Behavioral and neural evidence that robots are implicitly perceived as a threat

Zhengde Wei[1, +], Ying Chen[2, +], Jiecheng Ren[1], Yi Piao[1], Pengyu Zhang[1], Qian Zhao[1], Rujing Zha[1], Bensheng Qiu[3], Daren Zhang[1], Yanchao Bi[4], Shihui Han[5], Chunbo Li[6]*, Xiaochu Zhang[1, 2, 3, 7]*

1 School of Life Sciences, Division of Life Science and Medicine, University of Science & Technology of China, Hefei, Anhui 230027, China

2 School of Humanities & Social Science, University of Science & Technology of China, Hefei, Anhui 230026, China

3 Centers for Biomedical Engineering, School of Information Science and Technology, University of Science & Technology of China, Hefei, Anhui 230027, China

4 State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing 100875, China

5 School of Psychological and Cognitive Sciences, Peking University, Beijing, China

6 Shanghai Key Laboratory of Psychotic Disorders, Shanghai Mental Health Center, Shanghai Jiao Tong University School of Medicine, Shanghai 200030, China

7 Academy of Psychology and Behavior, Tianjin Normal University, Tianjin, 300387, China

+ These authors contributed equally to this work.

∗ Correspondence: licb@smhc.org.cn; zxcustc@ustc.edu.cn

Abstract

A deeper understanding of the human side of human-robot interaction determined by studying the human brain when we perceive robots should help solve the biggest challenges of successful social encounters with robots. However, current social neuroscience studies mainly focus on explicit perception of robots, and implicit perception of robots is rather unexplored. Here, our behavioral analysis indicated that despite self-reported positive attitudes, participants had negative implicit attitudes toward humanoid robots. Our neuroimaging analysis indicated that subthreshold presentation of humanoid robot vs. human images led to significant left amygdala activation that was associated with negative implicit attitudes toward robots. After successfully weakening the negative attitudes, the left amygdala response to subthreshold presentation of humanoid robot images decreased, and the decrease in the left amygdala response was positively associated with the decrease in negative attitudes. Our results reveal that the processing of information about humanoid robots displays automaticity with regard to the recruitment of amygdala activation. Our findings that people may implicitly perceive humanoid robots as a threat may guide more appropriate interaction with social robots.

Key words: robots; threat; implicit perception; amygdala

Significance statement

Social interactions with robots are one of the biggest challenges in robotics, which necessitates a deeper understanding of how people perceive robots. Our results reveal automaticity for processing information about humanoid robots similar to that previously evident for threats. Given the effort currently being put into the development of robots for daily assistance, studying implicit perception of robots could be a step toward building smooth human-robot social relationships.

Introduction

Given that robots are becoming increasingly present in the everyday environment and that social interactions with robots are one of the biggest challenges in robotics [1], a deeper understanding of how people perceive robots is necessary [2]. This is especially relevant given the effort currently being put into the development of robots for daily assistance [3, 4]. When a person perceives a social stimulus, information about that stimulus is immediately and spontaneously activated, including attitudes [5] and social stereotypes [6], and can consequently influence people's behavior [7]. Similarly, perceptions toward robots might automatically influence how we interact with robots. Studying the human brain with neuroimaging techniques when we perceive robots will shed light on a deeper understanding of the human side of human-robot interaction [8]. However, current social neuroscience studies mainly focus on explicit perception of robots [8, 9], and the implicit perception of robots is rather unexplored.

Researchers have proposed that our ideas about robots are mainly informed by science fiction media [10, 11], and robots in science fiction are generally described as a threat [10]. Recent statements from some influential industry leaders have strengthened these fears [12]. These ideas that robots tend to be dangerous entities that will threaten the survival of humanity might be rooted in our long-term socialization experiences, leading to the possibility that people might perceive robots as a threat. Empirical evidence has illustrated that people's explicit perception toward robots is inconsistent, ranging from enthusiasm to fear and anxiety [11, 13-18], possibly because people's explicit perceptions or attitudes are subject to recall bias and social desirability bias [19]. Notably, implicit perception is considered to be a highly stable evaluative representation stemming from long-term socialization experiences [20]. The most important feature is that implicit perception is irrespective of whether a person consciously considers the stimulus as good or bad [21]. Decades of work have identified implicit perception as having a crucial influence on thoughts and actions [22]. Here, we aimed to provide evidence from implicit behavioral and neural measures to support the possibility that people might implicitly perceive robots as a threat.

To accurately assess people's implicit perception of robots, it is important to implement implicit multimethod approaches. The implicit association test (IAT) is the most frequently used implicit approach [23] and is based on the response given by a participant when performing an association task [24]. IATs are commonly used to explore implicit attitudes toward sensitive issues, such as racism [25]. Furthermore, neuroimaging approaches might be more sensitive, as they reveal qualitative differences in processing without the need for distinct behavioral measures of implicit processing [22]. The preparedness model [26, 27] hypothesizes that the neural basis of automatic threat processing is the amygdala. The amygdala has long been known to play a key role in responding to emotionally relevant stimuli, activating in response to images containing threatening or highly arousing features [28, 29]. When automatic and controlled evaluations of threats differ, more positive controlled processing can moderate more negative automatic processing [30], which could account for the absence of significant activation in the amygdala in response to suprathreshold presentations of threats [31, 32]. Notably, although controlled processing can eliminate amygdala activation caused by consciously presented threats, the amygdala still showed significant activation caused by threats that were presented unconsciously [30-32]. We hypothesized that if humanoid robots are implicitly perceived as a threat, then participants would be expected to show negative implicit attitudes toward robots and an amygdala response to subthreshold presentation of robot images.

How people implicitly perceive humanoid robots is an intriguing question. People's perception of robots relies on the robots' behaviors, abilities and appearance [33]. More human characteristics are consistently perceived in robots that look and act like humans [34]. Here, we tested whether people implicitly perceive humanoid robots as a threat. We pursued this topic by first measuring participants' explicit and implicit attitudes toward humanoid robots. We then employed functional magnetic resonance imaging (fMRI) and intervention to test the amygdala response to humanoid robots and its causal relationship to implicit attitudes.

Material and Methods

Participants. One hundred fifty-two participants (age: 21.75±2.58 years; 61 females) took part in the original implicit association test (IAT), 22 participants (age: 22.77±1.97 years; 14 females) performed the Black IAT and the mixed IAT, 30 participants (age: 22.23±1.89 years; 20 females) performed the positive adjective IAT and the available humanoid IAT, and 38 participants (age: 21.77±1.67 years; 23 females) performed the humanoid weapon IAT and pet-robot weapon IAT. Ninety-nine participants took part in the intervention experiment. Among them, 69 participants were allocated to the positive intervention group, and 30 participants were allocated to the neutral intervention group. Ten participants in the positive intervention group were excluded because of excessive head movement during scanning (TRs with motion over 0.3 mm were censored, and participants who had more than 15% of their TRs censored were removed from further analysis); 36 participants discontinued the positive intervention after completing preintervention tasks due to personal reasons; and 3 participants discontinued the neutral intervention after completing the preintervention tasks because of withdrawal. Finally, data from 23 participants (age: 20.39±2.06 years; 15 females) in the positive intervention group and 27 participants (age: 22.13±1.49 years; 18 females) in the neutral group were analyzed. All participants were right-handed, with normal or corrected-to-normal vision, without alcohol or drug dependence disorders and without prior head injury. All participants and their immediate relatives had no psychiatric disorders or history of psychiatric disorders. All participants were nonpsychology undergraduates, and they were unaware of the purpose and contents of the study before they participated in the experiment.

The sample size was determined using a medium effect size (Cohen's d = 0.5) [35]. Power analyses using G*power suggested a sample size of 34 for 0.8 power. This sample size was implemented for all tasks except for the evaluating conditioning task. We performed power analyses using G*power with a priori effect size (Cohen's d = 0.55) and 0.8 power to determine the sample size for a one-tailed paired-sample t test, resulting in a sample size of 22 for the evaluating conditioning task. The study was approved by the Research Ethics Committee of the University of Science and Technology of China (NO. 2020-N(H)-099), and written informed consent was obtained from all participants consistent with the Declaration of Helsinki. The methods were carried out in accordance with the approved guidelines.

Robot-related questionnaires and scales

Participants were asked to rate their perceptions toward humanoid robots on a 5-point questionnaire (1-completely disagree, 5-completely agree; Table 1). Items 4, 5 and 6 were reversed items. The sum score of these six items indicates participants' opinion toward humanoid robots. We also obtained participants' explicit attitudes toward robots using the Negative Attitudes Toward Robots Scale (NARS) [36].

We determined participants' prior experience with robots by asking about the frequency of daily interactions with robots and which of 20 fictional films (*i.e.*, *Wall-E*, *The Terminator*, *Alita: Battle Angel*, *I, Robot*, *AI: Artificial Intelligence*, *Chappie*, *Bicentennial Man*, *RoboCop*, *Short Circuit*, *Ex Machina*, *Edward Scissorhands*, *Blade Runner*, *My Robot Girlfriend*, *Autómata*, *Real Steel*, *2001: A Space Odyssey*, *Moon*, *Star Wars*, *The Surrogates*, *Forbidden Planet*) portraying robots they have seen.

