



## OPEN Social perception of robots is shaped by beliefs about their minds

Ali Momen<sup>1,2</sup>, Kurt Hugenberg<sup>3</sup> & Eva Wiese<sup>2,4</sup>

Roboticians often imbue robots with human-like physical features to increase the likelihood that they are afforded benefits known to be associated with anthropomorphism. Similarly, deepfakes often employ computer-generated human faces to attempt to create convincing simulacra of actual humans. In the present work, we investigate whether perceivers' higher-order beliefs about faces (i.e., whether they represent actual people or android robots) modulate the extent to which perceivers deploy face-typical processing for social stimuli. Past work has shown that perceivers' recognition performance is more impacted by the inversion of faces than objects, thus highlighting that faces are processed holistically (i.e., as *Gestalt*), whereas objects engage feature-based processing. Here, we use an inversion task to examine whether face-typical processing is attenuated when actual human faces are labeled as non-human (i.e., android robot). This allows us to employ a task shown to be differentially sensitive to social (i.e., faces) and non-social (i.e., objects) stimuli while also randomly assigning face stimuli to seem real or fake. The results show smaller inversion effects when face stimuli were believed to represent android robots compared to when they were believed to represent humans. This suggests that robots strongly resembling humans may still fail to be perceived as "social" due pre-existing beliefs about their mechanistic nature. Theoretical and practical implications of this research are discussed.

In recent years, deepfakes—artificial intelligence (AI) generated images often involving celebrities or politicians—have gone viral on the internet, enabled by their hyper-photorealistic appearance. Such AI-produced content is powerful and ranges from synthesizing individual's speech and voice<sup>1</sup>, to creating images of fictional people<sup>2</sup>, to swapping a person's identity with another or altering what they are saying in a video<sup>3</sup>. Besides offering exciting opportunities, such technologies are widely recognized as a pressing threat to the believability of perceptual information<sup>4</sup> and many cases of misuse have been reported to date (e.g., deepfakes of Donald Trump being arrested; fake images of the pope in a puffer jacket<sup>5</sup>). Artificially recreating human appearance to an extent that it is not easily distinguishable from actual humans is also a major topic of interest in robotics, where androids very convincingly emulate their human counterparts (e.g., the android robot Sophia<sup>6</sup>).

Faces are most often the targets of such manipulation because they are a rich source of social information, including cues to identity and internal states<sup>7–10</sup>, and critically determine the effectiveness and believability of supposedly "human" content<sup>11,12</sup>. Significantly, it has been consistently shown people are particularly attuned to a face's orientation, with upright faces being recognized more readily than inverted faces. Since the seminal work by Yin in 1969, studies have shown that turning a face upside down interferes with our ability to process and recognize them far more than when the same is done to objects.<sup>13–15</sup> Multiple theoretical accounts have been made for the face inversion effect, with some arguing that faces are special and processed in a qualitatively different manner than non-social objects<sup>16</sup>; others argue that faces are not special but are rather processed in quantitatively different manners than non-face stimuli<sup>17,18</sup>. However, although there is some debate regarding the "special status" of faces, there is consensus that inversion does disrupt face-like processing more strongly than the processing of non-face objects<sup>13</sup>. This effect on face recognition occurs immediately after a face stimulus is presented and can be measured by activity in the N170 event-related potential—a neural indicator of face perception. It also triggers heightened fMRI activity in brain regions that are typically involved in recognizing objects<sup>19,20</sup>. Indeed, face processing is subserved by a core network that concerns the bottom-up processing of facial features (i.e., eyes, nose, mouth), as well as the spatial relations between these features: the inferior occipital gyrus or occipital face area (OFA) encodes isolated facial features (e.g., eye color), whereas the lateral fusiform

<sup>1</sup>United States Air Force Academy, Colorado Springs, CO, USA. <sup>2</sup>George Mason University, Fairfax, VA, USA. <sup>3</sup>Indiana University, Bloomington, IA, USA. <sup>4</sup>Berlin Institute of Technology, Berlin, Germany. ✉email: amomen425@gmail.com; eva.wiese@tu-berlin.de

gyrus (or fusiform face area; FFA) supports the configural processing of human faces and represents unchangeable facial information related to an individual's identity<sup>21</sup>. Changeable face features, on the other hand, such as changes in gaze direction to signal an action intention or facial expression to signal an emotional state, are processed by the posterior superior temporal sulcus (pSTS; see<sup>22</sup> for a review).

But how strongly will hyper-photorealistic yet synthetically generated face stimuli, such as images of deepfakes or androids, trigger face-typical processing<sup>23–25</sup>? Hyper-photorealistic depictions are easy to produce with deep-learning models, such as Generative Adversarial Networks (GANs), which pit a generator and a discriminator neural network against each other so that—over multiple iterations—the generator learns to synthesize increasingly realistic face stimuli until the discriminator is unable to distinguish them from real faces<sup>11</sup>; similar results can be obtained using diffusion models<sup>26</sup>. The generated images are often so visually convincing<sup>27</sup> that even sophisticated algorithms cannot reliably distinguish them from real human faces<sup>28–31</sup>. Indeed, whereas early deepfakes were still detectable by humans<sup>32–34</sup>,

the synthetic faces now being created are so lifelike that they are often indistinguishable from genuine human faces<sup>11,23,24</sup>. Due to their photorealistic visual similarity to human faces, one would expect that deepfakes activate the core face perception network in a similar manner as human faces. However, both behavioral and electrophysiological studies indicate that deepfake images are processed differently than human faces. ERPs like the N170<sup>35</sup> or steady-state visually evoked potentials (SSVEPs;<sup>36</sup>), for instance, appear sensitive to actual versus artificially generated faces. Furthermore, computer-generated faces do not always elicit face-processing phenomena typical of real faces (e.g., “other race effect”<sup>37</sup>). Thus, although highly realistic to the point of near indistinguishability from actual human faces, computer-generated face stimuli can differ in their manner of processing.

How might we understand why even hyper-realistic synthetic faces elicit a different manner of processing than actual human faces? We believe that this may be due to the core face network receiving top-down input from an extended network of other brain areas involved in social cognition (see<sup>22,38</sup>). This suggests that face perception may be the product of both perceptual characteristics of the face stimuli themselves and higher-order social cognitive influences reflecting perceivers' motives and beliefs<sup>39,40</sup>. Thus, even when synthetic face stimuli are perceptually indistinguishable from real human faces, they might still not be processed in the same way as human faces when perceivers know they are artificial. Perhaps simply believing a face to be artificial can influence the extent to which it is processed in a face-like manner. The most conservative test of this hypothesis would be to take actual human faces and make participants believe—through a cover story—that they are not human. Would such a top-down manipulation of beliefs be sufficient to attenuate face-typical processing, and as a consequence lead to a reduced face inversion effect?

