Humanoid robots are perceived as an evolutionary threat

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

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

+ These authors contributed equally to this work

* Correspondence: zxcustc@ustc.edu.cn

Abstract

Studying how we perceive humanoid robots will provide insights for a deeper understanding of human-robot interaction. Our ideas about humanoid robots are mainly informed by science fiction, and humanoid robots are generally described as an evolutionary threat in science fiction that has not been tested. The preparedness model has emphasized that the fear module is automatically activated by evolutionary threats, and its underlying neural circuit is centered on the amygdala. We hypothesized that if humanoid robots are perceived as an evolutionary threat even though humanoid robots are manmade, modern objects, we would expect to observe a monocular advantage for humanoid robots and an amygdala response to unconsciously presented humanoid robots that were previously only evident in 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 amygdala activation. Despite a positive humanoid robot-related association had been established by associative learning (as evidenced by results of successfully weakening the negative implicit attitude to humanoid robots and enhancing functional connectivity between the amygdala and hippocampus), the amygdala could still automatically and quickly detect humanoid robots. 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 apparently perceive humanoid robots as an evolutionary threat may help inform redefinition of human- robot interaction and robot ethics.

Introduction

Artificial intelligence advances have led to robots that look and behave like humans. Studying how we perceive humanoid robots will provide insights for a deeper understanding of human-robot interaction, which may facilitate successful social encounters between humans and humanoid robots. A growing number of studies has found that the perceptions toward humanoid robots are complex and inconsistent (*1-5*). We contend that acquiring a better understanding of human perceptions toward humanoid robots would be facilitated by identifying what type of threat(s) humanoid robots are perceived as; insights in this area could help in redefining human-humanoid robot interaction and humanoid robot ethics.

From an evolutionary perspective, threats can be divided into evolutionary threats and modern threats (*6*). Evolutionary threats relate to potential life-threatening stimuli and situations frequently encountered in the environments of our early evolutionary ancestors. Modern threats are fear-relevant stimuli that poses a problem today that was not prevalent throughout human evolutionary history (*7*). 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 (*8, 9*). It has been proposed that fear is more likely to result from threat stimuli related to survival in evolutionary history. The preparedness model (*6, 10*) 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 some 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 (*11*), 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) (*11, 12*).

The preparedness model (*6, 10*) 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 (*13, 14*). When automatic and controlled evaluations of threat sometimes differ, the more positive controlled processing can moderate more negative automatic processing (*15*), which could account for the absence of significant activation in amygdala in response to conscious presentations of threats (*16, 17*). 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 (*15-17*). The empirical evidence has demonstrated that amygdala responds more strongly to evolutionary threats than modern threats under unconscious presentation (*18*). It has also been proposed that there is enhanced resistance to extinction from evolutionary threats relative to modern threats (*10, 19*). Using classical conditioning paradigms, comparisons of aversive conditioning have been indicated enhanced resistance to extinction to snakes relative to guns (*19, 20*).

The humanoid robots are manmade, modern objects that have emerged recently in our cultural history. In previous studies, neither a monocular advantage (*11, 12*) nor an amygdala response (*21*) were detected upon encountering modern threats. Thus, we would anticipate that humanoid robots should fail to elicit a monocular advantage or an amygdala response. However, our ideas about humanoid robots are mainly informed by science fiction media (*22*). Science fiction has created the backdrop for how humanoid robots are interpreted and assessed (*22*). Humanoid robots are popular in science fiction (*23*) and are generally presented as artificial "living" entities with superhuman intelligence, digital emotions, and even consciousness(*22*). In many science fiction plots, these super-intelligent humanoid robots can exceed the abilities of human beings, and recent statements from some influential industry leaders have strengthened these fears (*24*). With the ideology that the intellectually superior are by nature masters and the intellectually inferior by nature slaves, it is comprehensible that we fear humanoid robots with superhuman intelligence will enslave us (*25*). So it is clear that some humanoid robots are perceived as more than "just tools"; rather, they are sometimes viewed as dangerous "living" entities. Thus, we proposed that humanoid robots are perceived as an evolutionary threat.