Original implicit association test. We assessed participants' automatic attitude using a standard implicit attitude measure paradigm — the IAT task [37, 38]. In the original IAT (Fig. 3a), stimuli consisted of attribute words and concept images. The word stimuli were six adjectives, including three adjectives—threatening, strikes and lethal—that indicate "threatening" and three adjectives—neutral, normal and average—that indicate "nonthreatening" [39]. The image stimuli were 20 images of fictional humanoid robots and 20 images of Caucasian individuals (Caucasian individuals were used because the fictional humanoid robot forms were Caucasian; we matched the action of each pair of images, and the gender ratio of fictional humanoid robots and humans was the same; the stimuli are available at https://rec.ustc.edu.cn/share/1f97a050-7745-11ec-bbe8-61ef14e3189d). To avoid confounding of character stereotypes, stimuli from the movies were excluded. These images were downloaded from the open source on the internet. Adobe Photoshop CS5 was used for all images to uniformly process the pixels and colors of the images so that the materials were consistent and the pixel size was 300×300. Before the formal experiment, participants were presented with an image set including all 20 images of fictional humanoid robots and 20 images of Caucasian individuals in a pilot experiment. For each image, participants provided a rating of affect valence (1 = extremely negative; 5 = extremely positive), arousal (1 = extremely calm; 5 = extremely aroused) and novelty (1 = not novel at all, 9 = extremely novel). The results showed that the affect valence and arousal for fictional humanoid robot images were neutral (valence: mean = 3.05, SD = 0.69; arousal: mean = 2.94, SD = 0.90). The affect valence and arousal for Caucasian images were neutral as well (valence: mean = 3.67, SD = 0.66; arousal: mean = 3.20, SD = 1.03). There was a significant difference in novelty between fictional humanoid robot images and Caucasian images (fictional humanoid robot: mean = 5.64, SD = 0.362; Caucasian: mean = 3.95, SD = 0.31; t24 = 16.99, p < 0.001).

The original IAT involved a series of five discrimination blocks.

Block 1 was a concept discrimination task. In this block, the concept labels (humanoid or human) were displayed on the top left and right sides of the screen, and participants were told to categorize the concept images (humanoid robot images or human images) by pressing a corresponding button. Each trial began with the presentation of a fixation cross for 1 s followed by a target image remaining on the screen until the participant responded or for a maximum of 3 s. There were 40 trials in this block, and each concept image was presented 1 time. The order of stimulus presentation was pseudorandomized across participants in all blocks. The label positions (left or right) of the concepts (humanoid or human) were counterbalanced between participants.

Block 2 was an attribute discrimination task. The participants were required to categorize the attribute words according to the attribute labels (threat or nonthreat), and the experimental procedure was consistent with the concept discrimination task. There were 18 trials in this block, and each attribute word was repeated 3 times.

Block 3 was a compatible combined task (humanoid + threat; human + nonthreat). In this block, humanoid robot images and threatening words shared the same response, and human images and nonthreatening words shared the same response. There were 76 trials in this block; each concept image was presented 1 time, and each attribute word was repeated 6 times.

Block 4 was a reversed attribute discrimination task. Participants performed the same task as in block 2, but the positions of the attribute labels (threatening words or nonthreatening words) were reversed. There were 18 trials in this block, and each attribute word was repeated 3 times.

Block 5 was an incompatible combined task. Participants performed the same task as in block 3, but the position of the attribute labels (threat or nonthreat) were swapped (humanoid + nonthreat; human + threat). There were 76 trials in this block; each concept image was presented 1 time, and each attribute word was repeated 6 times.

Brief instructions were presented on the screen at the beginning of each block until the participant responded.

Before analyzing the data from the IAT, the following data reduction procedure was applied [37, 38]: 1) eliminate data from subjects for whom more than 10% of trials have a latency shorter than 300 ms; 2) eliminate data from subjects for whom more than 20% of trials are incorrect; 3) compute the mean latencies for only correct trials for blocks 3 and 5; 4) compute the pooled standard deviation for all trials in blocks 3 and 5; 5) replace each incorrect trial's latency in blocks 3 and 5 with the block mean computed in step 3 plus 600 ms; 6) compute the difference between the mean latencies of blocks 3 and 5 using paired sample t test; 7) divide the difference computed in step 6 by the pooled standard deviation computed in step 4 to calculate the IAT effect size - the IAT score.

Black IAT. The procedure of the Black IAT was the same as that of the original IAT except the 20 images of Caucasian individuals were replaced with 20 images of Black individuals.

Mixed IAT. The procedure of the mixed IAT was the same as that of the original IAT except the 20 images of Caucasian individuals were replaced with 20 images of Black, Caucasian and Asian individuals.

Available humanoid IAT. The procedure of the available humanoid IAT was the same as that of the original IAT except the 20 images of fictional humanoid robots were replaced with 20 images of the currently available humanoid robots (*e.g.*, Pepper, Nao, and iCub).

Positive adjective IAT. The procedure of the positive adjective IAT was the same as that of the original IAT except the threat-related adjectives were replaced with positive adjectives (*i.e.*, salient, arousing, and affective).

The humanoid weapon IAT. The procedure of the humanoid weapon IAT was the same that of as the original IAT except the 20 images of Caucasian individuals were replaced with 20 images of weapons and the concept label of "human" was replaced with "weapon".

Pet-robot weapon IAT. The procedure of the pet-robot weapon IAT was the same as that of the humanoid weapon IAT except the 20 images of fictional humanoid robots were replaced with 20 images of pet robots.

Procedures in the intervention experiment. In this experiment, participants performed the original IAT before and after the intervention. In the positive intervention group, the positive evaluative conditioning task was used to intervene with implicit attitudes, whereas the neutral evaluative conditioning task was used in the neutral intervention group. In the positive intervention group, participants completed a backward masking task with fMRI scanning and a forced-choice detection task before intervention and then completed the backward masking task with fMRI scanning again after intervention (Fig. 2c). In the neutral intervention group, participants completed the positive adjective IAT and the available humanoid IAT before intervention. The implicit association test and evaluating conditioning task were presented by E-Prime 2.0, and the backward masking task and forced-choice detection task were presented by MATLAB (R2015a).

Materials in the fMRI experiment. The materials of the backward masking task and forced-choice detection task consisted of 20 images of fictional humanoid robots and 20 images of Caucasian individuals, the same images used in the original IAT. The materials of the evaluative conditioning task consisted of conditioned stimuli (CSs), unconditioned stimuli (USs), target stimuli and neutral fillers (Table 2). The CSs consisted of 20 images of fictional humanoid robots and 20 images of Caucasian individuals. Approximately half of the USs, target stimuli and neutral fillers were images, and half were words. We selected the US images, target images and neutral filler images from the Chinese Affective Image System (CAPS), including 3 target images, 5 positive images (USs), 5 neutral images (USs), and 16 neutral filler images. Adobe Photoshop CS5 was used for all images to uniformly process the pixels and colors of the images so that the materials were consistent and the pixel size was 300×300. The valence, arousal and dominance scores for these 29 images are displayed in Table 3 (on a scale from 1 = low valence/arousal/dominance to 9 = high valence/arousal/dominance). The word stimuli–3 target adjectives, 5 positive adjectives (USs) [40], 5 neutral adjectives (USs) [41] and 17 neutral filler nouns [41]–were selected from the pilot study. We translated these words into Chinese characters by using Collins COBUILD Advanced Learner's English-Chinese Dictionary [42]. The stimuli for this task are shown in Table 4.

Backward masking task. Participants completed the backward masking task in the MRI scanner (Fig. 2a-b) [43]. For the subthreshold presentation, the target image was presented for 17 ms followed by a mask for 183 ms and a fixation cross for 1800 ms. For the suprathreshold presentation, the target image was presented for 200 ms followed by a fixation cross for 1800 ms. These images were arranged in a block design consisting of 10 images (either humanoid robot or human) in a computer-generated pseudorandom order. To avoid any possible effects on subthreshold presentation, the six suprathreshold blocks (three humanoid blocks and three human blocks) were presented after the six subthreshold blocks (three humanoid blocks and three human blocks). Except for the first and last baseline blocks (a fixation cross displayed on the screen) lasting for 10 s, each target block was separated by a 20 s baseline block.

Forced-choice detection task. To confirm that participants were aware of the stimulus during suprathreshold presentation but not during subthreshold presentation in the backward masking task, a forced-choice detection task was used. The forced-choice detection task consisted of 80 trials; the first half were subthreshold presentations, and the second half were suprathreshold presentations. The stimuli were presented similarly to those in the backward masking task. The difference was that a 2000 ms forced-choice phase followed the target stimuli. Participants were informed that the target stimulus could have humanoid or human images and were told to recognize the content of each image. Data from two participants in this task were excluded because of program crashes. We compared the accuracy, response rate, and reaction time between suprathreshold and subthreshold presentations separately using a paired sample t test. We also compared the accuracy for both conditions to chance level (50%) using a one-sample t test.

