Humanoid robots are perceived as an evolutionary threat

Zhengde Wei[1, +], Ying Chen [2, +], Jiecheng Ren[1], Yi Piao[1], Pengyu Zhang[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 Hefei National Laboratory for Physical Sciences at the Microscale and 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

In the grand challenges of successful social encounters with socially sophisticated robots and shaping the future development of robots in socially acceptable ways, we need to quantify people perception to the robots. The critical assumption at the perception to humanoid robots, namely that people perceive humanoid robots as an evolutionary threat, has not been directly confirmed. We assume the existence of behavioral and neural automaticity for humanoid robots that were previously only evident for evolutionary threats. Here, we observed a monocular advantage for the perception of humanoid robots the same as an evolutionary threat (*i.e.*, snakes). Our neuroimaging analysis indicated that unconscious presentation of humanoid robot vs. human images led to significant left amygdala activation that was associated with negative implicit attitude to humanoid robots. After successfully weakening negative attitude, the left amygdala response to unconscious presentation of humanoid robot images decreased, and the decrease of left amygdala response was positively associated with the decrease of negative attitude. Our results reveal that processing of information about humanoid robots displays automaticity with regard to recruitment of visual pathway and amygdala activation. Our findings that humans may perceive humanoid robots as an evolutionary threat will guide the future direction of robots' development and bring us closer to interacting with socially sophisticated robots.

Key words: humanoid robot; evolutionary threat; amygdala; monocular advantage; robot morality

Introduction

Artificial intelligence advances have resulted in robots endowed with growingly sophisticated social abilities, which will make robots interact closely with humans as companions of everyday life. The prospect of socially interacting with these robots is on the horizon, and yet, remarkably little is known about how we perceive, interact with or accept these robots in social contexts [1, 2]. Quantifying people perception to robots will provide insights for a clearer and deeper understanding of the human side of human-robot interaction, and will facilitate successful social encounters with robots and shape the future development of robots in socially acceptable ways [3].

When perceive the robots, people treat robots as a threat [4-8]. From an evolutionary perspective, threats can be divided into evolutionary threats and modern threats [9]. Evolutionary threats relate to potential life-threatening stimuli and situations frequently encountered in the environments of our early evolutionary ancestors [9], while modern threats are fear-relevant stimuli that poses a problem today that was not prevalent throughout human evolutionary history [10]. When focusing on its manmade modern nature, robots should be categorized to modern threat. When focusing on the scene that robots will become fully autonomous "living" entities that threaten the survival of humanity [11, 12], robots should be categorized to evolutionary threat. We contend that acquiring a better understanding of human perceptions toward robots would be facilitated by identifying what type of threat(s) robots are perceived as; insights in this area could help in redefining human- robot interaction and robot ethics.

People perceive robots based upon context, cues, and cultural assumptions [13]. Our ideas about robots are mainly informed by science fiction media [12], and robots in science fiction are generally described as a super species with higher intelligence compared to human intelligence, attempting to eliminate humanity[12]. What's more, scientists want to build the better robots as artificial species [14, 15], even homo sapiens 2.0 [16]. We can expect the scene in the near future that robots have become "living" entities with fully autonomy, emotions, and even consciousness[12]. The two species with closest living habits have strongest competitiveness [17, 18], such as Neanderthal extinction by competition with anatomically modern humans [19-21]. Thus, the concept that fully autonomous robots are dangerous competitive "living" entities that will threaten the survival of humanity gradually becomes a deep impression, and recent statements from some influential industry leaders have strengthened these fears [22], supporting the possible idea that people perceive robots as an evolutionary threat.

Throughout human evolution, the ability to identify threatening stimuli has been critical to survival. Fear activates defensive behavior systems that help organisms to deal with different types of survival threats [23, 24]. It has

been proposed that fear is more likely to result from threat stimuli related to survival in evolutionary history. The preparedness model [9, 25] emphasizes that, compared to modern threats (*e.g.*, weapons), processing of evolutionary threats (*e.g.*, snakes) is assumed to show automaticity with regard to rapid recruitment of the behavioral and neural systems (see cartoon demonstrations in Fig. 1). This automaticity persists despite the fact that modern threats may be equally or even more closely related with trauma in our daily lives. A monocular advantage in responses to a particular category may reflect visual facilitation of automatic processing for that category [26], and there is empirical evidence supporting that a monocular advantage is present in responses to images of evolutionary threats (*e.g.*, snakes) but not in responses to images of modern threats (*e.g.*, guns) [26, 27].

The preparedness model [9, 25] also hypothesizes that the neural basis of automatic threat processing would be 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 threat sometimes differ, the more positive controlled processing can moderate more negative automatic processing [30], which could account for the absence of significant activation in amygdala in response to conscious presentations of threats [31, 32]. Notable, although controlled processing can eliminate amygdala activation caused by consciously presented threats, amygdala still showed significant activation by threats when threats were presented unconsciously [30-32]. The empirical evidence has demonstrated that amygdala responds more strongly to evolutionary threats than modern threats under unconscious presentation [33].

People perceive robots relied on their behaviors, abilities and appearance [13]. People consistently perceive more human characteristics in robots that look and act like humans [34]. Humanoid robots can be seen as machines with humanlike appearance, and mental capacities [3]. Here, we tested whether people perceive humanoid robots as an evolutionary threat. We hypothesized that if humanoid robots are perceived as an evolutionary threat, then participants would be expected to show a monocular advantage for humanoid robot images and an amygdala response to unconscious presentation of humanoid robot images. We pursued these topics by first measuring participants' explicit and implicit attitudes to humanoid robots which revealed that despite positive attitudes found from self-reporting in the questionnaire, participants actually had negative implicit attitudes to humanoid robots. We then used a Monocular Advantage Task which revealed that participants showed a monocular advantage for humanoid robots. We also employed fMRI and found that unconsciously presented images of humanoid robot trigger amygdala responses which was positively associated with implicit attitudes to humanoid robots, and later there was a decrease in amygdala activity after successfully weakening the negative implicit attitudes of

participants to humanoid robots. Our findings support the idea that people may perceive humanoid robots as an evolutionary threat even though humanoid robots are manmade, modern objects.

Results

Negative implicit attitudes to humanoid robots. People's attitudes to humanoid robots are complex. In this study, we used an explicit attitude questionnaire (see the details in Supplementary Information) and the Implicit Association Test (IAT, Supplementary Fig. 1) to assess participants' attitudes to humanoid robots. The IAT task reflects the degree of automatic connection between concept categories (*e.g.* humanoid or human) and attribute categories (*e.g.* threatening or non-threatening) by comparing the response time in combinations of "humanoid + threatening" and "human + non-threatening" and combinations of "humanoid + non-threatening" and "human + threatening". If there is a stronger association between humanoid robots and threatening meaning than between humans and threatening meaning, then we would expect participants to respond faster when humanoid images and threatening words share the same response.

In the explicit attitude questionnaire, participants (n = 66, age: 21.65±2.38 years; 41 females) showed a positive explicit attitude toward humanoid robots (t_{65} = 8.84, $p < 0.001$, Cohen's d = 1.10). In the IAT, participants (n = 127, age: 21.35±2.41 years; 80 females) had significant IAT scores (mean = 0.48; SD = 0.49; t_{126} = 10.95, $p < 0.0001$, Cohen's d = 0.97; Fig. 2a). Participants responded faster in combinations of "humanoid + threatening" and "human + non-threatening" than in combinations of "humanoid + non-threatening" and "human + threatening" (t_{126} = -8.81, $p < 0.0001$, Cohen's d = 0.78; Fig. 2a). We also recruited another group of participants (n = 22, age: 22.77±1.97 years; 14 females) to perform a Humanoid-Black IAT and a Humanoid-mixed (Caucasian, Black and Asian) IAT. Consistent results were found. Participants responded faster in compatible task than in incompatible task in Humanoid-Black IAT (RT: t_{21} = -6.40, $p < 0.001$, cohen's d = 1.36; mean IAT score = 0.61; Supplementary Fig. 2) and Humanoid-mixed IAT (RT: t_{21} = -7.33, $p < 0.001$, cohen's d = 1.57; mean IAT score = 0.60; Supplementary Fig. 2). Our IAT results indicated that, despite the findings from self reporting in the questionnaire, participants actually have negative implicit attitudes to humanoid robots.

Humanoid robots may be more competitive than animal robots, resulting in more negative implicit attitudes to humanoid robots as compared to animal robots. We recruited another group of participants (n = 38, age: 21.77±1.67 years; 23 females) to perform a Humanoid-weapon IAT and an Animal robot-weapon IAT. Participants displayed larger IAT scores in Humanoid-weapon IAT than in Animal robot-weapon IAT (t_{37} = 3.07, $p < 0.01$, Cohen's d = 0.50; Supplementary Fig. 3; Supplementary Table 1). These results indicate that participants have a more negative implicit attitude to humanoid robots compared to animal robots.

Monocular advantage for perception of humanoid robot images. A previous study indicated that a monocular advantage is shown in response to images of an evolutionary threat (*i.e.*, snakes) but not in response to images of a modern threat (*i.e.*, guns) [26]. Here, we used a Monocular Advantage Task (Supplementary Fig. 4) to investigate whether the visual pathway facilitates behavioral responses to humanoid robots. We hypothesized that if the humanoid robots are perceived as an evolutionary threat (as for snakes), then we would detect a monocular advantage for humanoid robots.

Sixty-two participants (age: 23.56±2.09 years; 40 females) were recruited. Each participant completed two blocks of trials for each of four categories (snake, gun, humanoid robot, and human), for a total of eight blocks. For each participant and condition, we measured mean accuracy and reaction time for each condition. We then computed inverse efficiency by dividing response time by accuracy [26, 27]. We carried out a repeated-measures factorial ANOVA with Image Category (snake, gun, humanoid robot, and human) and Eye Input (same and different eye) as within-subject factors, and with inverse efficiency in the condition of image mismatch as the dependent variable.