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. Here, 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 robots trigger amygdala responses, and later there was resistance to changes in amygdala activity after successfully weakening the negative implicit attitudes of participants to humanoid robots. Our findings support the idea that humans may perceive humanoid robots as an evolutionary threat.

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). 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 intelligent and 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 a 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 Figure 2; 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) (*11*). Here, we used a Monocular Advantage Task (Supplementary Figure 3) 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.

Thirty-two participants (age: 24.34±2.18 years; 21 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 (*11*, *12*). We carried out a repeated-measures factorial ANOVA with Image Category (snake, gun, humanoid robot, and human), Image Match (same and different), and Eye Input (same and different eye) as within-subject factors, and with inverse efficiency as the dependent variable.

There was a significant main effect of Image Match ($F(1,31) = 5.70$, $p = 0.017$), no main effect of Image Category nor Eye Input (all $p > 0.18$), and no significant interaction (all $p > 0.75$). In the planned paired-sample t-tests, we compared the differences between same and different eye input for each image category. As shown in Fig. 2b, in the different condition of Image Match, the inverse efficiency in same condition of Eye Input was smaller than in different condition of Eye Input for snake images ($t_{31} = -2.21$, $p = 0.035$, Cohen's $d = 0.39$), and for humanoid robot images ($t_{31} = -2.23$, $p = 0.028$, Cohen's $d = 0.41$), but not for human images or gun images (all $p > 0.24$). In the same condition of Image Match, the inverse efficiency differences between Eye Input were not significant for any image category (all $p > 0.10$). 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 different condition of Image Match 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 ($t(126) = 2.185$, $p = 0.031$, beta = 0.191, CI = [0.105, 2.122]). 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 ($p = 0.017$, Fig. 2c).

Greater amygdala activity induced by humanoid robot images compared to human images under unconscious presentation. Sixty-one participants (age: 21.03±2.43 years; 39 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-five (age: 20.44±2.00 years; 16 females) out of the 61 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 Figure 4), 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 Figure 5, 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 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 bilateral amygdala (Fig. 4a) under unconscious presentation (left amygdala: t_{60} = 3.76, $p < 0.001$, Cohen's d = 0.48; right amygdala: t_{60} = 3.32, $p < 0.005$, Cohen's d = 0.42; Fig. 4b); no such difference was observed upon conscious presentation (left amygdala: t_{60} = - 0.61, p = 0.55; right amygdala: t_{60} = - 0.79, p = 0.43). The activation differences between humanoid robot and human images under unconscious presentation in bilateral amygdala were significantly stronger than activation differences under conscious presentation (left amygdala: t_{60} = 3.34, $p < 0.001$, Cohen's d = 0.43; right amygdala: t_{60} = 2.67, $p < 0.01$, Cohen's d = 0.34; Fig. 4b). Our results indicate that greater amygdala activity is induced by humanoid robot images compared to human images under unconscious presentation.

Neural activity and functional connectivity changed 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_{24} = - 2.46, $p < 0.05$, Cohen's d = 0.49; Fig 4c), indicating that participants' negative implicit attitude to humanoid robots had been successfully weakened by the Evaluative Conditioning Task. We then analyzed neuroimaging data for the change of neural activity and functional connectivity after successfully weakening negative attitude. A whole-brain analysis demonstrated a significant time main effect (pre-test, post-test) in the activation of the right DLPFC (uncorrected $p < 0.005$; Supplementary Figure 6). Previous studies defined ROIs of bilateral DLPFC as a sphere with a radius of 5mm centered at specific coordinates (Talairach coordinates; left: x = - 47, y = 17, z = 28; right: x = 47, y = 20, z = 26) (*26*). Our ROI analysis revealed that bilateral DLPFC activation differences between humanoid robot and human images marginal significantly decreased in post-test compared to pre-test (left: t_{24} = 1.64, $p = 0.06$ (one-tailed), Cohen's d = 0.33; right: t_{24} = 1.92, $p < 0.05$ (one-tailed), Cohen's d = 0.38; Supplementary Figure 7a). Importantly, the activation change (post-test vs. pre-test) in the right DLPFC was positively correlated with IAT scores change (post-test vs. pre-test) ($r = 0.40$, $p < 0.05$; Supplementary Figure 7b).