Positive evaluative conditioning task. The evaluative conditioning task is a classic paradigm of changing implicit attitudes by pairing CSs and USs [44]. In the positive evaluative conditioning task, the participants' implicit negative attitude toward humanoid robots may be reduced by pairing the humanoid robot images (CSs) with the positive stimuli (USs) (Fig. 4b). As a control condition, the human images (CSs) were paired with neutral stimuli (USs). Participants were unaware of the repeated conditioned stimulus–unconditioned stimulus (CS-US) [45].

This task included 6 blocks of 61 trials each. For each block, the specific stimuli were arranged, and all stimuli were presented in pseudorandom order. All stimuli appeared for 1.5 s each. During the experiment, the participants were instructed to view a stream of images or words and respond as soon as possible whenever a prespecified target image or words appeared. To ensure that the participants carefully viewed the US-CS pairs during the entire experiment, the participants were told that the accuracy rate had to be at least 95% after the end of the experiment or they would have to restart the task. Before the formal evaluative conditioning task, a training evaluative conditioning task was performed.

Neutral evaluative conditioning task. The procedure of the neutral evaluative conditioning task was the same as that for the positive evaluating conditioning task except that the humanoid robot and human images were both paired with the neutral stimuli.

MRI acquisition. Gradient echo-planar imaging data were acquired using a 3.0 T GE Discovery MR750 with a circularly polarized head coil at the Information Science Center of the University of Science and Technology of China. A T2*-weighted echo-planar imaging sequence (FOV = 240 mm, TE = 30 ms, TR = 2000 ms, flip angle = 85°, matrix = 64 × 64) with 33 axial slices (no gaps, voxel size: 3.75 × 3.75 × 3.7 mm^3) covering the whole brain was used to acquire the functional MR images. High-resolution T1-weighted spin□echo imaging data were also acquired for anatomical overlays, and three-dimensional gradient-echo imaging data were acquired for stereotaxic transformations after functional scanning. Before entering the scanner, participants were instructed to keep their heads still during all scans. Participants were placed in a light head restraint within the scanner to limit head movement. Visual stimuli were projected on a screen and viewed through a mirror attached to the head coil. During the backward masking task, 4 functional scan runs occurred, each lasting 4 min. There was an interval of approximately 1 min between every two runs.

fMRI processing. Functional data were analyzed using the Analysis of Functional NeuroImages (version: AFNI_21.01.01) software. Time series were realigned to the second volume. The realigned images were normalized to the Talairach coordinate. Raw data were corrected for temporal shifts between slices and for motion (TRs with motion over 0.3 mm were censored, and participants who had more than 15% of their TRs censored were removed from further analysis), spatially smoothed with a Gaussian kernel (full width at half maximum = 8 mm), and temporally normalized (for each voxel, the signal of each volume was divided by the temporally averaged signal). High-pass temporal filtering (using a filter width of 128 s) was also applied to the data.

To elucidate neural responses that correlated with humanoid robot images and human images under subthreshold and suprathreshold conditions, a general linear model (GLM) was used. Regressors of interest were subthreshold humanoid blocks, subthreshold human blocks, suprathreshold humanoid blocks and suprathreshold human blocks. The regressor of no interest was the fixation block. These regressors were convolved with a hemodynamic response function (HRF) and simultaneously regressed against the blood oxygenation level-dependent (BOLD) signal in each voxel. The regressors were not orthogonalized, and there was no significant collinearity among the regressors. Six regressors for head motion were also included. Individual contrast images were analyzed for each regressor of interest's responses using one-sample t tests to generate statistical maps.

Region of interest (ROI) analysis. We conducted ROI analyses on the bilateral amygdala. The ROIs for the bilateral amygdala were identified from the AAL atlas [46]. We determined parameter estimates for each participant from the local average in a mask back-projected from the ROIs. The differential neural responses between the humanoid robot and human conditions in the ROIs were analyzed. Correlation analysis was performed between neural responses and implicit attitudes.

Results

Negative implicit attitudes toward humanoid robots. In the attitudes toward robots questions, participants showed a positive explicit attitude toward humanoid robots ($t_{65} = 8.84$, $p < 0.001$, Cohen's $d = 1.10$). In the original IAT, participants had significant IAT scores (baseline = 0.2; mean = 0.46; SD = 0.48; $t_{151} = 6.79$, $p < 0.0001$, Cohen's $d = 0.54$; Fig. 3b). Participants responded faster to combinations of "humanoid + threatening" and "human + nonthreatening" than to combinations of "humanoid + nonthreatening" and "human + threatening" ($t_{151} = -9.64$, $p < 0.0001$, Cohen's $d = 0.79$; Fig. 3b). Previous studies have indicated that participants showed implicit negative attitudes toward Black individuals relative to Caucasian individuals [30, 47]. We wondered whether the automatic negative associations to humanoid robots compared with Caucasian individuals would still survive when compared with a biased race (*i.e.*, Black individuals) and could be generalized to common races. Thus, we also recruited another group of participants to perform the Black IAT and the mixed IAT. Consistent results were found. Participants responded faster in the compatible task than in the incompatible task in the Black IAT (RT: $t_{21} = -6.40$, $p < 0.001$, Cohen's $d = 1.36$; mean IAT score = 0.61; Fig. 3c) and mixed IAT (RT: $t_{21} = -7.33$, $p < 0.001$, Cohen's $d = 1.57$; mean IAT score = 0.60; Fig. 3c). Our IAT results indicated that, despite the self-reported findings from the questionnaires, participants have negative implicit attitudes toward humanoid robots.

In the positive adjective IAT, participants had no significant IAT scores (baseline = 0.2; mean = 0.17; SD = 0.37; $t_{29} = -0.45$, $p = 0.65$). Participants did not respond faster to combinations of "humanoid + positive" and "human + neutral" than to combinations of "humanoid + neutral" and "human + positive" ($t_{29} = -1.30$, $p = 0.21$). These results indicated that humanoid robots were not implicitly more associated with positive words. In the available humanoid IAT, participants had significant IAT scores (baseline = 0.2; mean = 0.45; SD = 0.62; t_{29} = 2.17, p = 0.038). Participants responded faster to combinations of "humanoid + threatening" and "human + nonthreatening" than to combinations of "humanoid + nonthreatening" and "human + threatening" (t_{29} = -3.62, p = 0.001, Cohen's d = 0.66). These results indicated that participants also have negative implicit attitudes toward currently available humanoid robots. The more fictional films the participants had seen, the less explicitly negative attitudes toward robots they had (r = -0.439, p = 0.015). However, no significant relationships between participants' prior experience with robots and implicit attitudes toward robots (neither fictional humanoid robots nor currently available humanoid robots) were observed (all p > 0.60). Our results indicated that people's prior experience with robots might influence their explicit perception of robots rather than their implicit perception of robots.

Humanoid robots may be considered more competitive with humans than pet robots are, resulting in more negative implicit attitudes toward humanoid robots than toward pet robots. To test the possible influence of the robot's appearance on implicit attitudes, we focused on the implicit attitude differences between the humanoid and pet robots by using a well-known threatening stimulus (*i.e.*, weapons [48]) as a baseline. Our results indicated that participants displayed larger IAT scores in the humanoid weapon IAT than in the pet-robot weapon IAT (t_{37}= 3.07, p < 0.01, Cohen's d = 0.50; Fig. 4a). These results indicated that participants have a more negative implicit attitude toward humanoid robots than toward pet robots.

Greater left amygdala activity was induced by humanoid robot images than by human images under subthreshold presentation. As shown in Fig. 4c, the mean response rate in the forced-choice detection task was more than 90% under both presentations and was higher under suprathreshold presentation (t_{58}= 4.14, p < 0.001, Cohen's d = 0.54) than under subthreshold presentation. The response time under suprathreshold presentation was significantly shorter than that under subthreshold presentation (t_{58}= -10.91, $p < 0.001$, Cohen's d = 1.42). The accuracy under suprathreshold presentation was significantly higher than that under subthreshold presentation (t_{58}= 16.68, $p < 0.001$, Cohen's d = 2.17). Importantly, the accuracy under subthreshold presentation did not differ from random chance (t_{58}= -0.31, p = 0.76), whereas the accuracy under suprathreshold presentation was higher than random chance (t_{58}= 19.13, $p < 0.001$, Cohen's d = 2.49). These findings indicate that participants are aware of the stimuli under suprathreshold presentation but are unaware of the stimuli under subthreshold presentation.

In fMRI analysis, we first tested whether subthreshold presentation of images of humanoid robots leads to a greater amygdala (Fig. 5a) response compared to images of humans. Activation in response to humanoid robot images was significantly stronger than activation in response to human images in the anatomically defined left amygdala under subthreshold presentation (left amygdala: t_{58} = 2.61, p = 0.012, Cohen's d = 0.34; right amygdala: t_{58} = 0.26, p = 0.80; Fig. 5b); no such difference was observed under suprathreshold presentation (left amygdala: t_{58} = -0.76, p = 0.37; right amygdala: t_{58} = 0.26, p = 0.80). The activation differences between humanoid robot and human images under subthreshold presentation in the left amygdala were significantly stronger than activation differences under suprathreshold presentation (left amygdala: t_{58} = 2.23, p = 0.03, Cohen's d = 0.30; right amygdala: t_{58} = 1.32, p = 0.19; Fig. 5b). Importantly, a greater IAT score was associated with greater left amygdala activity under the subthreshold condition (r = 0.46, $p < 0.001$; Fig. 5c). Our results indicated that greater left amygdala activity induced by humanoid robot images compared to that induced by human images under subthreshold presentation is associated with negative implicit attitudes toward humanoid robots.