This question is highly relevant not just to understanding the processing involved with deepfakes and androids, but also applies to the perception of human faces themselves. In most traditional face perception studies, the perception of a stimulus (like a face) cannot easily be separated from people's beliefs about that stimulus (for instance, believing it represents a human). Consider, for instance, Yin's classical work<sup>14,15</sup> on the face inversion effect, where perceivers saw human faces and non-face objects, such as houses: they almost certainly thought that the faces represented actual people, and the houses real objects. However, with hyper-realistic face stimuli like deepfakes<sup>11</sup> or highly human-like robots like androids<sup>41</sup>, we live in an era where percepts and beliefs are not necessarily conflated. This creates the possibility of dissociating bottom-up effects associated with the percepts of human faces from top-down effects associated with beliefs about human faces, and allows us (i) to examine to what extent face processing is cognitively penetrable<sup>42,43</sup>, and (ii) whether beliefs about a face's essential human-ness influence the manner in which it is processed.

Indeed, there is evidence in the literature suggesting that manipulating perceivers' beliefs about faces can modulate face processing<sup>44</sup>. For instance, when the same face is believed to belong to an ingroup (i.e., fellow university students), it is afforded stronger configural processing than when the face is believed to belong to a social outgroup (i.e., a competing university<sup>45</sup>). Similarly, a stronger N170 component was found when face stimuli were categorized as belonging to the participants' own social group versus an outgroup<sup>46</sup>. Beliefs about a target's moral behavior can also modulate face processing. Fincher and Tetlock<sup>47</sup>, for instance, found that faces of targets believed to be engaged in immoral or inhumane behavior elicit less face-like processing across multiple measures (see<sup>48</sup>). Taken together, this suggests that beliefs regarding the human nature of faces can modulate the extent to which they are processed in a face-typical manner. What these experiments cannot answer, however, is to what extent the mere belief that a human face may represent an object—namely an artificially generated face or an android robot—may disrupt face-typical processing.

### Current research

We investigate the extent to which the face inversion effect—one of the longest standing and best replicated effects in face processing—depends on participants' beliefs that apparent human faces represent actual human or non-human agents (i.e., androids with hyper-realistic human-like appearance). To do so, we manipulate beliefs about images of actual human faces via instruction while holding the percept of faces constant: participants were led to believe they would see the faces of actual humans (labeled as *human*) or the faces of highly human-like robots (labeled as *android*); in actuality, all images depicted real humans. If merely believing that a face represents a human versus a non-human agent modulates face-typical processing, the face inversion effects should be stronger in the human than the android condition. If true, this would suggest that the face inversion effect is driven, at least in part, by beliefs about the humanness of the face stimulus itself.

The current research adds to the previous literature on face perception in multiple ways. First, past work has focused on how the perceptual characteristics of faces influence face-like processing of non-human faces—such as those of robots—in a bottom-up manner. For example, Momen, Hugenberg, and Wiese (2022) had participants complete a face inversion task with human and humanoid robot faces, finding that robot stimuli elicited smaller

inversion effects than human stimuli; it was also found that robot faces that physically resembled human faces (i.e., they contained more human-like facial features) were processed in a more face-like manner than were robots that appeared less human-like<sup>49</sup>. The present study builds upon this work in multiple ways. First, by keeping the bottom-up features of a stimulus constant but manipulating only participants' beliefs about the human nature of the stimulus. Second, although past work has already demonstrated that beliefs regarding one's social group status or morality can exert a top-down influence on face processing<sup>45,46</sup>, they cannot answer the question of whether beliefs regarding an agent's humanness also modulate face processing.

## Methods and materials

### Participants

A power analysis conducted with the effect size ( $n^2=0.07$ ) of a previous experiment utilizing the inversion task in a within-subjects design<sup>50</sup>; Experiment 2) indicated a sample size of 86 would yield more than 95% power. Given that we anticipated removing participants due to poor performance and/or failing the manipulation check at the end of the experiment (i.e., *Did you believe the android faces were creations of a robot designer we are collaborating with?*), we conservatively oversampled.

233 participants completed the study via Amazon's Mechanical Turk for pay. Of these, 44 participants were removed due to poor performance in the face recognition task ( $d' \leq 0$ ), and 99 were removed due to failing the manipulation check at the end of the experiment. This resulted in a final sample size of 90 participants ( $M_{\text{age}} = 32.90$  years; Range = 18–68 years; 47 females). Filtered participants accounted for 57.10% of our data. Informed consent was obtained from all participants, and all procedures were approved by the Office of Research Integrity and Assurance (ORIA) at George Mason University (GMU). The experiment was performed in accordance with relevant guidelines and regulations.

### Apparatus

The experiment was run using the Inquisit 5<sup>51</sup> platform online, which allows collection of behavioral data remotely over the web via participant keystroke. Participants completed the experiment locally on their computer after downloading the software. Attributes like screen size, keyboard or refresh rate depended on participant's individual computers and were not controlled.

### Stimuli

The sample of faces consisted of 80 White human male faces obtained from the Chicago Face Database<sup>52</sup>. All faces were converted to greyscale and presented on a white background that measured  $768 \times 768$  pixels. The longest point, from the top of the head to the bottom of the chin, nearly touched both the top and bottom edges of the  $768 \times 768$ -pixel frame. Faces were also presented with labels and colored frames depending on their condition. Faces in the human condition were presented with an "human" label in a green frame. Faces in the android condition were presented with an "android" label in a blue frame; see Fig. 1. Consent to publish these stimuli images was obtained from the Chicago Face Database.

For each participant, the 80 face images were randomly categorized into distinct groups based on their designated condition (human vs. android), phase (learning vs. recognition) and orientation (upright vs. inverted). For



**Figure 1.** Example face stimuli. For each participant, faces were presented in the human or android condition in a randomized fashion. Each face was presented together with a label and a colored frame indicating the experimental condition: faces in the human condition were presented with the label "human" in a green frame; faces in the android condition were presented with the label "android" in a blue frame.

the “human” condition, 40 randomly chosen images from the data base were allocated as follows: (1) *Presented in Learning Phase—Repeated Upright in Recognition Phase*: 10 images were randomly allocated to this category. These images were introduced upright during the learning phase and reappeared in upright orientation during the recognition phase. (2) *Presented in Learning Phase—Repeated Inverted in Recognition Phase*: Another 10 images were randomly assigned to this category. These images were presented upright in the learning phase and displayed inverted during the recognition phase. (3) *Not Presented in Learning Phase—Distractor in Recognition Phase Upright*: Another 10 images were randomly assigned to this category. The images were exclusively used as distractors in the upright orientation during the recognition phase and were not part of the learning task. (4) *Not Presented in Learning Phase—Distractor in Recognition Phase Inverted*: Consisting of another 10 randomly assigned images, this category mirrored the previous one but with the images appearing inverted during the recognition phase. The remaining 40 images followed the same distribution pattern as the human sets but were specifically used for the “android” condition. These images were similarly divided into repeated and distractor categories, with orientations corresponding to those in the human tasks. This random categorization into distinct groups was done before the beginning of the experiment separately for each participant to minimize the chances that any observed effects could be ascribed to the unique features of the faces.