There was a significant interaction between Image Category and Eye Input ($F_{(3, 183)} = 3.81$, $p = 0.011$), significant main effect of Eye Input ($F_{(1, 61)} = 13.22$, $p = 0.001$), but no main effect of Image Category ($p = 0.14$). We followed up the significant interaction with paired-sample t-tests comparing the differences between same and different eye input for each image category. As shown in Fig. 2b, the inverse efficiency in same condition of Eye Input was smaller than in different condition of Eye Input for snake images ($t_{61} = -3.35$, $p = 0.0014$, Cohen's d = 0.42), and for humanoid robot images ($t_{61} = -3.05$, $p = 0.0034$, Cohen's d = 0.37), but not for human images or gun images (all $p > 0.70$). The inverse efficiency differences between Eye Input for snake images were significantly larger than for human images ($t_{61} = 2.24$, $p = 0.029$, Cohen's d = 0.29). The inverse efficiency differences between Eye Input for humanoid images were significantly larger than for gun ($t_{61} = 2.29$, $p = 0.026$, Cohen's d = 0.29) and human images ($t_{61} = 2.52$, $p = 0.015$, Cohen's d = 0.32). These results indicate a monocular advantage for snakes and for humanoid robots.

We then used a general linear regression analysis to examine whether the inverse efficiency differences between Eye Input in the condition of image mismatch could successfully predict perceived threat category (perceived evolutionary threat (*i.e.*, snake and humanoid robot) and perceived non-evolutionary threat (*i.e.*, gun and human)). The inverse efficiency differences significantly predicted perceived threat category ($t_{246} = 3.31$, $p = 0.001$, beta = 0.207). To further test the significance of the effect size (beta), a permutation test was performed. We

exchanged the category labels of two randomly selected samples and calculated the effect size of the inverse efficiency differences, using a new general linear regression model each time. We repeated this step 1000 times and found that the original effect size survived the permutation test ($p = 0.007$, Fig. 2c).

To explore the origin of monocular advantage for humanoid robots, another group of thirty participants (age: 22.73±1.66 years; 19 females) were recruited to complete a new Monocular Advantage Task. Each participant completed two blocks of trials for each of four categories (humanoid robot, doll, non-humanoid robot and tool), for a total of eight blocks. We carried out a repeated-measures factorial ANOVA with Image Category (humanoid robot, doll, non-humanoid robot and tool) and Eye Input (same and different eye) as within-subject factors, and with inverse efficiency in the condition of image mismatch as the dependent variable. There was no significant interaction between Image Category and Eye Input ($p = 0.98$), no main effect of Eye Input ($p = 0.059$), nor main effect of Image Category ($p = 0.68$). In the planed paired-sample t-tests, we compared the differences between same and different eye input for each image category. As shown in Supplementary Fig. 5a, the inverse efficiency in same condition of Eye Input was smaller than in different condition of Eye Input for humanoid images ($t_{29} = -2.06$, $p = 0.048$, Cohen's $d = 0.39$), but not for doll images, non-humanoid robot images nor tool images (all $p > 0.25$). Notably, the monocular advantage for humanoid robots could not explained by novelty since no differences of novelty were found between humanoid robots and non-humanoid robots (Supplementary Fig. 5b).

Greater left amygdala activity induced by humanoid robot images compared to human images under unconscious presentation. Fifty-nine participants (age: 21.08±2.43 years; 37 females) performed the IAT, and then they finished a modified Backward Masking Task (Fig. 3a-b) with fMRI scanning to measure amygdala activity in response to conscious and unconscious presentations of humanoid robot images. Following the end of fMRI scanning, all participants took part in a Forced-choice Detection Task to confirm that participants were aware of the stimulus under conscious presentation but were unaware under unconscious presentation in the Backward Masking Task. Twenty-three (age: 20.39±2.06 years; 15 females) out of the 59 participants performed the Evaluating Conditioning Task to weaken their negative implicit attitude to humanoid robots (the modulation effect was established in a pilot experiment, see the details in Supplementary Information, Supplementary Fig. 6), and performed the modified Backward Masking Task with a second fMRI scanning. Following the end of fMRI scanning, participants took part in the IAT outside the scanner again (Fig. 3c).

As show in Supplementary Fig. 7, the mean response rate was more than 90% under both presentations, and

was higher under conscious presentation (t_{58}= 4.14, $p < 0.001$, Cohen's d = 0.54) than under unconscious presentation. The response time under conscious presentation was significantly shorter than that under unconscious presentation (t_{58}= -10.91, $p < 0.001$, Cohen's d = 1.42). The accuracy under conscious presentation was significantly higher than under unconscious presentation (t_{58}= 16.68, $p < 0.001$, Cohen's d = 2.17). Importantly, the accuracy under unconscious presentation did not differ from random chance (t_{58}= -0.31, p = 0.76), whereas the accuracy under conscious 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 conscious presentation but are unaware of the stimuli under unconscious presentation.

In fMRI analysis, we first tested whether unconscious presentation of images of humanoid robots leads to greater amygdala (Fig. 4a) response compared to images of humans. Activation in response to humanoid robot images was significantly stronger than in response to human images in anatomically defined left amygdala under unconscious presentation (left amygdala: t_{58} = 2.66, p = 0.01, Cohen's d = 0.35; right amygdala: t_{58} = 1.20, p = 0.24; Fig. 4b); no such difference was observed upon conscious presentation (left amygdala: t_{58} = -1.85, p = 0.07; right amygdala: t_{58} = -0.057, p = 0.96). The activation differences between humanoid robot and human images under unconscious presentation in left amygdala were significantly stronger than activation differences under conscious presentation (left amygdala: t_{58} = 3.40, p = 0.001, Cohen's d = 0.44; right amygdala: t_{58} = 0.77, p = 0.45; Fig. 4b). Importantly, greater IAT score was associated with greater left amygdala activity under unconscious condition (r = 0.42, $p < 0.001$; Fig. 4c). Our results indicate that greater left amygdala activity induced by humanoid robot images compared to human images under unconscious presentation is associated with negative implicit attitude to humanoid robots.

We found that there was a significant difference of novelty between humanoid robot images and human images. It has been reported that the novelty shows contribution to amygdala activation [35]. 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 (please see the details in Supplementary Fig. 8ab).

The left amygdala response to unconscious presentation of humanoid robot images changes after successfully weakening negative attitude. The Evaluative Conditioning Task was used to weaken the negative implicit attitude to humanoid robots. We found that the post-test IAT scores were significantly smaller than pre-test (t_{22} = -2.16, p = 0.042, Cohen's d = 0.45; Fig 4d), indicating that participants' negative implicit attitude to humanoid robots had been successfully weakened by the Evaluative Conditioning Task. We tested whether the left

amygdala response to unconscious presentation of humanoid robot images changes after successful weakening negative implicit attitude. We conducted a two factor (time factor: before-modulation, after-modulation; presentation factor: unconscious presentation, conscious presentation) repeated measure ANOVA on activation differences between humanoid robot and human images in the left amygdala. Significant time×presentation interaction effect was found ($F_{(1,22)} = 14.75$, $p < 0.001$), but no significant time or presentation main effects were found (all $p > 0.33$). *Post-hoc* analysis revealed that there was a significant decrease in left amygdala activation between humanoid robot and human images under the unconscious presentations between the post-test and pre-test scans ($t_{22} = -2.30$, $p = 0.031$, Cohen's $d = 0.48$; Fig. 4e). Correlation analysis revealed that there was a significant correlation between IAT scores change and activation value change in the left amygdala under unconscious presentation ($r = 0.52$, $p = 0.011$; Fig. 4f). These results demonstrate a causal relationship between implicit attitudes to humanoid robots and the left amygdala response to unconscious presentation of humanoid robot images. Same to the previous section, after controlling the novelty of images by adding this as a covariate in our fMRI general linear model analyses, the results did not change (See the details in Supplementary Fig. 8cd).

Neural activity and functional connectivity changed after successfully weakening negative attitude. We then explored the change of neural activity and functional connectivity after successfully weakening negative attitude by whole-brain analyses. A whole-brain analysis demonstrated significant time×presentation interaction effects in the activation of left middle frontal gyrus ($x = 28$, $y = -5$, $z = 35$; uncorrected $p = 0.005$, voxels > 10) and left insula ($x = 43$, $y = -8$, $z = 11$; Fig. 5a). We then used psychophysiological interaction analysis with a seed in the left amygdala to test functional connectivity changes after successfully weakening negative attitudes. A whole-brain connectivity analysis demonstrated a significant time main effect in functional connectivity of the left amygdala and the left superior frontal gyrus ($x = 40$, $y = -41$, $z = 26$), left insula ($x = 46$, $y = 22$, $z = 41$), and left postcentral gyrus ($x = 46$, $y = 25$, $z = 31$; Fig. 5b).

Discussion

We found automaticity for processing information about humanoid robots that were previously only evident for evolutionary threats. The automaticity is supported by a monocular advantage for humanoid robots and an amygdala response to unconsciously presented humanoid robots. The monocular advantage in responses to humanoid robots reflects visual facilitation of responses for humanoid robots. The left amygdala response to unconsciously presented humanoid robots, which was positively associated with implicit attitudes to humanoid robot, indicates the automatic and quick detection of humanoid robots. After successfully weakening negative attitude, the decrease of IAT scores was positively associated with the decrease of activation in the left amygdala under unconscious presentation demonstrating a causal relationship between implicit attitudes to humanoid robots and the left amygdala response to unconscious presentation of humanoid robot images.

Our results provided evidence for the idea that humanoid robots may be perceived as an evolutionary threat. A monocular advantage in responses to a particular category may reflect visual facilitation of responses for that category [26]. Inputs from one eye are mostly segregated throughout the subcortical visual pathway [36], whereas inputs from the two eyes appear to be integrated to a greater extent in the extrastriate cortex [37]. Two images presented to the same eye are likely to activate overlapping populations of monocular subcortical neurons, whereas two images presented to different eyes are not. In a previous study [26], the observation of monocular advantage for snakes and the absence of monocular advantage for guns indicate that visual facilitation of responses may be specific to evolutionary threats. Consistent with this previous study, we observed a monocular advantage for snakes but not for guns. Interestingly, the images of humanoid robots also showed a monocular advantage, potentially reflecting visual facilitation of responses for humanoid robots. The failure of showing monocular advantage for dolls with human-like face but without intelligence and for non-humanoid robots with intelligence but without human-like face indicate that neither human-like face nor intelligence could separately explain the origin of monocular advantage for humanoid robots, which supports the possibility that the monocular advantage for humanoid robots derive from the combination of human-like face and intelligence. Notably, the monocular advantage for humanoid robots could not explained by novelty since no differences of novelty were found between humanoid robots and non-humanoid robots, and only the humanoid robots showed a monocular advantage.