We found that there was a significant difference in novelty between humanoid robot images and human images. It has been reported that novelty contributes to amygdala activation [49]. We controlled for the novelty of images by adding this as a covariate in our fMRI general linear model analyses, and this did not alter the results related to amygdala activation (Fig. 6a-b).

The left amygdala response to subthreshold presentation of humanoid robot images changes after successfully weakening negative attitude. We conducted a two-factor (group factor: positive intervention group, neutral intervention group; time factor: pretest, posttest) repeated-measures ANOVA on IAT scores. A significant group×time interaction effect was found ($F_{(1,48)} = 4.99$, $p = 0.038$, $\eta^2 = 0.094$, Fig. 4d). *Post hoc* analysis revealed that the posttest IAT scores were significantly smaller than the pretest scores ($t_{22} = -2.16$, $p = 0.042$, Cohen's $d = 0.45$; Fig. 5d) in the positive intervention group but not in the neutral intervention group ($t_{26} = -0.88$, $p = 0.39$), indicating that participants' negative implicit attitudes toward humanoid robots had been successfully weakened by the positive evaluative conditioning task. We tested whether the left amygdala response to subthreshold presentation of humanoid robot images changed after successfully weakening negative implicit attitudes. We conducted a two-factor (time factor: before modulation, after modulation; presentation factor: subthreshold presentation, suprathreshold presentation) repeated-measures ANOVA on activation differences between humanoid robot and human images in the left amygdala. A significant time×presentation interaction effect was found ($F_{(1,22)} = 8.51$, $p = 0.008$), but no significant time or presentation main effects were found (all $p > 0.29$). *Post hoc* analysis revealed that there was a marginally significant decrease in left amygdala activation between humanoid robot and human images under the subthreshold presentation between the posttest and pretest scans ($t_{22} = -2.02$, $p = 0.056$, Cohen's $d = 0.38$; Fig. 5e). Correlation analysis revealed that there was a significant correlation between IAT score changes and activation value changes in the left amygdala under subthreshold presentation ($r = 0.58$, $p = 0.004$; Fig. 5f). These results demonstrated a causal relationship between implicit attitudes toward humanoid robots and the left amygdala response to subthreshold presentation of humanoid robot images. Similar to the previous section, after controlling for the novelty of images by adding it as a covariate in our fMRI general linear model analyses, the results did not change (Fig. 6c-d).

Discussion

Despite self-reported positive attitudes in the questionnaires, participants had negative implicit attitudes toward humanoid robots. The left amygdala response to subthreshold presentation of humanoid robots, which was positively associated with implicit attitudes toward humanoid robots, indicates the automatic and quick detection of humanoid robots. After successfully weakening negative attitudes, the decrease in IAT scores was positively associated with the decrease in activation in the left amygdala under subthreshold presentation, demonstrating a causal relationship between implicit attitudes toward humanoid robots and the left amygdala response to subthreshold presentation of humanoid robot images.

Our results provide evidence that humanoid robots may be implicitly perceived as a threat. People detect and respond rapidly to threatening stimuli [50, 51]. Nonhuman primates also respond more rapidly to threatening stimuli (at least some types) than to neutral stimuli [52, 53]. The amygdala, a subcortical structure in the anterior temporal lobe, is located in an evolutionarily old part of the brain and is shared by other mammals. It is assumed to be the neural basis of the hardwired "fear module" that allows us to automatically and quickly detect threatening stimuli [27]. Studies have documented that the amygdala responds selectively to threats, sometimes irrespective of the affective valence while using implicit measures, such as animate entities [54-56] and depictions of humans [57-59]. Our results showing no amygdala activity in response to suprathreshold presentation of humanoid robot images seemingly support that people did not perceive humanoid robots as a threat. However, when automatic and controlled evaluations of threat differ, more positive controlled processing can moderate more negative automatic processing [30]. With the positive explicit attitudes toward humanoid robots found in the present study, the absence of amygdala response to suprathreshold presentation of humanoid robot images is understandable. Interestingly, although controlled processing can eliminate amygdala activity caused by consciously presented threats, the amygdala still shows greater responses to threats that are presented unconsciously [30-32]. Several studies suggest that the left amygdala might be specifically involved in the processing of facial stimuli [60, 61]. It has also been reported that the left amygdala shows less habituation to fearful stimuli than the right amygdala, which might make it more likely to capture the blood oxygen level-dependent changes in this area [62, 63]. However, the lateralization of amygdala activation is still controversial [64, 65]. Our present study found that greater left amygdala activity was induced by humanoid robot images than by human images under subthreshold presentation. After successfully weakening negative attitudes, the decrease in IAT scores was positively associated with the decrease in activation in the left amygdala under subthreshold presentation. Our results indicate a causal relationship between implicit attitudes toward humanoid robots and the left amygdala response to subthreshold presentation of humanoid robot images. Note that this result did not change after controlling for the novelty of the humanoid robot and human images. These results potentially reflect the automaticity with rapid recruitment of the amygdala for humanoid robot-related stimuli processing.

Except for threatening stimuli, many studies have indicated that the amygdala plays an important role in emotional processing, especially negative emotions, including disgust, sadness and pain [66, 67]. The amygdala is also activated by stimuli involving social information [68, 69]. Thus, some might argue that the amygdala response to subthreshold presentation of humanoid robots in this study could be explained by factors other than threat or fear. In this study, we used humanoid robots with neutral faces and humans with neutral faces as control stimuli, which might eliminate the confounding factors of social information, such as facial expressions and emotions. Although there was a significant difference in novelty between humanoid robot and human images, the amygdala-related results were not altered after controlling for novelty. Furthermore, salience is also a possibility, and consequently, humanoid robots should be implicitly more associated with other salient adjectives (*e.g.*, positive words). However, our results of the positive adjective IAT were not significant, indicating that humanoid robots were not implicitly more associated with salient adjectives. Of note, the significantly positive correlation between the decrease in the left amygdala response and the decrease in negative attitudes indicates that there is a modulation effect of the positive evaluating conditioning task on amygdala activity, although the reduced saliency of stimuli might also contribute to the decrease in amygdala activity.

People perceive robots based upon context, cues, and cultural assumptions [33]. Our ideas about robots are mainly informed by science fiction media [10], and robots in science fiction are generally described as a super species with greater intelligence than that of humans, attempting to eliminate humanity[10]. The concept that fully autonomous robots are dangerous competitive “living” entities that will threaten the survival of humanity gradually becomes a deep impression. Two species with the closest living habits have the strongest competitiveness between them [70, 71], such as Neanderthal extinction by competition with anatomically modern humans [72-74]. Combined with our results, robots with human-like faces and intelligence are plausibly implicitly perceived as a new species or even race. Consequently, the fear of humanoid robots is likely similar to threat stimuli related to survival in evolutionary history. However, whether people perceive humanoid robots as an evolutionary threat is equivocal and is worth future research to uncover.

A growing body of evidence has indicated that the physical appearance of a robot has a strong impact on people’s perception [9, 75, 76]. However, our results suggest that a robot’s appearance is likely to influence explicit perception rather than implicit perception (at least the implicit attitudes toward robots). Fictional humanoid robots with highly anthropomorphic appearances and the currently available humanoid robots with less anthropomorphic appearances were used in our experiment. Of note, participants showed negative implicit attitudes toward both kinds of humanoid robots. Consistent with our results, previous studies have shown negative implicit attitudes toward robots by using robot silhouettes [77, 78]. Taken together, the physical appearance of a robot plausibly has a weak impact on people's implicit perception.

Our demonstration that people implicitly perceive humanoid robots as a threat might contribute to some negative biases for robots. Consistent with previous studies [77, 78], we found that people have negative implicit attitudes toward robots. A study suggests that the early top-down process of empathy is weaker for humanoid robots than for humans [79]. A systematic review [80] of anxiety and acceptance toward social robots reveals that people only slightly accept social robots and feel slightly anxious about them in general. It might be that the negative perception toward robots immediately and spontaneously activates negative attitudes and social stereotypes toward robots and consequently causes biased behaviors. The negative perception toward robots would be detrimental to the development of robots and successful social encounters with robots. Of note, using an evaluative conditioning task seems to weaken this negative perception toward humanoid robots. More research should address the issue of where negative perception toward humanoid robots comes from and how to modulate this negative perception.

In previous studies, a particular robot-related experience might have a significant influence on people's explicit perception of robots [36, 81, 82]. Consistent with these findings, our study found that the more fictional films portraying robots the participants had seen, the less explicitly negative attitudes toward robots they had. However, no significant relationships between participants' prior experience with robots and implicit attitudes toward robots were observed. Our results indicated that people's prior experience with robots might influence their explicit perception of robots rather than their implicit perception of robots. Thus, future research addressing the question of the impact of a priori experience on perception of robots might distinguish between explicit and implicit perception.