### Procedure and task

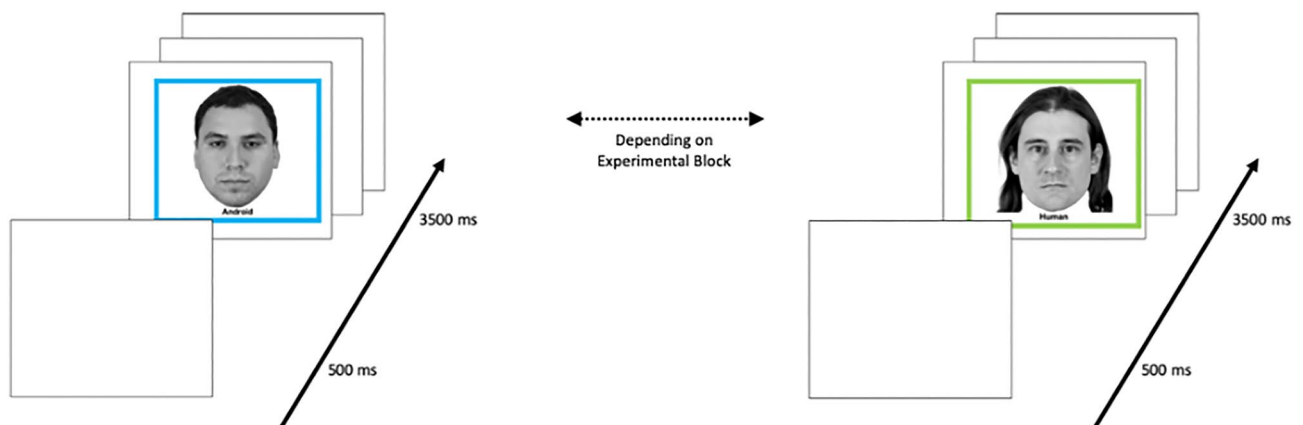
Participants first gave informed consent and were then presented with the cover story: “We are working with an android robot designer who designs robots nearly indistinguishable from humans. Here are some examples of his work”. This was followed by images of android robots difficult to distinguish from humans (e.g.,<sup>53</sup>).

Subsequently, participants completed the face inversion task. They were asked to memorize and recognize both human- and android-labeled faces. Agent type was blocked such that participants completed the learning and recognition task for one agent type (e.g., human) first before completing the learning and recognition task for the other agent type (e.g., android). Block order (android first vs. human first) was counter-balanced across participants with roughly half completing the android-labeled faces first (N = 44), and the other half completing the human-labeled faces first (N = 46).

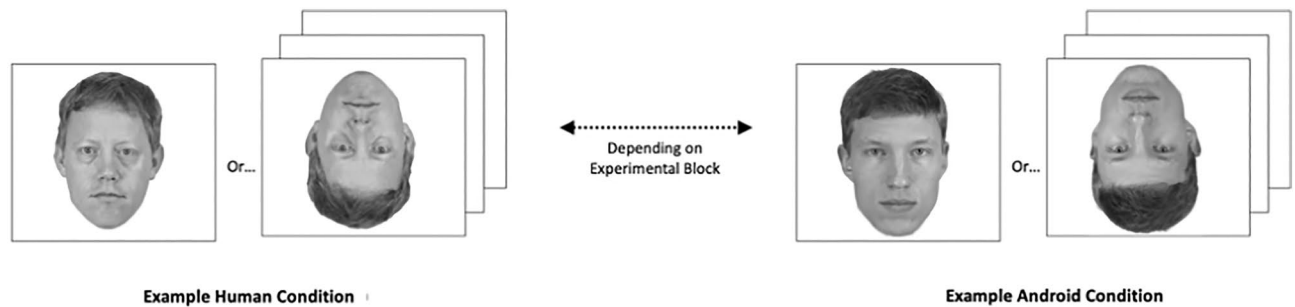
During the learning phase, participants viewed 20 target faces for 3500 ms each, with an inter-trial interval (ITI) of 500 ms; see Fig. 2. The order of presentation was random. Faces in the human condition were presented with a “human” label in a green frame; faces in the android condition were presented with an “android” label in a blue frame.

During the recognition phase that followed, participants were presented again with the 20 faces they had seen during the learning phase, interspersed with 20 new faces. All face stimuli were presented in a random order, without original labels or colored frames. Half of the faces—previously seen and new ones—were inverted. Participants indicated if they recognized a face from the learning phase by pressing the ‘D’ key; if they did not recognize it, they were asked to press the ‘K’ key. Each trial in the recognition phase lasted until a response was given by the participant; see Fig. 3.

After completing both the android and human blocks, participants were thanked, for their participation, debriefed about the experiment, and compensated.



**Figure 2.** Learning phase of the inversion task. Participants were instructed that they would see twenty upright-presented human or android faces and were asked to attend to the faces for subsequent recognition. Participants passively viewed the 20 randomly selected faces in a randomized order for 3500 ms each, with an inter-trial interval (ITI) of 500 ms. Each participant saw both human-labeled and android-labeled faces. Agent type (human vs. android) was blocked, and block order was counterbalanced, such that half of the participants started the experiment with the android-labeled faces and ended the experiment with the human-labeled faces; the other half of the participants started the experiment with the human-labeled faces and ended the experiment with the android-labeled faces. Faces were randomly assigned to the “android” or “human” labels across participants (i.e., whether a given face was labeled as “android” or “human” randomly varied across participants).



**Figure 3.** Recognition phase of the inversion task. At the beginning of the recognition phase, participants saw another series of faces, which included the 20 faces they had previously seen, interspersed with 20 new faces. These images were presented in random order without colored frames. Half of the new and half of the previously seen faces were presented upside down and participants were instructed that their task would be to report whether they had seen the face during the learning task.