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 hardwired “fear module” that allows us to automatically and quickly detect threatening stimuli [25]. Studies have well documented that the

amygdala responds selectively to evolutionary threats, irrespective of the affective valence, such as animate entities [38-40] and depictions of humans [41-43]. Our results showing that no amygdala activity in response to conscious presentation of humanoid robot images seemingly support perception of humanoid robots as a modern threat. However, when automatic and controlled evaluations of threat sometimes differ, the more positive controlled processing can moderate more negative automatic processing [30]. With the positive explicit attitude to humanoid robots found in present study, the absence of amygdala response to conscious presented 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 when threats were presented unconsciously [30-32]. Several researches suggest that the left amygdala might be specifically involved in processing of facial stimuli [44, 45]. It has also been reported that the left amygdala shows less habituation to fearful stimuli as compared with the right amygdala, which might make it more likely to capture the blood oxygen-level dependent changes in this area [46, 47]. However, the lateralization of amygdala activation is still controversial [48, 49]. Our present study found that greater left amygdala activity was induced by humanoid robot images compared to human images under unconscious presentation. After successfully weakening negative attitude, the decrease of IAT scores was positively associated with the decrease of activation in the left amygdala under unconscious presentation. our results indicate a causal relationship between implicit attitudes to humanoid robots and the left amygdala response to unconscious presentation of humanoid robot images. Note that these result did not alter after controlling for novelty of humanoid robot and human images. These results potentially reflect the automaticity with rapid recruitment of amygdala for humanoid robot-related stimuli processing.

Research emphasizes that people's minds are matters of perception [50], and we perceive robots based on the ascriptions such as some ability to think, remember, and exert self-control [13]. The more human-like a robot looks, the more people perceive it as having mental capacities [51]. We are entering an age where there will likely be a new type of entity that combines some properties of machines with some apparently mental capacities that were previously only evident in humans [52]. Combined with our demonstration that humans may perceive humanoid robots as an evolutionary threat, future humanoid robots with superhuman intelligence, emotions, and even consciousness may be perceived as a fully autonomous entity. By then, robots should hold fully moral responsibility for their wrongdoing. Although the issues of robot morality currently seem like science fiction, it is the best time to systematically explore robot morality while the robots and our perception to them are still in constant change [13]. We should consider security and ethics as early as possible, and put these considerations into

the technologies we develop. By understanding how people perceive the minds of robots and how people make sense of robot morality, we can help society think more about imminent rise of robot ethics.

In summary, this study demonstrates that humanoid robots are perceived as an evolutionary threat, even though humanoid robots are manmade, modern objects. 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 socially sophisticated robots.

Methods

Participants. One hundred and four participants took part in the IAT experiment. Among them, 66 participants (age: 21.65±2.38 years; 41 females) took part in a Humanoid-human IAT; 38 participants (age: 21.77±1.67 years; 23 females) performed a Humanoid-weapon IAT and a Animal robot-weapon IAT. Sixty-two participants (age: 23.56±2.09 years; 40 females) were recruited to perform the Monocular Advantage Task. Sixty-nine participants took part in the fMRI experiment. Potential participants in this experiment were excluded if they reported any contraindications for participating in a magnetic resonance imaging (MRI) study (e.g. non-removable metallic implants), or excessive head movement during scanning (TRs with motion over 0.3 mm were censored, participants who had more than 15% of their TRs censored were removed from further analysis). Ten participants were excluded because excessive head movement, ultimately data of 59 participants (age: 21.08±2.43 years; 37 females) were analyzed. Among them, 23 participants (age: 20.39±2.06 years; 15 females) took part in an Evaluative Conditioning Task and a second fMRI scanning. All participants were right-handed, with normal or corrected-to-normal vision, without alcohol or drug dependence disorders, without prior head injury. All participants and their immediate relatives had no psychiatric disorder or history of psychiatric disorder. All the participants were non-psychology undergraduates and they were unaware of the purpose and contents of the study before they participated in the experiment. The study was approved by the Research Ethics Committee of the University of Science and Technology of China, 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.

Procedures in fMRI experiment. In this experiment, participants performed the humanoid-human IAT, and then they finished a modified Backward Masking Task with fMRI scanning. Following the end of fMRI scanning, all participants took part in a Forced-choice Detection Task. Twenty-three out of the 59 participants performed the Evaluating Conditioning Task to weaken their negative implicit attitude to humanoid robots, and performed the modified Backward Masking Task with a second fMRI scanning. Following the end of fMRI scanning, participants took part in the Humanoid-human IAT outside scanner again (Fig. 3c).

Implicit Association Test. We assessed participants' automatic attitude using a standard implicit attitude measure paradigm — the IAT task [53, 54]. In the IAT, stimuli consisted of attribute words and concept images. The word stimuli were six adjectives, include three adjectives—threatening, strikes and lethal—that indicate "Threatening" and three adjectives—neutral, normal and average—that indicate "non-threatening" [55]. The image

stimuli were 20 images of humanoid robots, and 20 images of humans (Caucasians were used to represent humans, because the humanoid robot forms were Caucasians; we matched the action of each pair of images, and the gender ratio of humanoid robots and humans was same). In order to avoid confounding of stereotype of character, stimuli from movies were excluded. These images were downloaded on the open source in Internet. Adobe Photoshop CS5 was used for all images to uniformly process the pixels and colors of the images, so that the materials are consistent and the pixel size is 300×300. Before the formal experiment, participants were presented with an image set including all the 20 images of humanoid robot and 20 images of human in a pilot experiment. For each image, participant 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). Results showed that the affect valence and arousal of humanoid robot images were neutral (valence: mean = 3.05 scores, SD = 0.69; arousal: mean = 2.94, SD = 0.90). The affect valence and arousal of human images were neutral as well (valence: mean = 3.67 scores, SD = 0.66; arousal: mean = 3.20, SD = 1.03). There was a significant difference in novelty between humanoid robot images and human images (humanoid robot: mean score = 5.64, SD = 0.362; human: mean = 3.95, SD = 0.31; $t24 = 16.99$, $p < 0.001$). In Humanoid-weapon IAT, the image stimuli were 20 images of humanoid robot and 20 images of weapon. In Animal robot-weapon IAT, the image stimuli were 20 images of animal robot and 20 images of weapon.

The IAT involved a series of five discrimination blocks (Supplementary Fig. 1).

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 independently) of the screen, 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 1 s fixation cross following by a target image remained on the screen until the participant responded, or for a maximum of 3 s. There were 40 trials in this block, each concept image presented 1 time. The order of stimulus presentation was pseudo-randomized 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 non-threat), and the experimental procedure was consistent with the concept discrimination task. There were 18 trials in this block, each attribute word repeated 3 times.

Block 3 was a compatible combined task (humanoid + threat; human + non-threat). In this block, humanoid

robot images and threatening words share the same response, human images and non-threatening words share the same response. There were 76 trials in this block, each concept image presented 1 time and each attribute word 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 non-threatening words) were reversed. There were 18 trials in this block, each attribute word 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 non-threat) were as in block 4, and therefore swapped in side compared to block 3. There were 76 trials in this block, each concept image presented 1 time and each attribute word 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 IAT, the reduction-data procedure was applied [53, 54]. 1) eliminate data from subjects for whom more than 10% of trials has latency shorter than 300 ms; 2) eliminate data from subjects for whom has more than 20% error trials; 3) compute means of only correct trials' latencies for block 3 and 5; 4) compute the pooled standard deviation for all trials in block 3 and 5; 5) each error trial's latency in block 3 and 5 was replaced by the block mean computed at step 3 plus 600 ms; 6) compute the difference between mean latencies of block 3 and 5 using paired sample t-test; 7) the difference computed in step 6 was divided by the pooled standard deviation computed in step 4 to calculate the IAT effect size — the IAT score.

Monocular Advantage Task. The image stimuli consisted of 20 images of humanoid robots, 20 images of Humans, 20 images of snakes and 20 images of guns. A chin rest was used to stabilize the participant's head in front of a baffle. Two monitors were placed to the left and right of the center of the chin rest, perpendicular to the baffle. The baffle was placed in front of the participant's nose, so that the image from one of the monitors into one of the participant's eyes. The distance from the chin rest to the center of the mirror was 40 cm.

Each participant completed two blocks of trials for each of four categories of images (snakes, guns, humanoid robots and humans), for a total of eight blocks. The order of blocks was randomized for each participant. Participants completed 64 trials per block. On each trial, participants viewed a pair of images randomly selected from 20 images available in the current category (Supplementary Fig. 3). On half of the trials, both images were presented to the different eyes (half starting on the left, half starting on the right). The images were presented to

same eye on the other half of the trials (half to left eye, half to right eye). On each trial, a fixation cross appeared for 500 ms followed by the first image for 500 ms, then another fixation cross for 500 ms, then by the second image for 500 ms. Finally, a blank response screen was presented for 1 s. Participants were instructed to respond as quickly and accurately as possible to decide whether the two images were the same or different (i.e., match or mismatch). Participants were told that they could respond as soon as the second image appeared on the screen. Participants entered responses by pressing the "J" and "L" keys on a standard computer keyboard to indicate "same" and "different" responses, respectively. If there was an incorrect response or no response by 1.5 s was entered, a red '×' was presented for 1 s. If the participant entered a correct response within the limited time, a red '√' was presented for 1 s and the experiment proceeded to the next trial. A break was offered after each block. Within each block, trials were presented in a random order.

For each participant and condition, we measured mean accuracy and reaction time for each condition. We then computed inverse efficiency by dividing response time by accuracy [26, 27]. The inverse efficiency is likely to provide the most sensitive measure of performance because, unlike accuracy or reaction time alone, it captures speed-accuracy trade-offs in individual participants. Then, we carried out a repeated-measures factorial ANOVA with image category (snake, gun, humanoid robot and human) and eye condition (same and different) as within-subject factors and inverse efficiency in the condition of image mismatch as the dependent variable.