One limitation was that the humanoid robot stimuli were only in the form of images. Although fictional and currently available humanoid robot images were used in the present study, the ecological effect of researching human-robot interactions is somewhat weak. It is better to investigate how people perceive and interact with robots in a socially dynamic environment. Thus, stimuli in the form of videos and human-robot interaction research in the real world with the help of mobile neuroimaging should be considered in future studies.

Conclusions

In summary, this study demonstrates that humanoid robots are implicitly perceived as a threat. This sheds light on how people perceive and interact with robots that are increasingly entering our social environment. The future of social robots is undeniably exciting, and insights from neuropsychology research will guide the future direction of robot development and bring us closer to interacting with social robots.

Acknowledgments

This work was supported by grants from the National Key Basic Research Program (2018YFC0831101), the National Natural Science Foundation of China (71942003, 31771221, 61773360, 32100886, and 71874170), the Major Project of Philosophy and Social Science Research, Ministry of Education of China (19JZD010), the CAS-VPST Silk Road Science Fund 2021 (GLHZ202128), the Collaborative Innovation Program of Hefei Science Center, CAS (2020HSC-CIP001), and the China Postdoctoral Science Foundation (2016M592051). A portion of the numerical calculations in this study were performed with the supercomputing system at the Supercomputing Centre of USTC.

Author contributions

ZDW, YC, and XCZ conceived and designed the study. ZDW and YC obtained the findings. YC was responsible for the acquisition of data. ZDW and YC analyzed and interpreted the data. PYZ, YP, JCR, QZ, RJZ, BSQ, YCB, SHH, CBL and DRZ provided administrative, technical, or material support. CBL and XCZ supervised the study. ZDW and YC drafted the paper, and all authors contributed to critical revision for intellectual content.

Conflicts of interest

The authors declare no conflicts of interest.

Data and materials availability statement

The data and scripts are available at https://rec.ustc.edu.cn/share/1f97a050-7745-11ec-bbe8-61ef14e3189d.

References

1. Yang, G.Z., *et al.* The grand challenges of Science Robotics. *Science robotics* **3** (2018).
2. Bossi, F., *et al.* The human brain reveals resting state activity patterns that are predictive of biases in attitudes toward robots. *Science robotics* **5** (2020).
3. Scassellati, B., Admoni, H. & Mataric, M. Robots for Use in Autism Research. *Annu Rev Biomed Eng* **14**, 275-294 (2012).
4. Broadbent, E. Interactions With Robots: The Truths We Reveal About Ourselves. *Annual review of psychology* **68**, 627-652 (2017).
5. Fazio, R.H., Jackson, J.R., Dunton, B.C. & Williams, C.J. Variability in automatic activation as an unobtrusive measure of racial attitudes: a bona fide pipeline? *J Pers Soc Psychol* **69**, 1013-1027 (1995).
6. Hills, P.J., Lewis, M.B. & Honey, R.C. Stereotype priming in face recognition: Interactions between semantic and visual information in face encoding. *Cognition* **108**, 185-200 (2008).
7. Ferguson, M.J. & Bargh, J.A. How social perception can automatically influence behavior. *Trends in cognitive sciences* **8**, 33-39 (2004).
8. Henschel, A., Hortensius, R. & Cross, E.S. Social Cognition in the Age of Human-Robot Interaction. *Trends in neurosciences* **43**, 373-384 (2020).
9. Cross, E.S. & Ramsey, R. Mind Meets Machine: Towards a Cognitive Science of Human-Machine Interactions. *Trends in cognitive sciences* **25**, 200-212 (2021).
10. Cave, S. & Dihal, K. Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence* **1**, 74-78 (2019).
11. Broadbent, E., *et al.* Attitudes and reactions to a healthcare robot. *Telemedicine journal and e-health : the official journal of the American Telemedicine Association* **16**, 608-613 (2010).

12. Preparing for the Future of Artifcial Intelligence (Executive Ofce of the President National Science and Technology Council.

13. Timo, G. & Markus, A. Are robots becoming unpopular? Changes in attitudes towards autonomous robotic systems in Europe. *Computers in human behavior* **93**, 53-61 (2019).

14. Baobao, Z. & Dafoe, A. Artificial Intelligence: American Attitudes and Trends. *Oxford, UK: Center for the Governance of AI, Future of Humanity Institute, University of Oxford* (2019).

15. Liang, Y.H. & Lee, S.A. Fear of Autonomous Robots and Artificial Intelligence: Evidence from National Representative Data with Probability Sampling. *International Journal Of Social Robotics* **9**, 379-384 (2017).

16. Fast, E. & Horvitz, E. Long-Term Trends in the Public Perception of Artificial Intelligence. *Thirty-First Aaai Conference on Artificial Intelligence*, 963-969 (2017).

17. Smith & Anderson. Automation in everyday life. *Pew Research Center, Washington, DC* **Retrieved from http://www.pewinternet.org/2017/10/04/automation-in-everyday-life/ Google Scholar** (2017).

18. Social, T.O. Public attitudes toward robots. *Special Eurobarometer* **382** (2012).

19. Rosenman, R., Tennekoon, V. & Hill, L.G. Measuring bias in self-reported data. *International journal of behavioural & healthcare research* **2**, 320-332 (2011).

20. Petty, R.E., Tormala, Z.L., Brinol, P. & Jarvis, W.B. Implicit ambivalence from attitude change: an exploration of the PAST model. *J Pers Soc Psychol* **90**, 21-41 (2006).

21. Gawronski, B. & Bodenhausen, G.V. Associative and propositional processes in evaluation: an integrative review of implicit and explicit attitude change. *Psychological bulletin* **132**, 692-731 (2006).

22. Hannula, D.E., Simons, D.J. & Cohen, N.J. Imaging implicit perception: promise and pitfalls.

Nature reviews. Neuroscience **6**, 247-255 (2005).

23. Bar-Anan, Y. & Nosek, B.A. A comparative investigation of seven indirect attitude measures. *Behavior research methods* **46**, 668-688 (2014).

24. Schnabel, K., Asendorpf, J.B. & Greenwald, A.G. Assessment of Individual Differences in Implicit Cognition A Review of IAT Measures. *Eur J Psychol Assess* **24**, 210-217 (2008).

25. McConnell, A.R. & Leibold, J.M. Relations among the implicit association test, discriminatory behavior, and explicit measures of racial attitudes. *Journal of experimental social psychology* **37**, 435-442 (2001).

26. Mineka, S. & Ohman, A. Phobias and preparedness: The selective, automatic, and encapsulated nature of fear. *Biological psychiatry* **52**, 927-937 (2002).

27. Ohman, A. & Mineka, S. Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological review* **108**, 483-522 (2001).

28. Zald, D.H. The human amygdala and the emotional evaluation of sensory stimuli. *Brain research. Brain research reviews* **41**, 88-123 (2003).

29. Tamietto, M. & de Gelder, B. Neural bases of the non-conscious perception of emotional signals. *Nature Reviews Neuroscience* **11**, 697-709 (2010).

30. Cunningham, W.A.*, et al.* Separable neural components in the processing of black and white faces. *Psychological science* **15**, 806-813 (2004).

31. Williams, L.M.*, et al.* Amygdala-prefrontal dissociation of subliminal and supraliminal fear. *Human brain mapping* **27**, 652-661 (2006).

32. Felmingham, K.*, et al.* Dissociative responses to conscious and non-conscious fear impact underlying brain function in post-traumatic stress disorder. *Psychological medicine* **38**, 1771-1780

(2008).

33. Bigman, Y.E., Waytz, A., Alterovitz, R. & Gray, K. Holding Robots Responsible: The Elements of Machine Morality. *Trends in cognitive sciences* **23**, 365-368 (2019).

34. Waytz, A., Heafner, J. & Epley, N. The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of experimental social psychology* **52**, 113-117 (2014).

35. Cohen, J. *Statistical power analysis for the behavioral sciences* (Routledge, 2013).

36. Riek, L.D., Adams, A., & Robinson, P. Exposure to Cinematic Depictions of Robots and Attitudes Towards Them. (2011).

37. Greenwald, A., McGhee, D. & Schwartz, J. Measuring individual differences in implicit cognition: the implicit association test[J]. *JPSP* **74**, 1464-1480 (1998).

38. Kumaran, D., Banino, A., Blundell, C., Hassabis, D. & Dayan, P. Computations underlying social hierarchy learning: distinct neural mechanisms for updating and representing self-relevant information. *Neuron* **92**, 1135-1147 (2016).

39. Huijding, J. & de Jong, P. Beyond fear and disgust: The role of (automatic) contamination-related associations in spider phobia[J]. *Journal of Behavior Therapy and Experimental Psychiatry* **38**, 200-211 (2007).

40. Olson, M. & Fazio, R. Reducing automatically activated racial prejudice through implicit evaluative conditioning[J]. *Personality and Social Psychology Bulletin* **32**, 421-433 (2006).

41. Scott, G., O'Donnell, P., Leuthold, H. & Sereno, S. Early emotion word processing: Evidence from event-related potentials[J]. *Biol Psychol* **80**, 95-104 (2009).

42. Ke, K.e. Collins COBUILD Advanced Learner's English-Chinese Dictionary[M]. in *Beijing: Foreign Language Teaching and Research Press & Harper Collins Publishers Ltd* (2011).