## Results

Of primary interest was whether there was a differential face inversion effect for faces labeled as “human” versus “android.” To assess this, we first calculated *hits* (i.e., familiar face correctly identified), *misses* (i.e., familiar face not correctly identified), *correct rejections* (i.e., unfamiliar face correctly rejected) and *false alarms* (i.e., unfamiliar face falsely identified) during the recognition phase. We then used these measures to calculate the signal detection parameter *sensitivity* ( $d'$ )—a measure of recognition that accounts for both hit rates and false alarms<sup>54</sup>. To account for 100% hit-rates and/or 0% false alarm rates,  $d'$  scores were adjusted via a log-linear approach<sup>54,55</sup>.

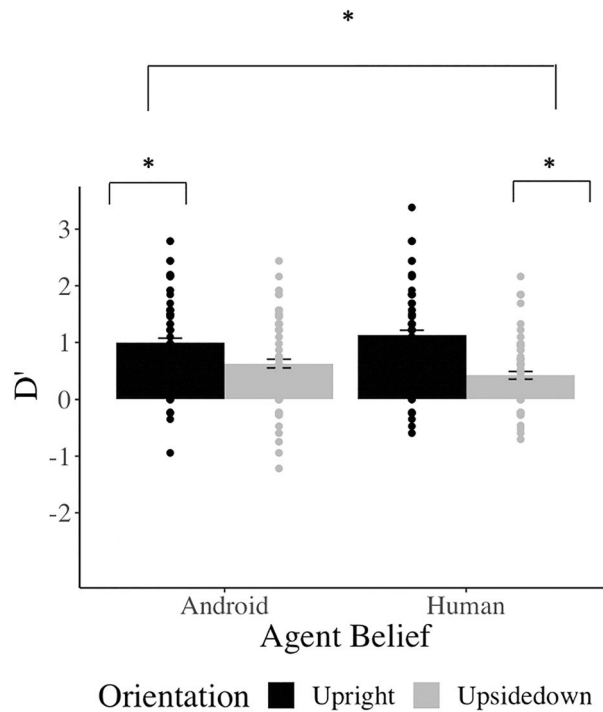
Participants'  $d'$  scores were entered into a 2 (*Agent Label*: Human vs. Android)  $\times$  2 (*Orientation*: Upright vs. Inverted) repeated-measures ANOVA. If face-typical processing was attenuated by labeling a human face as android, a significant interaction effect between agent label and orientation would be expected, such that the difference in  $d'$  between upright and inverted faces would be reduced for android-relative to human-labeled face stimuli.

The ANOVA revealed a main effect of *Orientation* ( $F(1,189) = 52.94, p < 0.001$ ), such that participants had better recognition performance for upright ( $M = 1.06, SD = 0.52$ ) compared to inverted ( $M = 0.52, SD = 0.68$ ) faces; this represents the classic face inversion effect. There was no significant main effect for *Agent Label* ( $F(1,189) = 0.29, p = 0.593$ ), indicating that the belief manipulation did not have an effect on participants' overall recognition performance. Most importantly, and in line with our hypothesis, the interaction effect between *Agent Label* and *Orientation* was significant ( $F(1,189) = 8.58, p = 0.004$ ), such that there was a larger difference in recognition performance for faces believed to be human when presented upright ( $M = 1.12, SD = 0.83$ ) vs. inverted ( $M = 0.42, SD = 0.61$ ) than for faces believed to be android (upright:  $M = 0.99, SD = 0.77$ ; inverted:  $M = 0.63, SD = 0.73$ ); see Fig. 4. Although the inversion effect was significant for both android ( $t(89) = 3.74, p < 0.001, d = 0.48$ ) and human conditions ( $t(89) = 7.93, p < 0.001, d = 0.97$ ), the effect was more than twice as strong in the “human” than the “android” belief condition.

## Ancillary analyses

We also conducted two ancillary analyses. *First*, we explored whether explicit beliefs about the android nature of the images (i.e., participants who believed the manipulation) was required for the observed pattern of effects, or if the effects held even when including participants in the data analysis that did not explicitly believe the manipulation. For this purpose, we re-ran the original analysis where participants had been excluded from the dataset who did not believe the instruction manipulation, this time with the data of the 99 participants who had originally failed the manipulation check (total  $N = 189$ ). The analysis showed that although there was a significant inversion effect across conditions (*Orientation*:  $F(1,188) = 192.33, p < 0.001$ ), the *Orientation*  $\times$  *Agent Label* interaction was no longer significant ( $F(1,188) = 1.91, p = 0.169$ ). The main effect of *Agent Label* was not statistically significant;  $F(1,188) = 1.25, p = 0.265$ . This indicates that explicit beliefs about the human versus non-human nature of face stimuli are required for the observed modulation of face-typical processing.

*Second*, we investigated the criterion ( $c$ ), which is the point at which individuals decide if a stimulus is familiar (‘old’) or unfamiliar (‘new’). A positive criterion indicates a tendency to not recognize stimuli, suggesting a cautious approach, whereas a negative criterion suggests a tendency to recognize stimuli, indicating a more lenient approach. We submitted the criterion scores to a two-way ANOVA to evaluate the impact of orientation and agent label on participants' decision thresholds. Neither the main effect of *Orientation* ( $F(1, 89) = 0.449, p = 0.504$ ), nor the main effect of *Agent Label* ( $F(1, 89) = 3.015, p = 0.086$ ) were significant. The interaction effect of *Orientation* and *Agent Label* was also not significant,  $F(1, 89) = 0.487, p = 0.487$ . In the context of our original  $d'$ -prime results, which indicated a larger difference in recognition performance for upright versus inverted faces with the ‘human’ label compared to the ‘android’ label, this criterion analysis suggests that although participants' ability to discriminate was affected by these factors, their overall threshold for deciding whether to recognize a face was not swayed by the orientation or the label. This implies that the observed differences in  $d'$ -prime are likely due to true perceptual discriminability rather than changes in decision bias.



**Figure 4.** Main results: The 2 (agent label: human vs android)  $\times$  2 (orientation: upright vs inverted) ANOVA revealed a significant main effect of orientation, indicating a significant inversion effect across conditions. Importantly, the agent label  $\times$  orientation interaction was also significant, indicating a stronger inversion effect when (perceptually identical) face stimuli were believed to represent human vs. android agents. Error bars represent 1 standard error above and below the mean.

## Discussion

Synthetic faces are increasingly realistic, to the point of near indistinguishability from actual human faces. However, such stimuli are not always processed in a fully face-like manner. In the present work, we sought to understand how believing that a face was human versus synthetic (i.e., non-human) would influence typical face perception processes. Specifically, we compared the extent to which human versus android labels affected face-typical processing in the face inversion task.