Another Monocular Advantage Task consisted of two blocks of trials for each of the four categories of images (humanoid robots, non-humanoid robots, doll and tool).

Materials in fMRI experiment. The materials of Backward Masking Task and Forced-choice Detection Task consisted of 20 images of humanoid robot and 20 images of human the same as used in the IAT. The materials of Evaluative Conditioning Task consisted of conditioned stimulus (CS), unconditioned stimulus (US), target stimulus and neutral fillers (Supplementary Table 2). CSs consisted of 20 images of humanoid robot and 20 images of human. Approximately half of the USs, target stimulus and neutral fillers were images, and half were words. We selected the US images, target images and neutral fillers images from the Chinese Affective Image System (CAPS), involving 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 are consistent and the pixel size is 300×300. The valence, arousal and dominance score of these 29 images were displayed in Supplementary Table 3 (on a scale from 1, low valence/arousal/dominance, to 9, high valence/arousal/dominance). The word stimulus–3 target adjectives, 5 positive adjectives (USs) [56], 5 neutral

adjectives (USs)[57] and 17 neutral filler nouns [57]–were selected from pilot study. We had translation these words into Chinese characters by using Collins COBUILD Advanced Learner's English-Chinese Dictionary [58] .The stimuli in this task were shown in Supplementary Table 4.

Backward Masking Task. Participants were finished the Backward Masking Task in MRI scanner (Fig. 3a-b) [59]. In unconscious presentation, the target image was presented for 17 ms followed by a mask for 183 ms and a fixation for 1800 ms. In conscious presentation, the target image was presented for 200 ms followed by a fixation 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. In order to avoid the possible effect on unconscious presentation, six conscious blocks (three humanoid blocks and three human blocks) followed the six unconscious 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 10s, each target block was separated by a 20 s baseline block.

Forced-choice Detection Task. To confirm that participants could be aware of the stimulus under conscious presentation but not under unconscious presentation in Backward Masking Task, a Forced-choice Detection Task was used. The Forced-choice Detection Task consisted of 80 trials, the first half were unconscious presentation, the second half were conscious presentation. The stimuli were presented closely in the same way as Backward Masking Task. The difference was that a 2000 ms forced-choice phase was following the target stimuli. Participants were informed that the target stimulus could have humanoid or human images and told to recognize the content of each image. Data from two participants in this task was excluded because of program crash. We compared the accuracy, response rate, and reaction time between conscious and unconscious presentations separately using paired sample t-test. We also compared the accuracy under both conditions to chance level (50%) using one sample t-test.

Evaluative Conditioning Task. The Evaluative Conditioning paradigm is a classic paradigm of changing implicit attitudes by pairing CS and US [60]. In this task, the participants' implicit negative attitude towards the humanoid robot may be reduced by pairing the humanoid robot images (CS) with the positive stimuli (US) (Supplementary Fig. 9). As a control condition, the human images (CS) were paired with the neutral stimuli (US). Participants were unaware of the repeated conditioned stimulus– unconditioned stimulus (CS-US) [61].

This task includes 6 blocks of 61 trials each. For each block, the specific stimuli were arranged, and all stimuli are presented in pseudo-random 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. In order to ensure that the participants carefully view the US-CS pairs during the whole experiment, the participants were told that the accuracy rate must be at least 95% after the end of the experiment, or to restart the task.

Before formal Evaluative Conditioning Task, a training Evaluative Conditioning Task was preceded. The implicit attitude change modulated by Evaluative Conditioning was analyzed by paired sample t-test.

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 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 for stereotaxic transformations after functional scanning. Before entering the scanner, participants were instructed to keep their heads still during all scans. During the backward masking task, 4 functional scan runs occurred with each lasting 4 min. There was an interval of approximately 1 min between every two runs.

fMRI processing. Functional data 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, 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).

To elucidate neural responses that correlated with humanoid robot images and human images under unconscious and conscious conditions, a general linear model (GLM) was used. Regressors of interest were unconscious humanoid blocks, unconscious human blocks, conscious humanoid blocks and conscious human blocks. Regressor of no interest was fixation blocks. 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 and second-level random effect analyses were conducted using one-sample t-tests to generate statistical maps.

Region of interest (ROI) analysis. We conducted ROI analyses on bilateral amygdala. The ROIs for bilateral amygdala were identified from the Talairach Daemon atlas [62]. We took parameter estimates in each participant from local average in a mask back-projected from the ROIs. The differential neural responses between humanoid robot and human conditions in the ROIs were analyzed.

Psychophysiological interaction (PPI) analysis. To investigate functional connectivity of the left amygdala, we used the PPI analysis [63]. We extracted the entire experimental time series of each participant for the left amygdala cluster. To create the PPI regressors of interest, we multiplied these normalized time series data by condition vectors containing ones for 2 second during each stimulus blocks (unconscious humanoid blocks, unconscious human blocks, conscious humanoid blocks and conscious human blocks) and zeros under all other conditions. These regressors were used in a separate regression analysis. The resulting parameter estimates represented the degree to which activity in each voxel correlated with activity in left amygdala. A 2 (image: humanoid versus human) × 2 (presentation: unconscious versus conscious) × 2 (time: before-modulation versus after-modulation) ANOVA was used to obtain connectivity map ($p < 0.005$, uncorrected).

Acknowledgements

This work was supported by grants from The National Key Basic Research Program (2018YFC0831101), The National Natural Science Foundation of China (71942003, 31771221, 61773360, and 71874170), Major Project of Philosophy and Social Science Research, Ministry of Education of China (19JZD010), CAS-VPST Silk Road Science Fund 2021 (GLHZ202128), Collaborative Innovation Program of Hefei Science Center, CAS (2020HSC-CIP001), and 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 acquisition of data. ZDW, YC analyzed and interpreted the data. PYZ, YP, JCR, 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.

Competing interests

The authors declare no competing interests.

Data availability statement

The complete dataset is available from the corresponding author.

Code availability statement

The MATLAB and AFNI code is available from the corresponding author.

Referenes

1. Wykowska, A., Chaminade, T. & Cheng, G. Embodied artificial agents for understanding human social cognition. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **371** (2016).
2. Cross, E.S., Hortensius, R. & Wykowska, A. From social brains to social robots: applying neurocognitive insights to human-robot interaction. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **374**, 20180024 (2019).
3. Broadbent, E. Interactions With Robots: The Truths We Reveal About Ourselves. *Annual review of psychology* **68**, 627-652 (2017).
4. 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).
5. 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).
6. 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).
7. Fast, E. & Horvitz, E. Long-Term Trends in the Public Perception of Artificial Intelligence. *Thirty-First Aaai Conference on Artificial Intelligence*, 963-969 (2017).
8. 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).
9. Mineka, S. & Ohman, A. Phobias and preparedness: The selective, automatic, and encapsulated nature of fear. *Biological psychiatry* **52**, 927-937 (2002).
10. Subra, B., Muller, D., Fourgassie, L., Chauvin, A. & Alexopoulos, T. Of guns and snakes: testing a modern threat superiority effect. *Cogn Emot* **32**, 81-91 (2018).
11. Yang, G.Z.*, et al.* The grand challenges of Science Robotics. *Science robotics* **3** (2018).
12. Cave, S. & Dihal, K. Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence* **1**, 74-78 (2019).
13. 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).
14. Kim, J.H., Lee, K.H. & Kim, Y.D. The Origin of Artificial Species: Genetic robot. *Int J Control Autom* **3**, 564-570 (2005).
15. Burton, H. Robo sapiens: The evolution of a new species. *Libr J* **125**, 110-110 (2000).
16. Dietrich, E. Homo sapiens 2.0: why we should build the better robots of our nature. *J Exp Theor Artif In* **13**, 323-328 (2001).
17. Hardin, G. The competitive exclusion principle. *Science* **131**, 1292-1297 (1960).
18. Wangersky, P.J. Lotka-Volterra Population Models. *Annu Rev Ecol Syst* **9**, 189-218 (1978).
19. Klein, R.G. Neanderthals and modern humans - An ecological and evolutionary perspective. *Science* **305**, 45-45 (2004).
20. Mellars, P. Neanderthals and the modern human colonization of Europe. *Nature* **432**, 461-465 (2004).
21. Zilhao, J. Neandertals and moderns mixed, and it matters. *Evol Anthropol* **15**, 183-195 (2006).
22. Preparing for the Future of Artifcial Intelligence (Executive Ofce of the President National Science and Technology Council.

23. Blanchard, D.C. & Blanchard, R.J. Ethoexperimental approaches to the biology of emotion. *Annual review of psychology* **39**, 43-68 (1988).
24. Paradiso, S. Affective neuroscience: The foundations of human and animal emotions. *Am J Psychiat* **159**, 1805-1805 (2002).
25. Ohman, A. & Mineka, S. Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological review* **108**, 483-522 (2001).
26. Vida, M.D. & Behrmann, M. Subcortical Facilitation of Behavioral Responses to Threat. *Scientific reports* **7**, 13087 (2017).
27. Gabay, S., Nestor, A., Dundas, E. & Behrmann, M. Monocular advantage for face perception implicates subcortical mechanisms in adult humans. *Journal of cognitive neuroscience* **26**, 927-937 (2014).
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. 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).
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. 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).
36. Menon, R.S., Ogawa, S., Strupp, J.P. & Ugurbil, K. Ocular dominance in human V1 demonstrated by functional magnetic resonance imaging. *Journal of neurophysiology* **77**, 2780-2787 (1997).
37. Bi, H*., et al.* Neuronal responses in visual area V2 (V2) of macaque monkeys with strabismic amblyopia. *Cerebral cortex* **21**, 2033-2045 (2011).
38. Mormann, F*., et al.* A category-specific response to animals in the right human amygdala. *Nature neuroscience* **14**, 1247-1249 (2011).
39. Rutishauser, U*., et al.* Single-Unit Responses Selective for Whole Faces in the Human Amygdala. *Current Biology* **21**, 1654-1660 (2011).
40. Yang, J.J., Bellgowan, P.S.F. & Martin, A. Threat, domain-specificity and the human amygdala. *Neuropsychologia* **50**, 2566-2572 (2012).
41. Breiter, H.C*., et al.* Response and habituation of the human amygdala during visual processing of facial expression. *Neuron* **17**, 875-887 (1996).
42. 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).
43. 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).