43. Zhang, X., *et al.* Masked smoking-related images modulate brain activity in smokers. *Hum. Brain Mapp.* **30**, 896-907 (2009).

44. McNulty, J.K., Olson, M.A., Jones, R.E. & Acosta, L.M. Automatic associations between one's partner and one's affect as the proximal mechanism of change in relationship satisfaction: Evidence from evaluative conditioning. *Psychol. Sci.* **28**, 1031-1040 (2017).

45. Olson, M.A. & Fazio, R.H. Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin* **32**, 421-433 (2006).

46. Rolls, E.T., Huang, C.C., Lin, C.P., Feng, J. & Joliot, M. Automated anatomical labelling atlas 3. *NeuroImage* **206**, 116189 (2020).

47. Mori, K. Assessment of the Implicit Attitude of Japanese People toward Blacks and Little Black Sambo. *Open Journal of Social Sciences* **06. 1-13. 10.4236** (2018).

48. Anderson, C.A., Benjamin, A. J., & Bartholow, B. D. Does the gun pull the trigger? Automatic priming effects of weapon pictures and weapon names. *Psychological science* **9**, 308–314 (1998).

49. Blackford, J.U., Buckholtz, J.W., Avery, S.N. & Zald, D.H. A unique role for the human amygdala in novelty detection. *NeuroImage* **50**, 1188-1193 (2010).

50. Gomes, N., Silva, S., Silva, C.F. & Soares, S.C. Beware the serpent: the advantage of ecologically-relevant stimuli in accessing visual awareness. *Evol Hum Behav* **38**, 227-234 (2017).

51. Soares, S.C. The Lurking Snake in the Grass: Interference of Snake Stimuli in Visually Taxing Conditions. *Evol Psychol-Us* **10**, 187-197 (2012).

52. Kawai, N. & Koda, H. Japanese monkeys (Macaca fuscata) quickly detect snakes but not spiders: Evolutionary origins of fear-relevant animals. *Journal of comparative psychology* **130**, 299-303 (2016).

53. Shibasaki, M. & Kawai, N. Rapid detection of snakes by Japanese monkeys (Macaca fuscata): an

evolutionarily predisposed visual system. *Journal of comparative psychology* **123**, 131-135 (2009).

54. Mormann, F., *et al.* A category-specific response to animals in the right human amygdala. *Nature neuroscience* **14**, 1247-1249 (2011).

55. Rutishauser, U., *et al.* Single-Unit Responses Selective for Whole Faces in the Human Amygdala. *Current Biology* **21**, 1654-1660 (2011).

56. Yang, J.J., Bellgowan, P.S.F. & Martin, A. Threat, domain-specificity and the human amygdala. *Neuropsychologia* **50**, 2566-2572 (2012).

57. Breiter, H.C., *et al.* Response and habituation of the human amygdala during visual processing of facial expression. *Neuron* **17**, 875-887 (1996).

58. Wheatley, T., Milleville, S.C. & Martin, A. Understanding animate agents - Distinct roles for the social network and mirror system. *Psychological science* **18**, 469-474 (2007).

59. Bonda, E., Petrides, M., Ostry, D. & Evans, A. Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *Journal Of Neuroscience* **16**, 3737-3744 (1996).

60. Morris, J.S., *et al.* A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature* **383**, 812-815 (1996).

61. Blair, R.J.R., Morris, J.S., Frith, C.D., Perrett, D.I. & Dolan, R.J. Dissociable neural responses to facial expressions of sadness and anger. *Brain : a journal of neurology* **122**, 883-893 (1999).

62. Wright, C.I., *et al.* Differential prefrontal cortex and amygdala habituation to repeatedly presented emotional stimuli. *Neuroreport* **12**, 379-383 (2001).

63. Phillips, M., *et al.* Time courses of left and right amygdalar responses to fearful facial expressions. *NeuroImage* **13**, S458-S458 (2001).

64. Baas, D., Aleman, A. & Kahn, R.S. Lateralization of amygdala activation: a systematic review of

functional neuroimaging studies. *Brain Res Rev* **45**, 96-103 (2004).

65. Wager, T.D., Phan, K.L., Liberzon, I. & Taylor, S.F. Valence, gender, and lateralization of functional brain anatomy in emotion: a meta-analysis of findings from neuroimaging. *NeuroImage* **19**, 513-531 (2003).

66. Tamietto, M. & de Gelder, B. Neural bases of the non-conscious perception of emotional signals. *Nature reviews. Neuroscience* **11**, 697-709 (2010).

67. Fang, Z.Y., Li, H., Chen, G. & Yang, J.J. Unconscious Processing of Negative Animals and Objects: Role of the Amygdala Revealed by fMRI. *Frontiers in human neuroscience* **10** (2016).

68. Adolphs, R. What does the amygdala contribute to social cognition? *Year In Cognitive Neuroscience 2010* **1191**, 42-61 (2010).

69. Frith, C.D. & Frith, U. Mechanisms of Social Cognition. *Annual Review Of Psychology, Vol 63* **63**, 287-313 (2012).

70. Hardin, G. The competitive exclusion principle. *Science* **131**, 1292-1297 (1960).

71. Wangersky, P.J. Lotka-Volterra Population Models. *Annu Rev Ecol Syst* **9**, 189-218 (1978).

72. Klein, R.G. Neanderthals and modern humans - An ecological and evolutionary perspective. *Science* **305**, 45-45 (2004).

73. Mellars, P. Neanderthals and the modern human colonization of Europe. *Nature* **432**, 461-465 (2004).

74. Zilhao, J. Neandertals and moderns mixed, and it matters. *Evol Anthropol* **15**, 183-195 (2006).

75. Baraka, K.A.-O., Patrícia & Ribeiro, Tiago. An Extended Framework for Characterizing Social Robots. (2020).

76. Bartneck, C., Croft, E., & Kulic, D. Measurement Instruments for the Anthropomorphism, Animacy,

Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal of Social Robotics* **1(1), 71-81** (2009).

77. MacDorman, K.F., Vasudevan, S.K. & Ho, C.-C. Does Japan really have robot mania? Comparing attitudes by implicit and explicit measures. *Ai & Society* **23**, 485-510 (2008).

78. Graaf., M.M.A.d., Allouch., S.B. & Lutfi., S. What are People's Associations of Domestic Robots?: Comparing Implicit and Explicit Measures. *25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN) August 26-31, 2016. Columbia University, NY, USA*, 1077-1083 (2016).

79. Suzuki, Y., Galli, L., Ikeda, A., Itakura, S. & Kitazaki, M. Measuring empathy for human and robot hand pain using electroencephalography. *Scientific reports* **5** (2015).

80. Naneva, S., Gou, M.S., Webb, T.L. & Prescott, T.J. A Systematic Review of Attitudes, Anxiety, Acceptance, and Trust Towards Social Robots. *International Journal Of Social Robotics* **12**, 1179-1201 (2020).

81. Nomura, T., Suzuki, T., Kanda, T. & Kato, K. Measurement of negative attitudes toward robots. *Interact Stud* **7**, 437-454 (2006).

82. Bartneck, C., Kulic, D., Croft, E. & Zoghbi, S. Measurement Instruments for the Anthropomorphism, Animacy, Likeability, Perceived Intelligence, and Perceived Safety of Robots. *International Journal Of Social Robotics* **1**, 71-81 (2009).

bioRxiv preprint doi: https://doi.org/10.1101/2021.08.13.456053; this version posted November 22, 2022.

Figure Legends

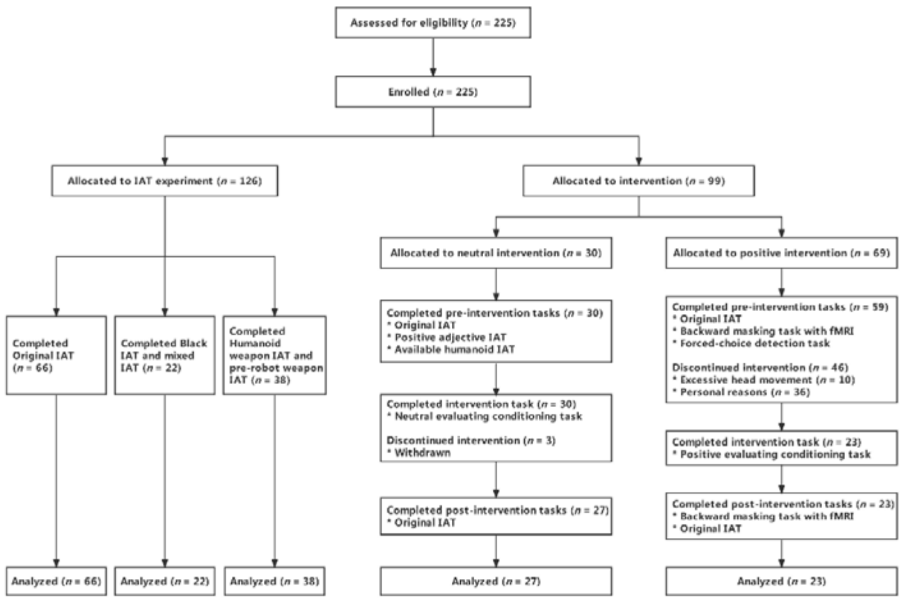

Figure 1. Overview diagram of the study flow.