We found that when participants believed a face was that of an android, it elicited significantly smaller inversion effects than when they believed it was a human face. This was true even though all stimuli (i) were pictures of actual human faces and (ii) were randomly assigned to the human-versus-android labels, thus ruling out the possibility that the effects were driven by perceptual differences in the stimuli themselves. Indeed, while previous studies have demonstrated that visual differences between robot and human faces can affect performance in inversion tasks<sup>56</sup> these findings cannot be fully explained by visual factors. Instead, it appears that the differences in how human and android faces are processed stem from the observer's beliefs rather than visual differences inherent to the stimuli.

Human-versus-android labeling muted but did not eliminate the effect of face inversion on recognition: even amongst ostensibly android stimuli there was still a robust effect of inversion on recognition performance. This makes sense given that all stimuli were actually human faces, and perceivers had significant prior expertise with this group of stimuli<sup>57,58</sup>. Even believing a human face to be non-human does not eliminate this prior perceptual experience, highlighting that person perception is best understood as a confluence of both top-down characteristics of perceivers (e.g., beliefs about targets; motives of perceivers) and bottom-up cues of the targets themselves (e.g., facial features; see<sup>44</sup> for a review). Although we were able to influence face processing by altering participants' beliefs (top-down effects), the impact of face inversion on recognition, even when faces were labeled as androids, mean that actual human facial characteristics (bottom-up cues) still influence performance on face recognition tasks. This occurs despite believing that these faces are not human. This indicates that both the perceivers' top-down beliefs and the bottom-up visual features of the faces contribute to how we perceive and recognize individuals (see also<sup>56</sup>).

Beyond the implications for face perception, our results also have implications for research on human–robot interaction (HRI), mind perception (i.e., ascribing human-like traits to human and non-human agents<sup>59</sup>) and animacy (i.e., ascribing aliveness to human and non-human agents<sup>60</sup>), and social cognition research more broadly. Speaking first to HRI, the present work indicates that even highly human-like robots may not be fully perceived as social stimuli with emotions, identities and intentions, which has been suggested to influence various aspects of HRI, ranging from action understanding to acceptance<sup>13,61–63</sup>. Thus, robot designers should not focus solely on

physical robot features, but also on factors outside of appearance to make robots appear “social”, such as behaviors or beliefs about non-human agents themselves (see<sup>64</sup>, for a review). Believing that an agent belongs to a social ingroup, or is similar to oneself in general, is an effective method of modulating perceptions of humanness. In HRI, it may be effective to emphasize similarities between a robot and an interacting human to boost face-like processing<sup>59,65</sup>. This could be accomplished via matching physical facial characteristics (e.g., eye color), behaviors (e.g., mirroring gestures) or personality characteristics (e.g., ways of speaking). Furthermore, previous research suggests that the mere belief a non-human entity has human-like capacities can cause perceivers to treat them as social entities, which could further influence face-like processing with non-human agents<sup>66</sup>. Thus, research in HRI would benefit from considering how such motives can influence the perception of robots.

The present research contributes to our understanding of how we perceive animacy and mental states in non-human entities. Research suggests that the extent to which stimuli are processed as mindless machines versus as mindful agents is modulated by perceivers’ motives for social connection. Past research demonstrates that humans’ innate desire to connect socially can motivate them to perceive intentionality in a robot in order satisfy this motivation<sup>67,68</sup>. Furthermore, when loneliness is induced in perceivers (i.e., via experimental manipulation), they are more likely to ascribe inner states and animacy to ambiguously animate agents (e.g., human-doll morphs<sup>60</sup>). This raises the possibility that such motives might modulate face perception processes in non-human agents as well. Perhaps when robots are designed for social interaction with human partners, this may motivate participants to attribute human-like traits to them, influencing how they perceive robot faces. Alternately, perhaps lonely perceivers might engage face-typical processing even for android robot faces (see<sup>20</sup>; for similar results with animals).

Our findings add to research examining the impact of top-down influences on early perceptual processes, such as face perception. However, the present work is distinct from other recent demonstrations in important ways. First, much recent research has focused more on social categories such as race and gender, and less on direct manipulations of the perceived humanness of agents. Second, whereas previous studies showed that inhumane behavior can lead to a decline in face-typical processing<sup>47</sup>, our research highlights that already beliefs about the non-human nature of a stimulus is sufficient to attenuate face perception. This is an important distinction as the capacity for inhumane acts is intrinsically linked to humanity; for instance, a lion hunting its prey is not considered inhumane, but a human harming another reflects a violation of human ethics. In our experiments, androids are treated as distinctly non-human rather than unethical or inhumane, highlighting that our findings are fundamentally different from previous findings due to the nature of the manipulation we employed. To our knowledge, our study is the first to find that believing a face lacks a human essence affects face-typical processing even when controlling for perceptual features of the stimuli. Although past work has hinted at this possibility<sup>69,70</sup>, the present work is the first that has manipulated the essential humanness of targets in a categorical manner.

The limitations of the current study provide avenues for future studies. First, we elected to manipulate the apparent humanness of faces via a human-versus-android label manipulation. This was intentional because we believed that the distinction between humans and androids would be plausible to some participants, while holding both the percept of the agents and some apparent mental capacities of the agents constant. For example, although androids are seen as lacking experiential capacities (i.e., emotional depth), both androids and humans are typically seen as sharing the capacity for mental agency (i.e., the ability to act on the world). However, highly realistic non-human faces generated via artificial intelligence could instead vary more widely in terms of their perceived capacities. For example, deepfakes may be seen as lacking minds altogether (i.e., experiential and agential capacities; e.g.,<sup>71</sup>) whereas androids have the ability to remember stimuli and to act on the world. Future research would benefit from understanding how inducing perceptions of different capacities in non-human stimuli may influence face perception.

Second, while it is clear that participants engaged in face-typical processing less strongly when believing stimuli to be androids, it is unclear where in the cognitive stream this took effect. Indeed, top-down effects on perceptual processing can take place at several levels, such as (i) the user’s focus of attention, (ii) which percepts are selected for further perceptual processing, (iii) perceptual organization of selected percepts (i.e., configural processing), and (iv) the representation of stimuli at a higher cognitive level<sup>72</sup>. Effects on all these levels could have manifested in smaller inversion decrements for android- versus human-believed stimuli. Since it is unclear from the current study what level of perception caused these differences in processing, future research would benefit from unpacking the temporal dynamics of these effects.