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

45. 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).

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

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

48. 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).

49. 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).

50. Weisman, K., Dweck, C.S. & Markman, E.M. Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 11374-11379 (2017).

51. de Visser, E.J.*, et al.* Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents. *J Exp Psychol-Appl* **22**, 331-349 (2016).

52. Prescott, T.J. Robots are not just tools. *Connection Science*, 142-149 (2017).

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

54. 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).

55. 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).

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

57. 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).

58. 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).

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

60. 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).

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

62. Lancaster, J.L.*, et al.* Automated Talairach atlas labels for functional brain mapping. *Human brain mapping* **10**, 120-131 (2000).

63. Wei, Z.D.*, et al.* Chronic nicotine exposure impairs uncertainty modulation on reinforcement learning in anterior cingulate cortex and serotonin system. *NeuroImage* **169**, 323-333 (2018).

Figures

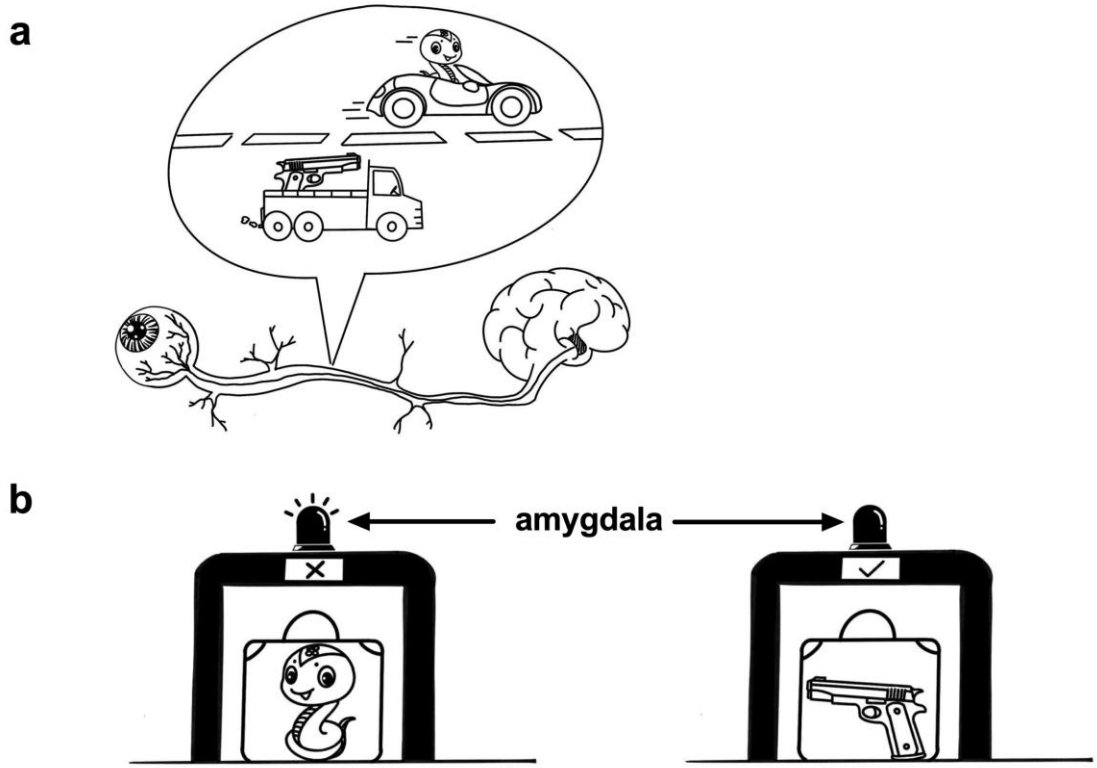


Figure 1. Automaticity for evolutionary threats as compared to modern threats. (a) Monocular advantage. The monocular input information of evolutionary threats (*e.g.*, snake) is faster than information of modern threats (*e.g.*, gun) to the brain for perception. (b) As the security devices can detect invisible dangerous goods in suitcases, the amygdala responds to unconsciously presented images of evolutionary threats, but do not respond to unconsciously presented images of modern threats.

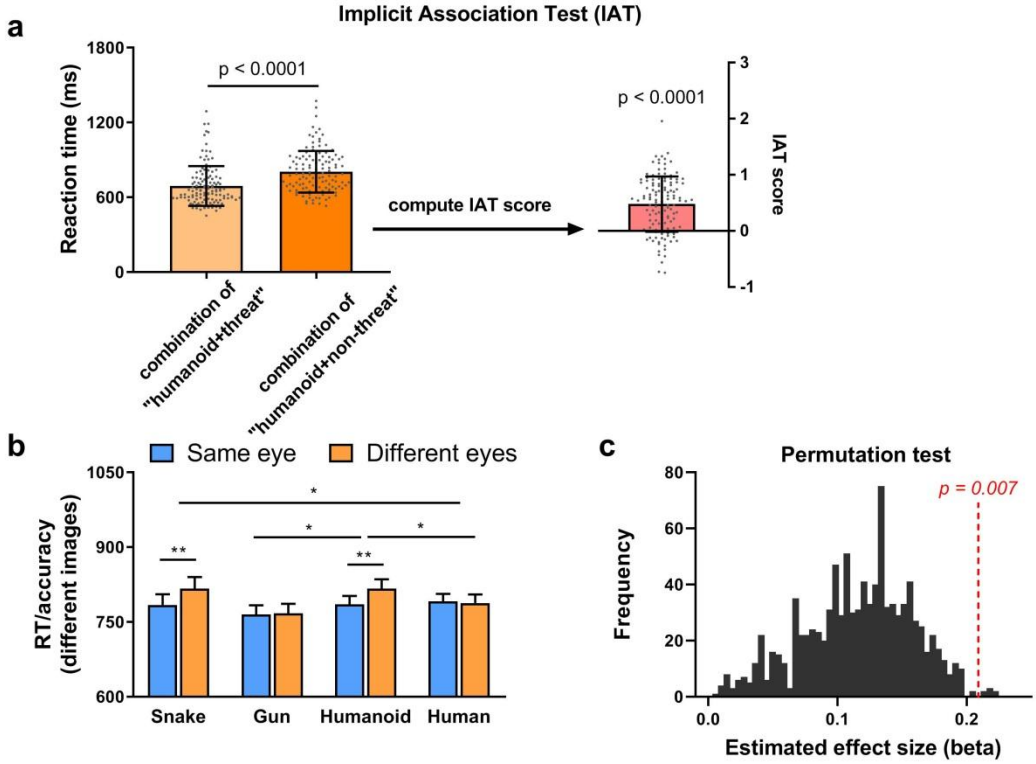


Figure 2. Implicit Association Test and Monocular Advantage Task. (a) Negative implicit attitude to humanoid robots. In the Implicit Association Test, participants responded faster in combination of "humanoid + threat" than in combination of "humanoid + non-threat", and the computed IAT scores (effect size) were significant. Plotted data represent the mean ± SD across participants. (b) A monocular advantage for the perception of humanoid robot images and snake images. In the condition of image mismatch, the inverse efficiency (reaction time (RT)/accuracy) in the same eye input condition was smaller than in different eye input condition for humanoid robot images, and for snake images. Plotted data represent the mean ± s.e.m. across participants. (c) The inverse efficiency differences between Eye Input in the condition of image mismatch could successfully predict perceived threat category (perceived evolutionary threat (*i.e.*, snake and humanoid robot) and perceived non-evolutionary threat (*i.e.*, gun and human)). To test the significance of the effect size of prediction, a permutation test was performed. We exchanged the category labels of two randomly selected samples and calculated the effect size of the inverse efficiency differences, using a new general linear regression model each time. We repeated this step 1000 times and found that the original effect size survived the permutation test ($p = 0.007$). * $p < 0.05$.

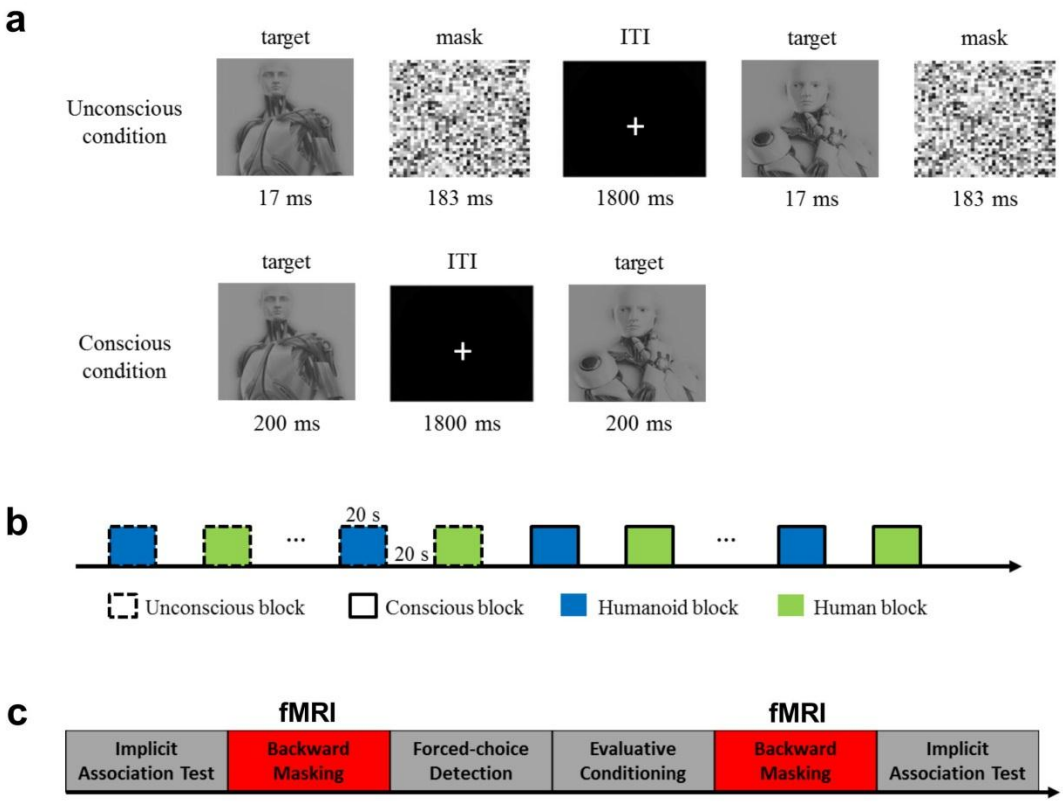


Figure 3. Description of the Backward Masking Task and the procedure for the fMRI experiment. (a) Time setting of the Backward Masking Task. In the unconscious condition, the target image was presented for 17 ms followed by a mask for 183 ms and a fixation for 1800 ms. In the conscious condition, the target image was presented for 200 ms followed by a fixation 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 an Implicit Association Test outside scanner, and then finished a modified Backward Mask Task with fMRI scanning followed by a Forced-choice Detection Task and an Evaluating Conditioning Task in scanner, and then they performed the modified Backward Mask Task with a second fMRI scan. Following the end of fMRI scanning, participants finished the Implicit Association Test outside scanner again.

