether the effect of human-likeness on categorical perception is mainly perceptual or motivational in nature.

5 VITAE

Patrick Weis is a PhD student in the Human Factors and Applied Cognition Program at George Mason University. He received an MS in Neuroscience from the University of Tuebingen in 2014.

Eva Wiese is an Assistant Professor in the Human Factors and Applied Cognition Program at George Mason University. She received her PhD in Neuroscience from the Ludwig-Maximilian University of Munich, Germany in 2013.

6 REFERENCES

Assal, G., Favre, C., & Anderes, J. P. (1984). [Nonrecognition of familiar animals by a farmer. Zooagnosia or prosopagnosia for animals]. *Revue neurologique*, *140*(10), 580–584.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*(3), 379–384.

Burleigh, T. J., & Schoenherr, J. R. (2015). A reappraisal of the uncanny valley: categorical perception or frequency-based sensitization? *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.01488

Burleigh, T. J., Schoenherr, J. R., & Lacroix, G. L. (2013). Does the uncanny valley exist? An empirical test of the relationship between eeriness and the human likeness of digitally created faces. *Computers in Human Behavior, 29*(3), 759–771.

Campbell, R., Pascalis, O., Coleman, M., Wallace, S. B., & Benson, P. J. (1997). Are faces of different species perceived categorically by human observers? *Proceedings of the Royal Society of London B: Biological Sciences, 264*(1387), 1429–1434. https://doi.org/10.1098/rspb.1997.0199

Castelli, F., Happé, F., Frith, U., & Frith, C. (2000). Movement and Mind: A Functional Imaging Study of Perception and Interpretation of Complex Intentional Movement Patterns. *NeuroImage, 3*(12), 314–325. https://doi.org/10.1006/nimg.2000.0612

Chattopadhyay, D., & MacDorman, K. F. (2016). Familiar faces rendered strange: Why inconsistent realism drives characters into the uncanny valley. *Journal of Vision*, *16*(11), 7–7.

Cheetham, M., Pavlovic, I., Jordan, N., Suter, P., & Jancke, L. (2013). Category Processing and the human likeness dimension of the Uncanny Valley Hypothesis: Eye-Tracking Data. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00108

Cheetham, M., Suter, P., & Jäncke, L. (2011). The human likeness dimension of the “uncanny valley hypothesis”: behavioral and functional MRI findings. *Frontiers in Human Neuroscience*, *5*, 126.

Cheetham, M., Suter, P., & Jancke, L. (2014). Perceptual discrimination difficulty and familiarity in the uncanny valley: more like a “Happy Valley.” *Frontiers in Psychology*, *5*, 1219.

Ciechanowski, L., Przegalinska, A., Magnuski, M., & Gloor, P. (2018). In the shades of the uncanny valley: An experimental study of human–chatbot interaction. *Future Generation Computer Systems*. https://doi.org/10.1016/j.future.2018.01.055

Cleveland, W. S., Grosse, E., & Shyu, W. M. (1992). Local regression models. In J. M. Chambers & T. J. Hastie (Eds.), *Statistical models in S* (pp. 309–376).

Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annu. Rev. Neurosci.*, (18), 30.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological Review*, *114*(4), 864.

Eysenck, M. W., & Calvo, M. G. (1992). Anxiety and performance: The processing efficiency theory. *Cognition & Emotion*, *6*(6), 409–434.

Ferrey, A. E., Burleigh, T. J., & Fenske, M. J. (2015). Stimulus-category competition, inhibition, and affective devaluation: a novel account of the uncanny valley. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.00249

Forde, E. M. E., & Humphreys, G. W. (2002). Dissociations in Routine Behaviour across Patients and Everyday Tasks. *Neurocase*, *8*(1–2), 151–167. https://doi.org/10.1093/neucas/8.1.151

Freeman, J. B., & Ambady, N. (2010). MouseTracker: Software for studying real-time mental processing using a computer mouse-tracking method. *Behavior Research Methods*, *42*(1), 226–241.

Freeman, J. B., & Johnson, K. L. (2016). More Than Meets the Eye: Split-Second Social Perception. *Trends in Cognitive Sciences*, *20*(5), 362–374.

Gao, T., McCarthy, G., & Scholl, B. J. (2010). The Wolfpack Effect: Perception of Animacy Irresistibly Influences Interactive Behavior. *Psychological Science*, *21*(12), 1845–1853. https://doi.org/10.1177/0956797610388814

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, *315*(5812), 619–619.