Third, while we interpret our findings within the context of embodied robotic agents, our study only examined static images. This limitation raises the question of whether the reported findings apply exclusively to static images and might not hold for physical robots. Indeed, past work has shown that static face images can have different effects than dynamic facial images, especially in situations where beliefs about the minds of targets (i.e., human versus robot) are being manipulated (e.g.<sup>73</sup>). Future research should explore whether embodied agents are also subject to disengagements in face-typical processing that occurs due to beliefs about targets.

Finally, we employed the face inversion paradigm to index face-typical processing. Although it is well established that faces are more sensitive to inversion than are most other objects, the inversion paradigm does not bear directly on the question of whether faces are processed in a qualitatively different manner than non-face stimuli<sup>18</sup>. Further, inversion could affect both perceptual processing broadly and face-like processing specifically. However, given that past work has shown that faces are more strongly affected by inversion than are most non-face objects, and given that we observe human-labeled faces were more strongly affected by inversion than were android-labeled faces, it seems reasonable to conclude that the android-label did reduce face-typical processing. However, it will be important in future research to compare inversion to other manipulations that affect image processing but not face orientation to differentiate where in the perceptual-cognitive stream that these effects emerge.

## Conclusion

In the present study, we find that manipulating beliefs about whether a face represents a real human versus a synthetic android influences face-typical processing, even when holding constant the faces themselves. To our knowledge, this is the first demonstration that believing a face represents a human or a robot, holding the stimulus itself constant, influences face perception. This research has implications not just for our understanding of face perception, but also for research in HRI, mind perception, and social cognition.

## Data availability

The datasets used and/or analyzed during the current study available from the corresponding author on request.

Received: 31 July 2023; Accepted: 29 January 2024

Published online: 05 March 2024

## References

- van den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A. & Kavukcuoglu, K. *WaveNet: A Generative Model for Raw Audio*. <https://doi.org/10.48550/arXiv.1609.03499> (2016).
- Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J. & Aila, T. *Analyzing and Improving the Image Quality of StyleGAN*. 8110–8119. [https://openaccess.thecvf.com/content\\_CVPR\\_2020/html/Karras\\_Analyzing\\_and\\_Improving\\_the\\_Image\\_Quality\\_of\\_StyleGAN\\_CVPR\\_2020\\_paper.html](https://openaccess.thecvf.com/content_CVPR_2020/html/Karras_Analyzing_and_Improving_the_Image_Quality_of_StyleGAN_CVPR_2020_paper.html). Accessed 21 June 2023 (2020).
- Suwajanakorn, S., Seitz, S. & Kemelmacher, I. Synthesizing Obama: Learning lip sync from audio. *ACM Trans. Graph.* **36**, 1–13. <https://doi.org/10.1145/3072959.3073640> (2017).
- The Rise of the Deepfake and the Threat to Democracy | Technology | The Guardian*. <https://www.theguardian.com/technology/ng-interactive/2019/jun/22/the-rise-of-the-deepfake-and-the-threat-to-democracy>. Accessed 21 June 2023.
- Farid, H., & McGregor, J. We have the technology to fight manipulated images and videos. It's time to use it. *Fast Company*. <https://www.fastcompany.com/90575763/we-have-the-technology-to-fight-manipulated-images-and-videos-its-time-to-use-it>. Accessed 24 June 2023 (2020).
- Sophia. *Hanson Robotics*. <https://www.hansonrobotics.com/sophia/>. Accessed 21 June 2023.
- Hugenberg, K., Wilson, J. P., See, P. E. & Young, S. G. Towards a synthetic model of own group biases in face memory. *Vis. Cognit. Sci.* **119**(8), 1392–1417. <https://doi.org/10.1080/13506285.2013.821429> (2013).
- Johnson, K. & Hugenberg, K. Perception of faces and bodies. In *The Oxford Handbook of Social Cognition*. 2nd edn. (in press)
- Piepers, D. & Robbins, R. A review and clarification of the terms “holistic”, “configural”, and “relational” in the face perception literature. *Front. Psychol.* <https://doi.org/10.3389/fpsyg.2012.00559> (2012).
- Willis, J. & Todorov, A. First impressions: Making up your mind after a 100-ms exposure to a face. *Psychol. Sci.* **17**(7), 592–598. <https://doi.org/10.1111/j.1467-9280.2006.01750.x> (2006).
- Nightingale, S. J. & Farid, H. AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proc. Natl. Acad. Sci.* **119**(8), e2120481119. <https://doi.org/10.1073/pnas.2120481119> (2022).
- Selvaraj, N. *Real Face or AI Generated Fake? Medium*. <https://towardsdatascience.com/real-face-or-ai-generated-fake-d95b30cf86f>. Accessed 24 June 2023 (2021).
- Maurer, D., Le Grand, R. & Mondloch, C. J. The many faces of configural processing. *Trends Cognit. Sci.* **6**(6), 255–260. [https://doi.org/10.1016/S1364-6613\(02\)01903-4](https://doi.org/10.1016/S1364-6613(02)01903-4) (2002).
- Yin, R. K. Looking at upside-down faces. *J. Exp. Psychol.* **81**(1), 141–145. <https://doi.org/10.1037/h0027474> (1969).
- Yin, R. K. Face recognition by brain-injured patients: A dissociable ability?. *Neuropsychologia* **8**(4), 395–402. [https://doi.org/10.1016/0028-3932\(70\)90036-9](https://doi.org/10.1016/0028-3932(70)90036-9) (1970).
- Taubert, J., Apthorp, D., Aagten-Murphy, D. & Alais, D. The role of holistic processing in face perception: Evidence from the face inversion effect. *Vis. Res.* **51**(11), 1273–1278. <https://doi.org/10.1016/j.visres.2011.04.002> (2011).
- Sekuler, A. B., Gaspar, C. M., Gold, J. M. & Bennett, P. J. Inversion leads to quantitative, not qualitative, changes in face processing. *Curr. Biol.* **14**(5), 391–396. <https://doi.org/10.1016/j.cub.2004.02.028> (2004).
- Valentine, T. Upside-down faces: A review of the effect of inversion upon face recognition. *Br. J. Psychol.* **79**(4), 471–491. <https://doi.org/10.1111/j.2044-8295.1988.tb02747.x> (1988).
- Rousselet, G. A., Macé, M.-M. & Fabre-Thorpe, M. Animal and human faces in natural scenes: How specific to human faces is the N170 ERP component?. *J. Vis.* **4**(1), 2. <https://doi.org/10.1167/4.1.2> (2004).
- Young, S. G., Goldberg, M. H., Rydell, R. J. & Hugenberg, K. Trait anthropomorphism predicts acquiring human traits to upright but not inverted chimpanzee faces. *Soc. Cognit.* **37**(2), 105–121. <https://doi.org/10.1521/soco.2019.37.2.105> (2019).
- Kanwisher, N., McDermott, J. & Chun, M. M. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. Neurosci.* **17**(11), 4302–4311. <https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997> (1997).
- Kawakami, K., Amodio, D. M. & Hugenberg, K. Chapter One—Intergroup perception and cognition: An integrative framework for understanding the causes and consequences of social categorization. In *Advances in Experimental Social Psychology* (Olson, J. M. ed.), 1–80. <https://doi.org/10.1016/bs.aesp.2016.10.001> (Academic Press, 2017).
- Hulzebosch, N., Ibrahim, S. & Worring, M. Detecting CNN-generated facial images in real-world scenarios. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2729–2738. <https://doi.org/10.1109/CVPRW50498.2020.00329> (2020).
- Lago, F. et al. More real than real: A study on human visual perception of synthetic faces. *IEEE Signal Process. Mag.* **39**(1), 109–116. <https://doi.org/10.1109/MSP.2021.3120982> (2022).
- Sofer, C., Dotsch, R., Wigboldus, D. H. J. & Todorov, A. What is typical is good: The influence of face typicality on perceived trustworthiness. *Psychol. Sci.* **26**(1), 39–47. <https://doi.org/10.1177/0956797614554955> (2015).
- DALL-E 2*. <https://openai.com/product/dall-e-2>. Accessed 27 Mar 2023.
- Groth, C., Tauscher, J.-P., Castillo, S., Magnor, M. *Altering the Conveyed Facial Emotion Through Automatic Reenactment of Video Portraits*. 128–135 [https://doi.org/10.1007/978-3-030-63426-1\\_14](https://doi.org/10.1007/978-3-030-63426-1_14) (2020).
- Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K. & Li, H. *Protecting World Leaders Against Deep Fakes*.
- Farid, H. Digital forensics in a post-truth age. *For. Sci. Int.* **289**, 268–269. <https://doi.org/10.1016/j.forsciint.2018.05.047> (2018).
- Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., & Guo, B. *Face X-Ray for More General Face Forgery Detection*. 5001–5010. [https://openaccess.thecvf.com/content\\_CVPR\\_2020/html/Li\\_Face\\_X-Ray\\_for\\_More\\_General\\_Face\\_Forgery\\_Detection\\_CVPR\\_2020\\_paper.html](https://openaccess.thecvf.com/content_CVPR_2020/html/Li_Face_X-Ray_for_More_General_Face_Forgery_Detection_CVPR_2020_paper.html). Accessed 24 June 2023 (2020).
- Wang, S.-Y., Wang, O., Zhang, R., Owens, A. & Efros, A. A. CNN-generated images are surprisingly easy to spot... for now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8692–8701. <https://doi.org/10.1109/CVPR42600.2020.00872> (2020).
- Farid, H. & Bravo, M. J. Perceptual discrimination of computer generated and photographic faces. *Digit. Invest.* **8**(3), 226–235. <https://doi.org/10.1016/j.diin.2011.06.003> (2012).