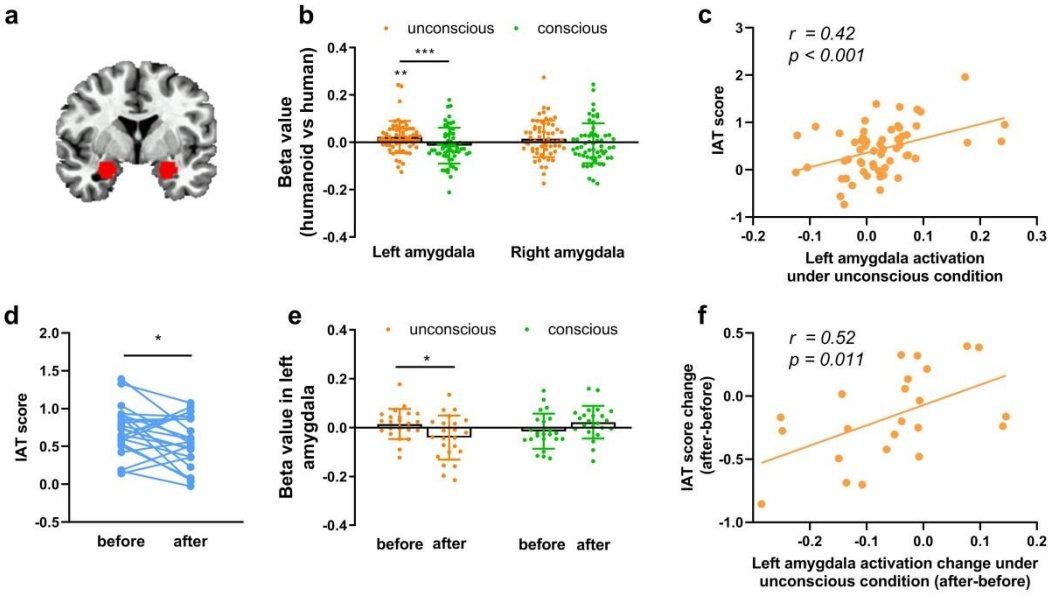


Figure 4. Humanoid robot image related amygdala activity and amygdala activity change. (a) The region of interest for bilateral amygdala. (b) Although no amygdala activity differences were detected for consciously presented humanoid robot vs. human images, greater left amygdala activity was induced by humanoid robot images compared to images of humans under unconscious presentation. (c) Greater IAT score was associated with greater left amygdala activity under unconscious condition. (d) Significantly smaller IAT scores were found in after-modulation compared to before-modulation, indicating that participants' negative implicit attitude to humanoid robots was successfully weakened. (e) The left amygdala activity differences of humanoid robot vs. human images did change under unconscious presentations after successfully weakening the negative implicit attitudes to humanoid robots. (f) There was a significant correlation between IAT scores change and activation change 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.

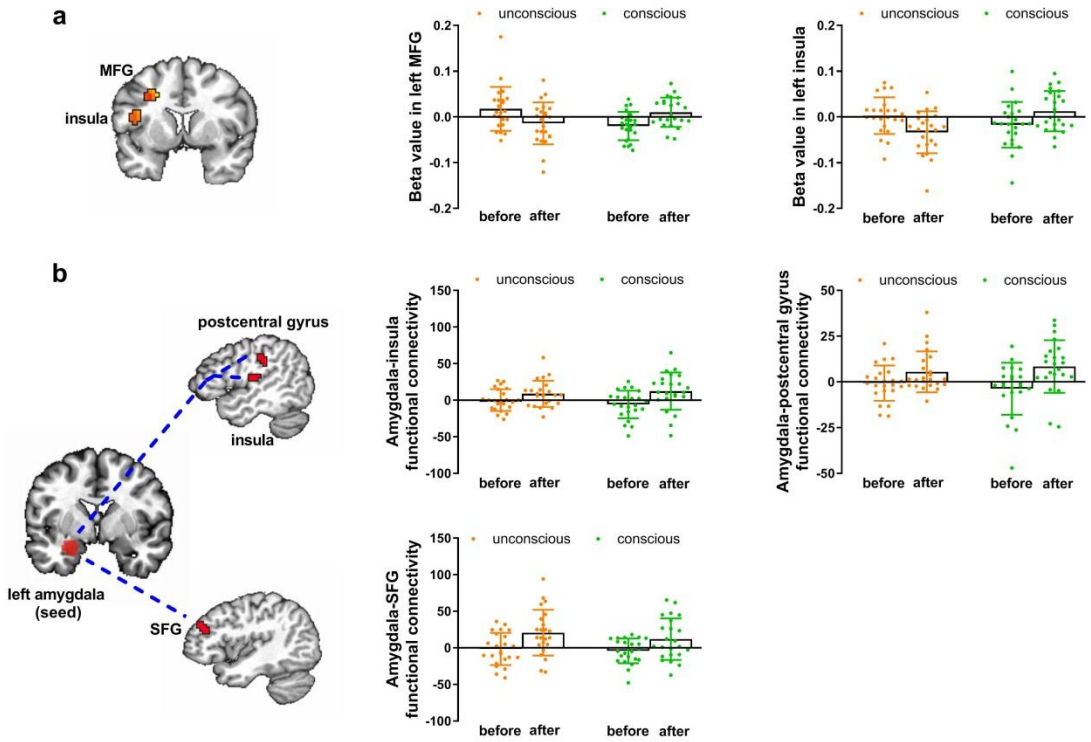


Figure 5. Change of brain activation and functional connectivity after attitude modulation. (a) A whole brain exploratory analysis demonstrated significant time×presentation interaction effects in the activation of left middle frontal gyrus (MFG) and left insula (uncorrected $p < 0.005$). (b) A whole brain exploratory analysis demonstrated significant time main effects in functional connectivity of the left amygdala and the left superior frontal gyrus (SFG), left insula, and left postcentral gyrus (uncorrected $p < 0.005$). Plotted data represent the mean ± SD. across participants.

Supplementary materials

Supplementary Tables

Supplementary Table 1. Results of Humanoid- and Animal robot-weapon Implicit Association Test.

Task	Condition	Mean	SD	t test
Humanoid-weapon IAT	Compatible (Humanoid robot + threat, weapon + non-threat)	769.61	207.81	1.89
	Noncompatible (Humanoid robot + non-threat, weapon + threat)	721.17	154.52	
Animal robot-weapon IAT	Compatible (Animal robot + threat, weapon + non-threat)	828.02	238.61	3.59^{***}
	Noncompatible (Animal robot + non-threat, weapon + threat)	710.56	140.54	

Supplementary Table 2. Stimulus of Evaluative Conditioning.

Label	Stimulus
CSs	20 humanoid images, 20 human images
USs	
Positive	
Images	Chicken[a](14[b]), Dog2(18), Island1(29), Apple(77), Cat8(781)
Words	Fantastic, Enjoyable, Fabulous, Excellent, Magnificent
Neutral	
Images	Bug 8(234), Clothes Rack1(318), Plastic Cup(386), Butterfly1(451), Graph2(724)
Words	Odd, Stiff, Cold, Material, Muddy
Target	
Images	Antique 1(292), Antique 2(293), Antique 3(295)
Words	Three meaning of antique
Fillers	
Images	Dock(424), Iron Bridge(406), Wolf1(547), Insect10(603), Plant2(842), Tool(785), Tree4(534), Locust(310), Frame(298), Shanghai2(387), City4(665), Tortoise(601), Hippo(482), Hair Drier(329), River(401), Mountain7(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.

Supplementary Table 3. Stimulus score of Evaluative Conditioning.