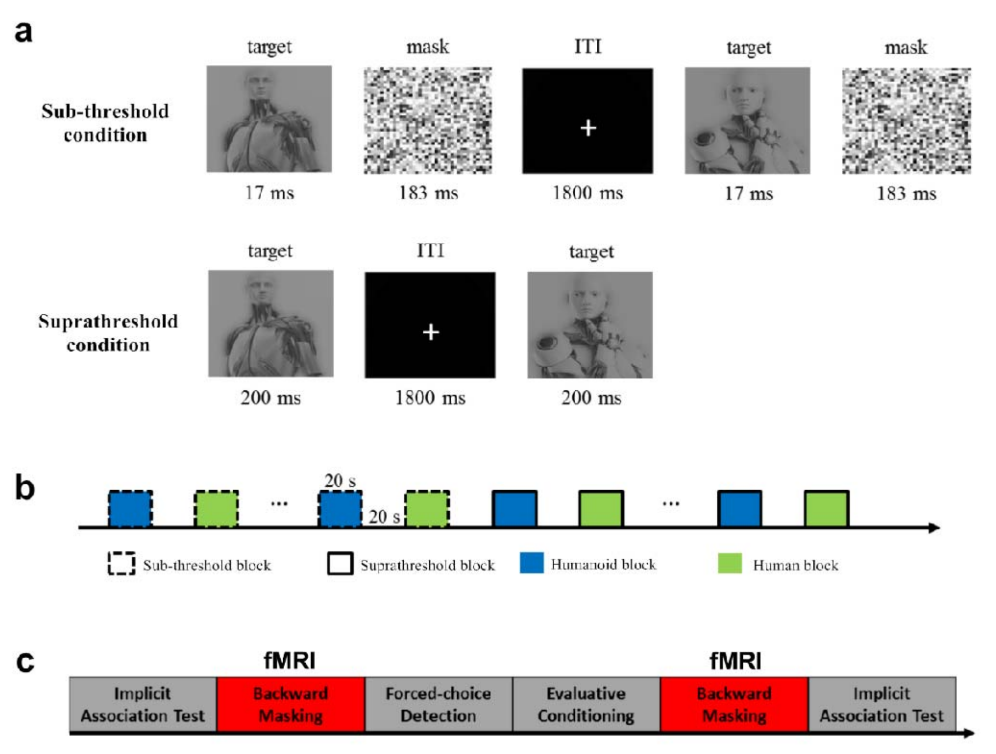

Figure 2. Description of the backward masking task and the procedure for the fMRI experiment. (a) Time setting of the backward masking task. In the unconscious condition, the target image was presented for 17 ms followed by a mask for 183 ms and a fixation cross for 1800 ms. In the conscious condition, the target image was presented for 200 ms followed by a fixation cross for 1800 ms. (b) Block design of the backward masking task. There were six unconscious blocks (three humanoid blocks and three human blocks) followed by six conscious blocks (three humanoid blocks and three human blocks). (c) Procedure for the fMRI experiment. Participants performed the original implicit association test outside the scanner and then completed a backward masking task with fMRI scanning followed by a forced-choice detection task and an evaluating conditioning task in the scanner; then, they performed the backward masking task during a second fMRI scan. Following the end of fMRI scanning, participants completed the original implicit association test outside the scanner again.

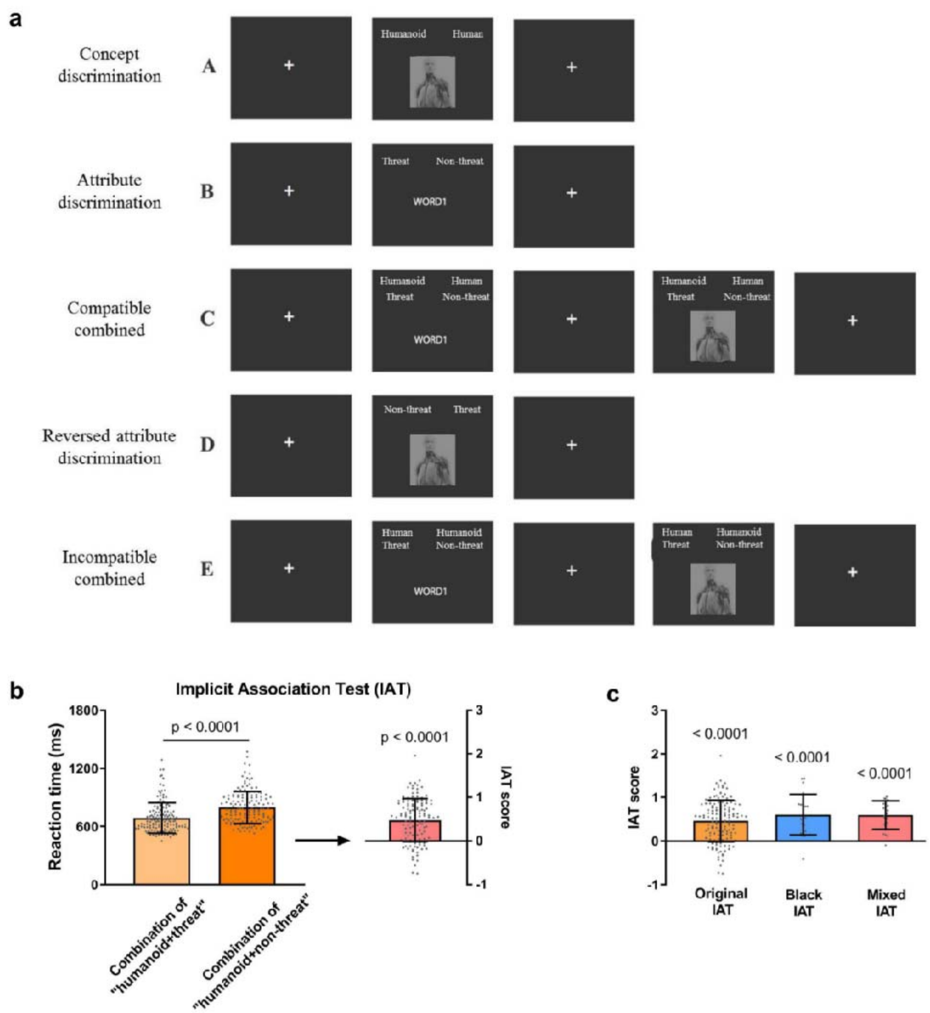

Figure 3. Implicit association test and its results. (a) Procedure of the original implicit association test. (b) Negative implicit attitudes toward humanoid robots. In the original implicit association test (IAT), participants responded faster to the combination of "humanoid + threat" than to the combination of "humanoid + nonthreat", and the computed IAT scores (effect size) were significant. (c) Participants displayed larger IAT scores in the original, Black, and mixed IATs. Plotted data represent the mean ± SD across participants.

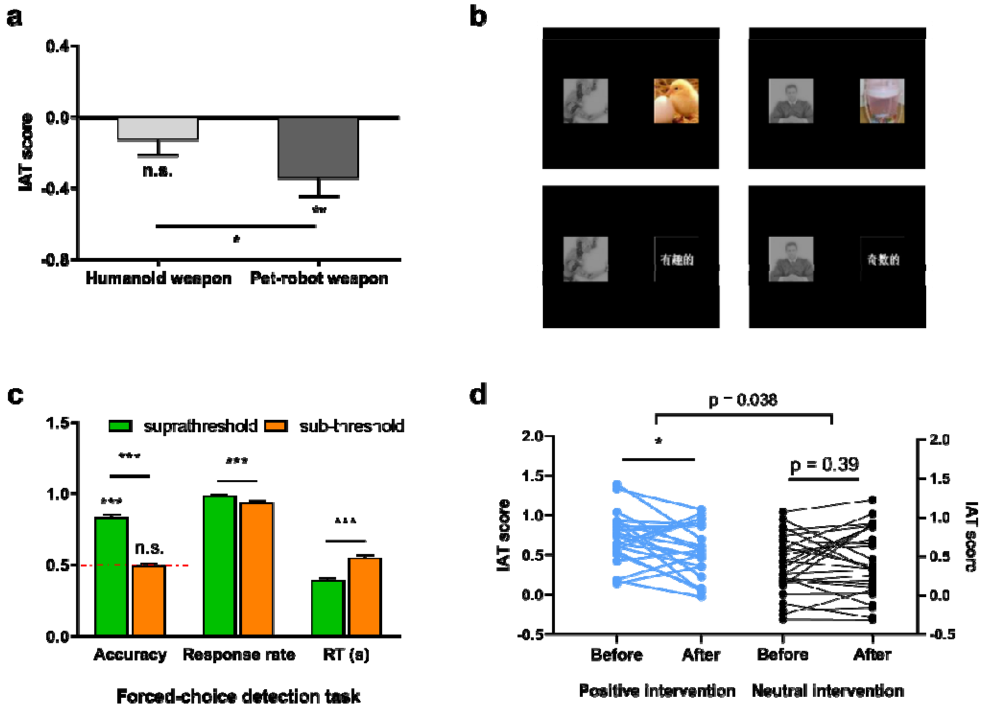

Figure 4. Evaluative conditioning task and its results. (a) More negative implicit attitudes toward humanoid robots compared to animal robots. Participants displayed larger IAT scores in the humanoid weapon IAT than in the pet-robot weapon IAT ($t_{37} = 3.07$, $p < 0.01$, Cohen's $d = 0.50$). (b) Procedure of evaluative conditioning task. (c) Results of the forced-choice detection task. (d) A significant group×time interaction effect was found, indicating that the weakened negative implicit attitudes toward humanoid robots were not due to the practice effect. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Plotted data represent the mean ± s.e.m. across participants. IAT = implicit association test.