33. Holmes, O., Banks, M.S. & Farid, H. Assessing and improving the identification of computer-generated portraits. *ACM Trans. Appl. Percept.* **13**(2), 71–712 <https://doi.org/10.1145/2871714> (2016).
34. Mader, B., Banks, M. S. & Farid, H. Identifying computer-generated portraits: The importance of training and incentives. *Perception* **46**(9), 1062–1076. <https://doi.org/10.1177/0301006617713633> (2017).
35. Mustafa, M., Guthe, S., Tauscher, J.-P., Goesle, M. & Magnor, M. How human am I? EEG-based evaluation of virtual characters. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. 5098–5108 <https://doi.org/10.1145/3025453.3026043> (2017).
36. Bagdasarian, M.T., Hilsmann, A., Eisert, P., Curio, G., Müller, K.-R., Wiegand, T. & Bosse, S. EEG-Based Assessment of Perceived Realness in Stylized Face Images. In *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*. 1–4 <https://doi.org/10.1109/QoMEX48832.2020.9123145> (2020).
37. Crookes, K. *et al.* How well do computer-generated faces tap face expertise?. *PLOS ONE* **10**(11), e0141353. <https://doi.org/10.1371/journal.pone.0141353> (2015).
38. Barnett, B. O., Brooks, J. A. & Freeman, J. B. Stereotypes bias face perception via orbitofrontal–fusiform cortical interaction. *Soc. Cognit. Affect. Neurosci.* **16**(3), 302–314. <https://doi.org/10.1093/scan/nsaa165> (2021).
39. Freeman, J. B., Stolier, R. M. & Brooks, J. A. Dynamic interactive theory as a domain-general account of social perception. *Adv. Exp. Soc. Psychol.* **61**, 237–287. <https://doi.org/10.1016/bs.aesp.2019.09.005> (2020).
40. Oh, D. *Person Knowledge Shapes Face Identity Perception*.
41. Balkenius, C., & Johansson, B. Almost alive: Robots and androids. In *Frontiers in Human Dynamics*. Vol. 4. <https://doi.org/10.3389/fhumd.2022.703879>. Accessed 25 June 2023 (2022).
42. Ventura, P., Domingues, M., Ferreira, L., Madeira, M., Martins, A., Neto, M.L. & Pereira, M. *Holistic Word Processing is Involved in Fast Parallel Reading* (2019).
43. Weston, N. J. & Perfect, T. J. Effects of processing bias on the recognition of composite face halves. *Psychon. Bull. Rev.* **12**(6), 1038–1042. <https://doi.org/10.3758/BF03206440> (2005).
44. Brooks, J. & Freeman, J. *Psychology and Neuroscience of Person Perception*. <https://doi.org/10.1002/9781119170174.epcn413> (2018).
45. Hugenberg, K. & Corneille, O. Holistic processing is tuned for in-group faces. *Cognit. Sci.* **33**(6), 1173–1181. <https://doi.org/10.1111/j.1551-6709.2009.01048.x> (2009).
46. Ratner, K. G. & Amodio, D. M. Seeing “us vs. them”: Minimal group effects on the neural encoding of faces. *J. Exp. Soc. Psychol.* **49**(2), 298–301 <https://doi.org/10.1016/j.jesp.2012.10.017> (2013).
47. Fincher, K. M. Perceptual dehumanization of faces is activated by norm violations and facilitates norm enforcement. *J. Exp. Psychol. Gen.* <https://doi.org/10.1037/xge0000132> (2016).
48. Fincher, K. M., Tetlock, P. E. & Morris, M. W. Interfacing with faces: Perceptual humanization and dehumanization. *Curr. Dir. Psychol. Sci.* **26**(3), 288–293. <https://doi.org/10.1177/0963721417705390> (2017).
49. Momen, A., Hugenberg, K. & Wiese, E. *Robot Faces Engage Face-Typical Processing Less Strongly Than Human Faces*. In *Review*. Vol. 6 (2020).
50. Young, S. G., Slepian, M. L., Wilson, J. P. & Hugenberg, K. Averted eye-gaze disrupts configural face encoding. *J. Exp. Soc. Psychol.* **53**, 94–99. <https://doi.org/10.1016/j.jesp.2014.03.002> (2014).
51. *Download Inquisit 5 Player (Free)*. <https://www.millisecond.com/download/inquisitweb5>. Accessed 27 Nov 2023.
52. Ma, D. S., Correll, J. & Wittenbrink, B. The Chicago face database: A free stimulus set of faces and norming data. *Behav. Res. Methods* **47**(4), 1122–1135. <https://doi.org/10.3758/s13428-014-0532-5> (2015).
53. *ISHIGURO Symbiotic Human-Robot Interaction Project*. <https://www.jst.go.jp/erato/ishiguro/en/index.html>. Accessed 25 Mar 2019.
54. Stanislaw, H. & Todorov, N. Calculation of signal detection theory measures. *Behav. Res. Methods Instrum. Comput.* **31**(1), 137–149. <https://doi.org/10.3758/BF03207704> (1999).
55. Hautus, M. J. Corrections for extreme proportions and their biasing effects on estimated values of  $d'$ . *Behav. Res. Methods Instrum. Comput.* **27**(1), 46–51. <https://doi.org/10.3758/BF03203619> (1995).