	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

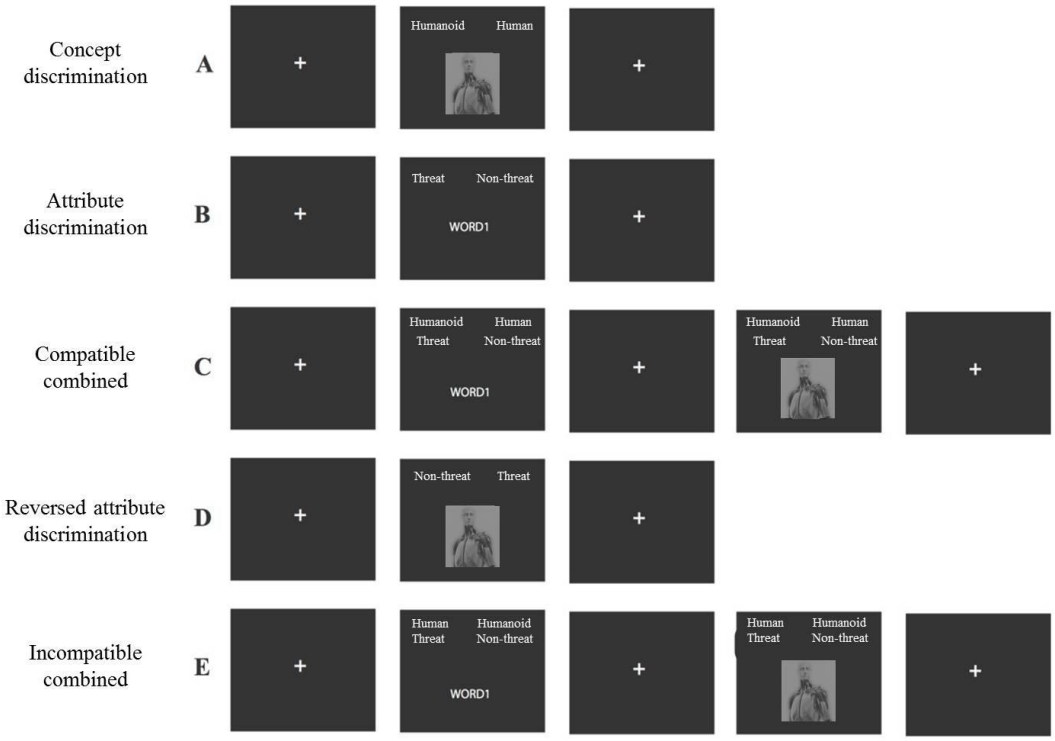
Supplementary Table 4. Stimulus arrangement in each block of Evaluative Conditioning

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(precede and follow CS-US pairings)

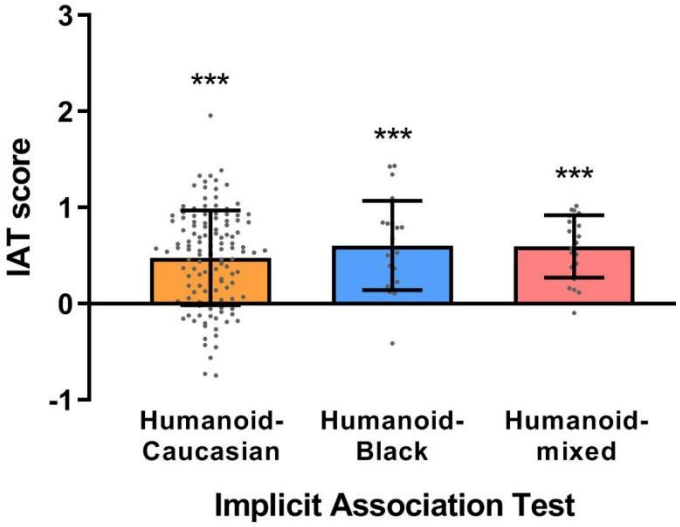
[a] US appeared simultaneously with CS for 10 trials in each block

[b] Target stimulus appeared alone for 3 trials in each block

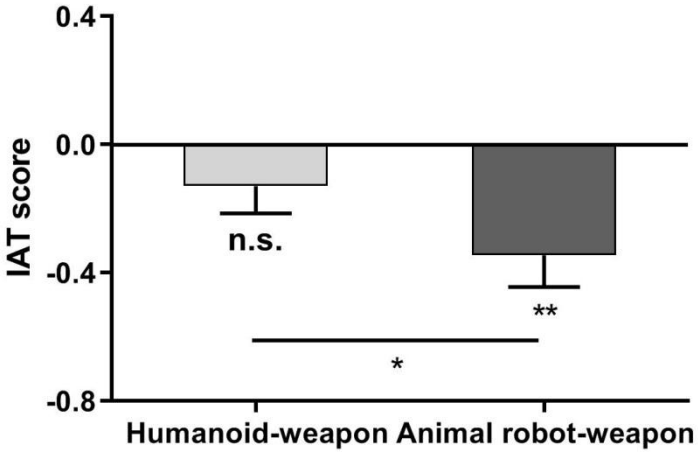
Supplementary Figures



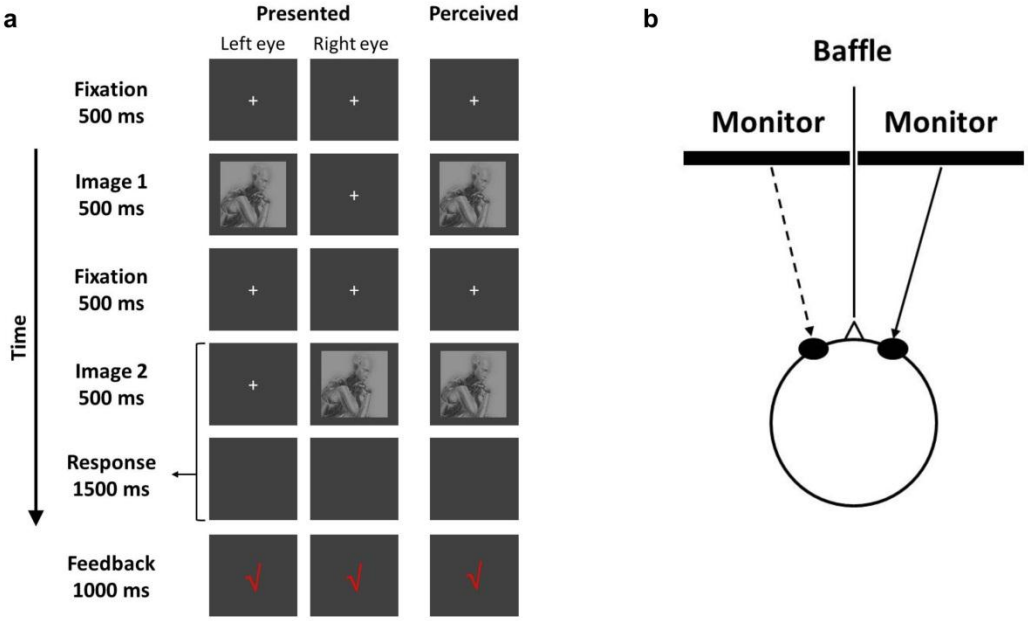
Supplementary Figure 1. Procedure of Implicit Association Test.



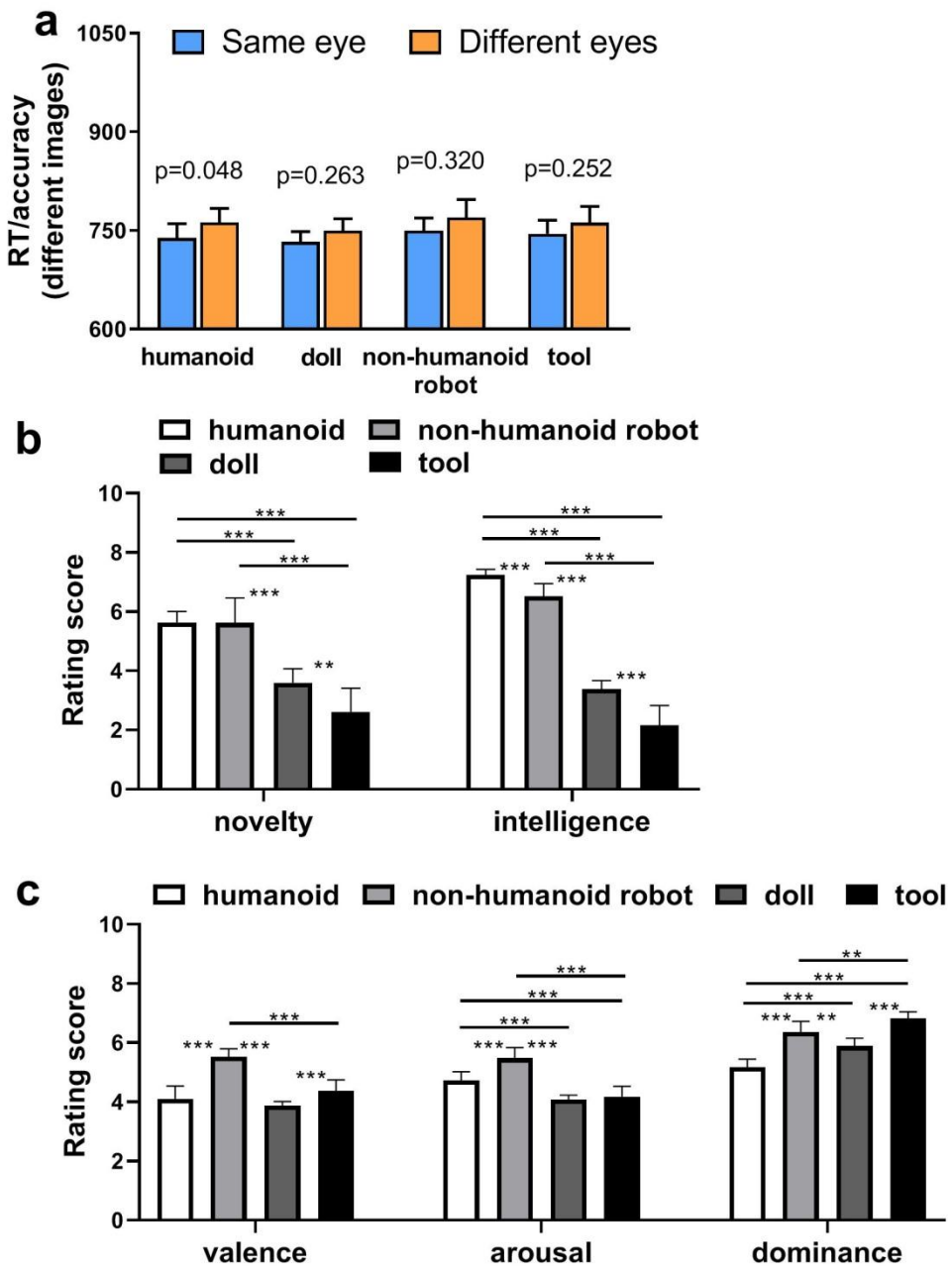
Supplementary Figure 2. More negative implicit attitude to humanoid robots as compared to humans. Participants displayed larger IAT scores in humanoid-Caucasian, humanoid-Black, and humanoid-mixed IAT. *** $p < 0.001$. Plotted data represent mean ± s.e.m. across participants. IAT = implicit association test.