Hackel, L. M., Looser, C. E., & Van Bavel, J. J. (2014). Group membership alters the threshold for mind perception: The role of social identity, collective identification, and intergroup threat. *Journal of Experimental Social Psychology*, *52*, 15–23. https://doi.org/10.1016/j.jesp.2013.12.001

Hugenberg, K., Wilson, J. P., See, P. E., & Young, S. G. (2013). Towards a synthetic model of own group biases in face memory. *Visual Cognition*, *21*(9–10), 1392–1417. https://doi.org/10.1080/13506285.2013.821429

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorization-individuation model: an integrative account of the other-race recognition deficit. *Psychological Review*, *117*(4), 1168–1187. https://doi.org/10.1037/a0020463

Kätsyri, J., Förger, K., Mäkäräinen, M., & Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: Support for perceptual mismatch as one road to the valley of eeriness. *Frontiers in Psychology*, *6*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4392592/

Kätsyri, J., Mäkäräinen, M., & Takala, T. (2017). Testing the ‘uncanny valley’ hypothesis in semirealistic computer-animated film characters: An empirical evaluation of natural film stimuli. *International Journal of Human-Computer Studies*, *97*, 149–161. https://doi.org/10.1016/j.ijhcs.2016.09.010

Khosla, A., Jayadevaprakash, N., Yao, B., & Li, F.-F. (2011). Novel dataset for fine-grained image categorization: Stanford dogs. *Proc. CVPR Workshop on Fine-Grained Visual Categorization (FGVC)*, *2*, 1.

Looser, C. E., Guntupalli, J. S., & Wheatley, T. (2013). Multivoxel patterns in face-sensitive temporal regions reveal an encoding schema based on detecting life in a face. *Social Cognitive and Affective Neuroscience*, *8*(7), 799–805. https://doi.org/10.1093/scan/nss078

Looser, C. E., & Wheatley, T. (2010). The Tipping Point of Animacy: How, When, and Where We Perceive Life in a Face. *Psychological Science*, *21*(12), 1854–1862.

MacDorman, K. F., & Chattopadhyay, D. (2016). Reducing consistency in human realism increases the uncanny valley effect; increasing category uncertainty does not. *Cognition*, *146*, 190–205. https://doi.org/10.1016/j.cognition.2015.09.019

MacDorman, K. F., Green, R. D., Ho, C.-C., & Koch, C. T. (2009). Too real for comfort? Uncanny responses to computer generated faces. *Computers in Human Behavior*, *25*(3), 695–710.

Mäkäräinen, M., Kätsyri, J., & Takala, T. (2014). Exaggerating Facial Expressions: A Way to Intensify Emotion or a Way to the Uncanny Valley? *Cognitive Computation*, *6*(4), 708–721. https://doi.org/10.1007/s12559-014-9273-0

Mandell, A. R., Smith, M., & Wiese, E. (2017). Mind Perception in Humanoid Agents has Negative Effects on Cognitive Processing. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*, 1585–1589. https://doi.org/10.1177/1541931213601760

Martini, M. C., Buzzell, G. A., & Wiese, E. (2015). Agent Appearance Modulates Mind Attribution and Social Attention in Human-Robot Interaction. *Social Robotics*, 431–439. https://doi.org/10.1007/978-3-319-25554-5_43

Martini, M. C., Gonzalez, C. A., & Wiese, E. (2016). Seeing Minds in Others – Can Agents with Robotic Appearance Have Human-Like Preferences? *PLOS ONE*, *11*(1), e0146310. https://doi.org/10.1371/journal.pone.0146310

Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition*, *146*, 22–32. https://doi.org/10.1016/j.cognition.2015.09.008

McMains, S., & Kastner, S. (2011). Interactions of Top-Down and Bottom-Up Mechanisms in Human Visual Cortex. *Journal of Neuroscience, 31*(2), 587–597. https://doi.org/10.1523/JNEUROSCI.3766-10.2011

McNeil, J. E., & Warrington, E. K. (1993). Prosopagnosia: A face-specific disorder. *The Quarterly Journal of Experimental Psychology Section A*, *46*(1), 1–10. https://doi.org/10.1080/14640749308401064

Meng, M., & Tong, F. (2004). Can attention selectively bias bistable perception? Differences between binocular rivalry and ambiguous figures. *Journal of Vision*, *4*(7), 2. https://doi.org/10.1167/4.7.2

Milborrow, S., Morkel, J., & Nicolls, F. (2010). The MUCT landmarked face database. *Pattern Recognition Association of South Africa, 201*(0). Retrieved from http://www.dip.ee.uct.ac.za/~nicolls/publish/sm10-prasa.pdf

Misselhorn, C. (2009). Empathy with Inanimate Objects and the Uncanny Valley. *Minds and Machines*, *19*(3), 345. https://doi.org/10.1007/s11023-009-9158-2

Mitchell, W. J., Szerszen, K. A., Lu, A. S., Schermerhorn, P. W., Scheutz, M., & MacDorman, K. F. (2011). A Mismatch in the Human Realism of Face and Voice Produces an Uncanny Valley. *I-Perception*, *2*(1), 10–12. https://doi.org/10.1068/i0415

Moore, R. K. (2012). A Bayesian explanation of the 'Uncanny Valley' effect and related psychological phenomena. *Scientific Reports, 2,* 864. https://doi.org/10.1038/srep00864

Özdem, C., Wiese, E., Wykowska, A., Müller, H., Brass, M., & Van Overwalle, F. (2016). Believing androids – fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents. *Social Neuroscience*, *12*(5), 582–593. https://doi.org/10.1080/17470919.2016.1207702

R Core Team. (2013). *R: A language and environment for statistical computing*.

Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Social Cognitive and Affective Neuroscience*, *7*(4), 413–422. https://doi.org/10.1093/scan/nsr025

Schein, C., & Gray, K. (2015). The Unifying Moral Dyad: Liberals and Conservatives Share the Same Harm-Based Moral Template. *Personality and Social Psychology Bulletin*, *41*(8), 1147–1163. https://doi.org/10.1177/0146167215591501

Scherbaum, S., & Kieslich, P. J. (2017). Stuck at the starting line: How the starting procedure influences mouse-tracking data. *Behavior Research Methods*, 1–14.

Seyama, J., & Nagayama, R. S. (2007). The Uncanny Valley: Effect of Realism on the Impression of Artificial Human Faces. *Presence: Teleoperators and Virtual Environments*, *16*(4), 337–351. https://doi.org/10.1162/pres.16.4.337

Sigala, R., Logothetis, N. K., & Rainer, G. (2011). Own-species bias in the representations of monkey and human face categories in the primate temporal lobe. *Journal of Neurophysiology*, *105*(6), 2740–2752. https://doi.org/10.1152/jn.00882.2010

Steckenfinger, S. A., & Ghazanfar, A. A. (2009). Monkey visual behavior falls into the uncanny valley. *Proceedings of the National Academy of Sciences*, *106*(43), 18362–18366. https://doi.org/10.1073/pnas.0910063106

Sterzer, P., Kleinschmidt, A., & Rees, G. (2009). The neural bases of multistable perception. *Trends in Cognitive Sciences*, *13*(7), 310–318. https://doi.org/10.1016/j.tics.2009.04.006

Tovée, M. J., & Cohen-Tovée, E. M. (1993). The neural substrates of face processing models: A review. *Cognitive Neuropsychology*, *10*(6), 505–528. https://doi.org/10.1080/02643299308253471

Wagner, D. D., Kelley, W. M., & Heatherton, T. F. (2011). Individual differences in the spontaneous recruitment of brain regions supporting mental state understanding when viewing natural social scenes. *Cerebral Cortex (New York, N.Y.: 1991)*, *21*(12), 2788–2796. https://doi.org/10.1093/cercor/bhr074

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

Weis, P. P., & Wiese, E. (2017). Cognitive conflict as possible origin of the uncanny valley. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*, 1599–1603. SAGE Publications Sage CA: Los Angeles, CA.

Wheatley, T., Weinberg, A., Looser, C., Moran, T., & Hajcak, G. (2011). Mind Perception: Real but Not Artificial Faces Sustain Neural Activity beyond the N170/VPP. *PLOS ONE*, *6*(3), e17960. https://doi.org/10.1371/journal.pone.0017960

Wiese, E., Buzzell, G. A., Abubshait, A., & Beatty, P. J. (2018). Seeing minds in others: Mind perception modulates low-level social-cognitive performance and relates to ventromedial prefrontal structures. *Cognitive, Affective, & Behavioral Neuroscience*, *18*(5), 837–856. https://doi.org/10.3758/s13415-018-0608-2

Wiese, E., Mandell, A., Shaw, T., & Smith, M. (2019). Implicit mind perception alters vigilance performance because of cognitive conflict processing. *Journal of Experimental Psychology: Applied*, *25*(1), 25.

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots As Intentional Agents: Using Neuroscientific Methods to Make Robots Appear More Social. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01663

Winkielman, P., Schwarz, N., Fazendeiro, T., & Reber, R. (2003). The hedonic marking of processing fluency: Implications for evaluative judgment. In *The psychology of evaluation: Affective processes in cognition and emotion* (pp. 189–217). Mahwah, NJ: Lawrence Erlbaum.

Wykowska, A., Wiese, E., Prosser, A., & Müller, H. J. (2014). Beliefs about the minds of others influence how we process sensory information. *PLoS One*, *9*(4), e94339.

Yamada, Y., Kawabe, T., & Ihaya, K. (2012). Can you eat it? A link between categorization difficulty and food likability. *Advances in Cognitive Psychology*, *8*(3), 248–254. https://doi.org/10.2478/v10053-008-0120-2

Yamada, Y., Kawabe, T., & Ihaya, K. (2013). Categorization difficulty is associated with negative evaluation in the “uncanny valley” phenomenon. *Japanese Psychological Research*, *55*(1), 20–32. https://doi.org/10.1111/j.1468-5884.2012.00538.x

Young, M. P., & Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. *Science*, *256*(5061), 1327–1331. https://doi.org/10.1126/science.1598577