56. Momen, A., Hugenberg, K. & Wiese, E. Robots engage face-processing less strongly than humans. *Front. Neuroergon.* <https://doi.org/10.3389/fnrgo.2022.959578> (2022).
57. Farah, M. J., Wilson, K. D., Drain, M. & Tanaka, J. N. What is “special” about face perception?. *Psychol. Rev.* **105**(3), 482–498. <https://doi.org/10.1037/0033-295X.105.3.482> (1998).
58. Hills, P. J. & Lewis, M. B. The development of face expertise: Evidence for a qualitative change in processing. *Cognit. Dev.* **48**, 1–18. <https://doi.org/10.1016/j.cogdev.2018.05.003> (2018).
59. Waytz, A., Gray, K., Epley, N. & Wegner, D. M. Causes and consequences of mind perception. *Trends Cognit. Sci.* **14**(8), 383–388. <https://doi.org/10.1016/j.tics.2010.05.006> (2010).
60. Powers, K. E., Worsham, A. L., Freeman, J. B., Wheatley, T. & Heatherton, T. F. Social connection modulates perceptions of animacy. *Psychol. Sci.* **25**(10), 1943–1948. <https://doi.org/10.1177/0956797614547706> (2014).
61. Deska, J. C. & Hugenberg, K. The face-mind link: Why we see minds behind faces, and how others’ minds change how we see their face. *Soc. Pers. Psychol. Compass* **11**(12), e12361. <https://doi.org/10.1111/spc3.12361> (2017).
62. Deska, J. C., Paige Lloyd, E. & Hugenberg, K. Facing humanness: Facial width-to-height ratio predicts ascriptions of humanity. *J. Pers. Soc. Psychol.* **114**(1), 75–94. <https://doi.org/10.1037/pspi0000110> (2018).
63. Haslam, N. Dehumanization: An integrative review. *Pers. Soc. Psychol. Rev.* **10**(3), 252–264. [https://doi.org/10.1207/s15327957psr1003\\_4](https://doi.org/10.1207/s15327957psr1003_4) (2006).
64. Wiese, E., Metta, G. & Wykowska, A. Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Front. Psychol.* <https://doi.org/10.3389/fpsyg.2017.01663> (2017).
65. Hugenberg, K., Young, S. G., Bernstein, M. J. & Sacco, D. F. The categorization-individuation model: An integrative account of the other-race recognition deficit. *Psychol. Rev.* **117**(4), 1168. <https://doi.org/10.1037/a0020463> (2010).
66. Almaraz, S. M., Hugenberg, K. & Young, S. G. Perceiving sophisticated minds influences perceptual individuation. *Pers. Soc. Psychol. Bull.* **44**(2), 143–157. <https://doi.org/10.1177/0146167217733070> (2018).
67. Epley, N., Akalis, S., Waytz, A. & Cacioppo, J. T. Creating social connection through inferential reproduction: Loneliness and perceived agency in gadgets, gods, and greyhounds. *Psychol. Sci.* **19**(2), 114–120. <https://doi.org/10.1111/j.1467-9280.2008.02056.x> (2008).
68. Epley, N., Waytz, A., Akalis, S. & Cacioppo, J. T. When we need a human: Motivational determinants of anthropomorphism. *Soc. Cognit.* **26**(2), 143–155. <https://doi.org/10.1521/soco.2008.26.2.143> (2008).
69. Bernard, P., Gervais, S. J., Allen, J., Campomizzi, S. & Klein, O. Integrating sexual objectification with object versus person recognition: The sexualized-body-inversion hypothesis. *Psychol. Sci.* **23**(5), 469–471. <https://doi.org/10.1177/0956797611434748> (2012).
70. Hugenberg, K. *et al.* The face of humanity: Configural face processing influences ascriptions of humanness. *Soc. Psychol. Pers. Sci.* **7**(2), 167–175. <https://doi.org/10.1177/1948550615609734> (2016).
71. Gray, H. M., Gray, K. & Wegner, D. M. Dimensions of mind perception. *Science* **315**(5812), 619–619. <https://doi.org/10.1126/science.1134475> (2007).
72. Jenny Xiao, Y., Coppin, G. & Van Bavel, J. J. Perceiving the world through group-colored glasses: A perceptual model of intergroup relations. *Psychol. Inquiry* **27**(4), 255–274. <https://doi.org/10.1080/1047840X.2016.1199221> (2016).

73. Krumhuber, E. G., Lai, Y.-K., Rosin, P. L. & Hugenberg, K. When facial expressions do and do not signal minds: The role of face inversion, expression dynamism, and emotion type. *Emotion* **19**(4), 746–750. <https://doi.org/10.1037/emo0000475> (2019).

### Author contributions

A.M. & E.W. conceived the idea for the study. A.M. collected data and analyzed data. A.M., E.W., and K.H. all helped interpret results and write main manuscript text. All authors reviewed the manuscript.

### Funding

Open Access funding enabled and organized by Projekt DEAL.

### Competing interests


The authors declare no competing interests.

### Additional information

**Correspondence** and requests for materials should be addressed to A.M. or E.W.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2024