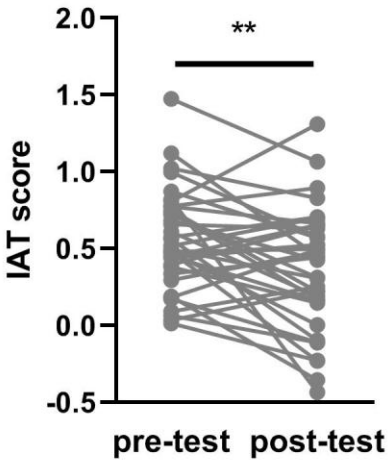
Supplementary Figure 3. More negative implicit attitude to humanoid robots as compared to animal robots. Participants displayed larger IAT scores in humanoid-weapon IAT than that in animal robot-weapon IAT ($t_{37} = 3.07$, $p < 0.01$, cohen's d = 0.50). * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. Plotted data represent mean ± s.e.m. across participants. IAT = implicit association test.



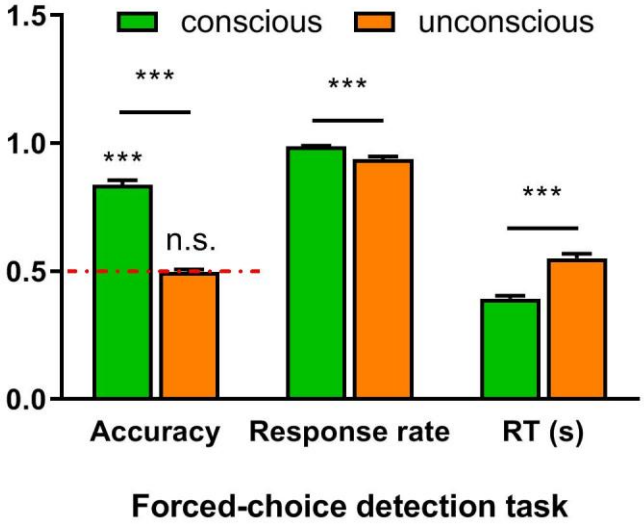
Supplementary Figure 4. Monocular Advantage Task. (a) Sequence of events in Monocular Advantage Task. Two images were presented to the same eye (left or right) or different eyes sequentially. Participant was asked to indicated whether the two images were the same. (b) A chin rest was used to stabilize the participant's head in front of a baffle.



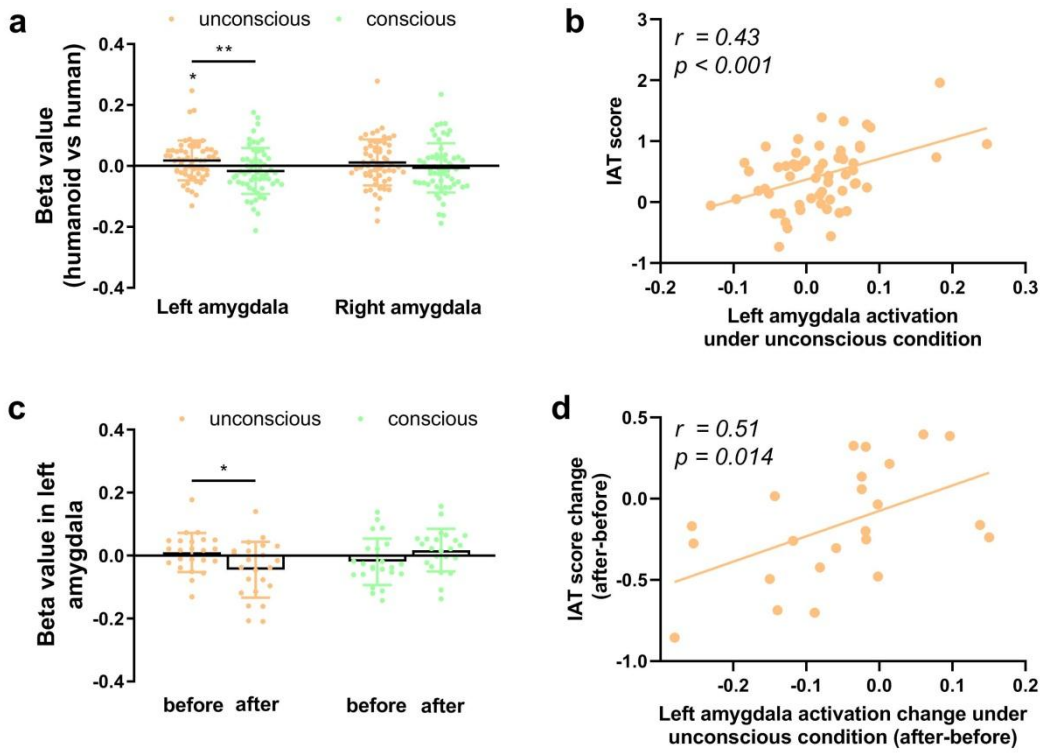
Supplementary Figure 5. (a) A monocular advantage for the perception of humanoid robot images. In the condition of image mismatch, the inverse efficiency (reaction time (RT)/accuracy) in the same eye input condition was smaller than in different eye input condition for humanoid robot images. Plotted data represent the mean ± s.e.m. across participants. (b) Rating score for the humanoid, non-humanoid robot, doll and tool images in dimensions including valence, arousal, dominance. (c) Rating score for the humanoid, non-humanoid robot, doll and tool images in dimensions including novelty and intelligence.



Supplementary Figure 6. Negative implicit attitude to humanoid robot significantly decreased. Significant smaller IAT score was found in post-test than in pre-test ($t_{36} = -3.13$, $p < 0.01$, cohen's d = 0.51). ** $p < 0.01$. IAT = implicit association test.



Supplementary Figure 7. Results of Forced-choice Detection Task. The accuracy under conscious condition was significantly higher than that under unconscious condition (t_{58}= 16.68 , $p < 0.001$, cohen's d = 2.17). Importantly, the accuracy under unconscious condition was equal to chance level (t_{58}= −0.31, p = 0.76), but the accuracy under conscious condition was higher than chance level (t_{58}= 19.13 , $p < 0.001$, cohen's d = 2.49). The mean response rate was more than 90% under both conditions, and higher under conscious condition (t_{58}= 4.14 , $p < 0.001$, cohen's d = 0.54) compared with under unconscious condition. The response time under conscious condition was significantly shorter than that under unconscious condition (t_{58}= −10.91 , $p < 0.001$, cohen's d = 1.42).



Supplementary Figure 8. Results related to left amygdala after controlling for novelty of humanoid robot and human images.(a) Greater left amygdala activity was induced by humanoid robot images compared to images of humans under unconscious presentation after controlling for novelty of humanoid robot and human images (t_{58} = 2.10, p = 0.04, Cohen's d = 0.27). (b) Greater IAT score was associated with greater left amygdala activity under unconscious condition after controlling for novelty of humanoid robot and human images (r = 0.43, p < 0.001). (c) Controlling for novelty of humanoid robot and human images, the left amygdala activity differences of humanoid robot vs. human images did change under unconscious presentations after successfully weakening the negative implicit attitudes to humanoid robots (t_{22} = -2.28, p = 0.033, Cohen's d = 0.47). (d) There was a significant correlation between IAT scores change and activation change in the left amygdala under unconscious presentation after controlling for novelty of humanoid robot and human images (r = 0.51, p = 0.014).

Supplementary Figure 9. Procedure of Evaluative Conditioning.

Explicit attitude questionnaire

Participants were asked to rate their perceptions of some opinions toward humanoid robot on a 5-point scales (1-completely disagree, 5-completely agree). Opinions were as follows:

1. Humanoid robot could improve work efficiency;

2. Humanoid robot could do jobs that human can't finish;

3. Humanoid robot could improve the quality of life for humans;

4. Humanoid robot would consume a lot of resources;

5. Humanoid robot could have unexpected dangers;

6. Humanoid robot would disrupt the original human life.

Item 4, 5 and 6 were reversed items. The sum score in these six items indicate participants' opinion toward humanoid robot.

Pilot experiment: negative implicit attitude to humanoid robots could be weakened via Evaluative Conditioning Task

Participants. Forty-one participants took part in this experiment, among them four participants were excluded because their IAT scores before the evaluative conditioning procedure were positive, so finally 37 participants' data were analyzed.

Procedures. In this experiment, the IAT task was employed before and after an evaluative conditioning procedure.

Results. In the pre-test IAT, participants express negative attitude toward humanoid robot (RT: $t_{36} = -8.32$, $p < 0.001$, cohen's d = 1.37; IAT score: mean D = 0.55). After Evaluative Conditioning, we found significant smaller IAT scores in post-test than in pre-test ($t_{36} = -3.13$, $p < 0.01$, cohen's d = 0.51; Supplementary Figure 4).

References

1. Greenwald, A., McGhee, D. & Schwartz, J. Measuring individual differences in implicit cognition: the implicit association test[J]. *JPSP* **74**, 1464-1480 (1998).
2. 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).
3. 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).
4. Vida, M.D. & Behrmann, M. Subcortical Facilitation of Behavioral Responses to Threat. *Scientific reports* **7**, 13087 (2017).
5. Gabay, S., Nestor, A., Dundas, E. & Behrmann, M. Monocular advantage for face perception implicates subcortical mechanisms in adult humans. *Journal of cognitive neuroscience* **26**, 927-937 (2014).
6. Olson, M. & Fazio, R. Reducing automatically activated racial prejudice through implicit evaluative conditioning[J]. *Personality and Social Psychology Bulletin* **32**, 421-433 (2006).
7. 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).
8. 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).
9. Zhang, X.*, et al.* Masked smoking‐related images modulate brain activity in smokers. *Hum. Brain Mapp.* **30**, 896-907 (2009).
10. 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).
11. Olson, M.A. & Fazio, R.H. Reducing automatically activated racial prejudice through implicit evaluative conditioning. *Personality and Social Psychology Bulletin* **32**, 421-433 (2006).
12. Lancaster, J.L.*, et al.* Automated Talairach atlas labels for functional brain mapping. *Human brain mapping* **10**, 120-131 (2000).
13. Wei, Z.D.*, et al.* Chronic nicotine exposure impairs uncertainty modulation on reinforcement learning in anterior cingulate cortex and serotonin system. *NeuroImage* **169**, 323-333 (2018).