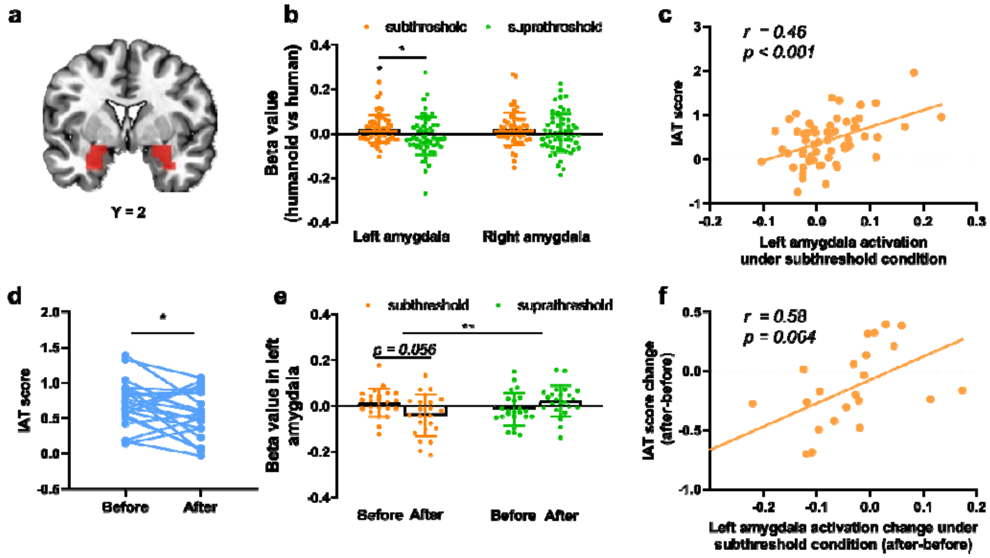

Figure 5. Humanoid robot image-related amygdala activity and amygdala activity changes. (a) The region of interest for the bilateral amygdala. (b) Although no amygdala activity differences were detected for consciously presented humanoid robot vs. human images, greater left amygdala activity was induced by humanoid robot images than by images of humans under unconscious presentation. (c) A greater IAT score was associated with greater left amygdala activity under unconscious conditions. (d) Significantly smaller IAT scores were found after modulation than those before modulation, indicating that participants' negative implicit attitude toward humanoid robots was successfully weakened. (e) The left amygdala activity differences of humanoid robot vs. human images did change under unconscious presentation after successfully weakening the negative implicit attitudes toward humanoid robots. (f) There was a significant correlation between IAT score changes and activation changes in the left amygdala under unconscious presentation. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. For b and e, plotted data represent the mean ± SD across participants. IAT = implicit association test.

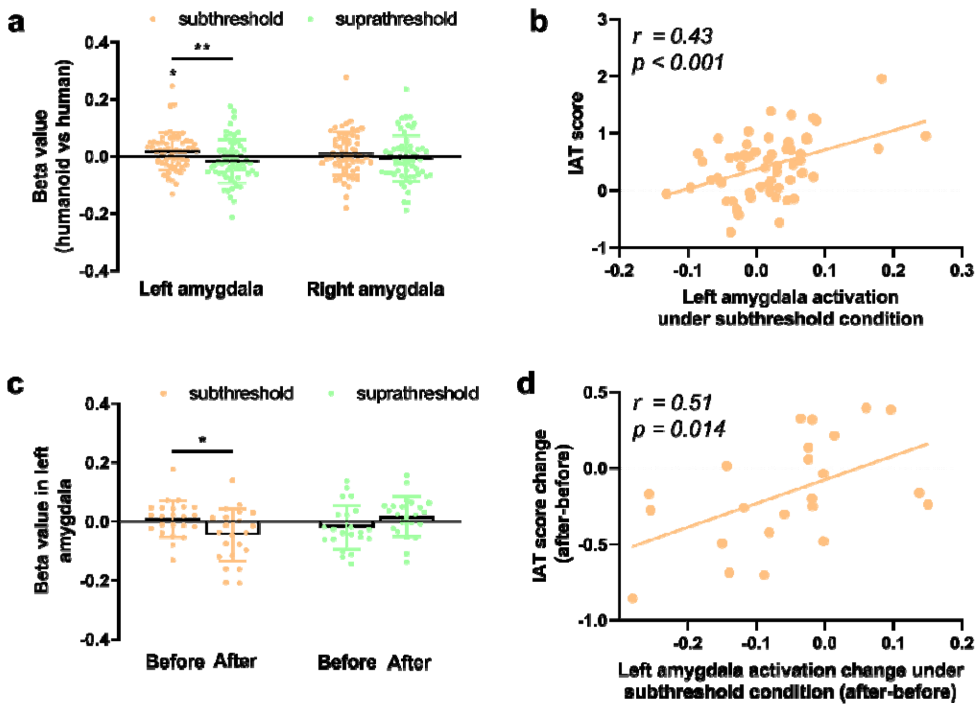

Figure 6. Results related to the left amygdala after controlling for novelty. (a) Greater left amygdala activity was induced by humanoid robot images than by images of humans under subthreshold presentation after controlling for the novelty of the humanoid robot and human images (t_{58} = 2.10, p = 0.04, Cohen's d = 0.27). (b) A greater IAT score was associated with greater left amygdala activity under the subthreshold condition after controlling for the novelty of the humanoid robot and human images (r = 0.43, p < 0.001). (c) Controlling for the novelty of the humanoid robot and human images, the left amygdala activity differences of humanoid robot vs. human images did change under subthreshold presentations after successfully weakening the negative implicit attitudes toward humanoid robots (t_{22} = -2.28, p = 0.033, Cohen's d = 0.47). (d) There was a significant correlation between IAT score change and activation change in the left amygdala under subthreshold presentation after controlling for the novelty of the humanoid robot and human images (r = 0.51, p = 0.014).

Tables

Table 1. Attitudes toward robots questions.

Items	5-point rating (1- completely disagree, 5-completely agree)
1. Humanoid robots could improve work efficiency.	
2. Humanoid robots could do jobs that human can't finish.	
3. Humanoid robots could improve the quality of life for humans.	
4. Humanoid robots would consume a lot of resources.	
5. Humanoid robots could have unexpected dangers.	
6. Humanoid robots would disrupt human life.	

Table 2. Stimulus of evaluative conditioning task.

Label	Stimulus
CSs	20 humanoid images, 20 human images
USs	
Positive	
Images	Chicken[a] (14[b]), Dog 2 (18), Island 1 (29), Apple (77), Cat 8 (781)
Words	Fantastic, Enjoyable, Fabulous, Excellent, Magnificent
Neutral	
Images	Bug 8 (234), Clothes Rack 1 (318), Plastic Cup (386), Butterfly 1 (451), Graph 2 (724)
Words	Odd, Stiff, Cold, Material, Muddy
Target	
Images	Antique 1 (292), Antique 2 (293), Antique 3 (295)
Words	Three meanings of antique
Fillers	
Images	Dock (424), Iron Bridge (406), Wolf 1 (547), Insect 10 (603), Plant 2 (842), Tool (785), Tree 4 (534), Locust (310), Frame (298), Shanghai 2 (387), City 4 (665), Tortoise (601), Hippo (482), Hair Drier (329), River (401), Mountain 7 (732)
Words	Bowl, Wine, Rock, Bench, Glass, Avenue, Boxer, Trunk, Rattle, Spray, Icebox, Ketchup, Radiator, Whistle, Nursery, Pamphlet, Thermometer

[a] CAPS image name.

[b] CAPS image numbers.

Table 3. Stimulus score of evaluative conditioning task.

	Valence		Arousal		Dominance	
	Mean	SD	Mean	SD	Mean	SD
USs						
Positive	7.27	0.18	6.25	0.45	7.11	0.42
Neutral	5.06	0.01	4.33	0.43	5.91	0.62
Filler	5.01	0.20	4.51	0.53	5.24	0.59
Target	5.24	0.08	4.10	0.16	5.59	0.01

Table 4. Stimulus arrangement in each block of evaluative conditioning task.

Trials	Arrangement
10	US + CS[a]
3	Target stimulus[b]
3	Target stimulus + Neutral filler
10	Neutral filler
10	Neutral filler + Neutral filler
5	Blank screen
20	Blank screen (preceding and following CS-US pairings)

[a] US appeared simultaneously with CS for 10 trials in each block.
[b] Target stimulus appeared alone for 3 trials in each block.