We then used psychophysiological interaction analysis with a seed in the 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 right amygdala and the right hippocampus (Fig. 5a), and thalamus (Supplementary Figure 8). We conducted a two factor (time factor: pre-test, post-test; image factor: humanoid robot, human) repeated measure ANOVA on functional connectivity under unconscious presentation of right amygdala-right hippocampus, and right amygdala-thalamus. There were significant time × image interaction effects in both couplings (amygdala-hippocampus: $F_{(1, 24)}$ = 11.45, $p < 0.001$, Fig 5b; amygdala-thalamus: $F_{(1, 24)}$ = 16.05, $p < 0.001$). There was a significant time main effect ($F_{(1, 24)}$ = 10.47, $p < 0.01$) in the right amygdala-right hippocampus. There was a marginal significant image main effect ($F_{(1, 24)}$ = 3.80, $p = 0.054$) in the right amygdala-thalamus. Importantly, the right amygdala-right hippocampus connectivity change (post-test vs. pre-test) was negatively correlated with the IAT scores change (post-test vs. pre-test) ($r = -0.53$, $p < 0.01$; Fig. 5c). These findings suggest that a positive humanoid robot-related association has been established by associative learning as evidenced by results of successfully weakening the negative implicit attitude to humanoid robots and enhancing functional connectivity between the amygdala and hippocampus.

The amygdala response to unconscious presentation of humanoid robot images does not change after successfully weakening negative attitude. We next tested whether the amygdala response to unconscious

presentation of humanoid robot images changes after successful weakening negative implicit attitude. We conducted a two factor (time factor: pre-test, post-test; presentation factor: unconscious presentation, conscious presentation) repeated measure ANOVA on activation differences between humanoid robot and human images in the left and right amygdala. No significant time×presentation interaction effects were found in any brain region (all $p > 0.39$). There were no significant time or presentation main effects in the left or right amygdala (all $p > 0.55$). *Post-hoc* analysis revealed no changes in bilateral amygdala activation between humanoid robot and human images under the conscious or unconscious presentations between the post-test and pre-test scans (all $p > 0.48$; Fig. 4d). Correlation analyses revealed that there was a significant correlation between IAT scores change and activation value change in the left amygdala under conscious presentation ($r = 0.39$, $p = 0.050$). However, we found no significant correlation between IAT scores change and activation value change in bilateral amygdala under unconscious presentation (all $p > 0.13$). These results demonstrate that despite a positive humanoid robot-related association had been established by associative learning, the amygdala could still automatically and quickly detect humanoid robots.

Discussion

We found automaticity for processing information about humanoid robots that were previously only evident in 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 amygdala response to unconsciously presented humanoid robots indicates the automatic and quick detection of humanoid robots. Despite a positive humanoid robot-related association has been established by associative learning, the amygdala can still automatically and quickly detect humanoid robots.

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 (*11*). Inputs from one eye are mostly segregated throughout the subcortical visual pathway (*27*), whereas inputs from the two eyes appear to be integrated to a greater extent in the extrastriate cortex (*28*). 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 (*11*), 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 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 (*10*). Studies have well documented that the amygdala responds selectively to evolutionary threats, irrespective of the affective valence, such as animate entities (*21, 29, 30*) and depictions of humans (*31-33*). 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 (*15*). 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 (*15-17*). Our present study found that greater amygdala activity was induced by humanoid robot images compared to human images under unconscious presentation. These results potentially reflect the automaticity with rapid recruitment of amygdala for humanoid robot-related stimuli processing.

It has been proposed that there is enhanced resistance to extinction to evolutionary threats relative to modern threats (*10*). Modern threatening stimuli like pictures of guns have not induced the same resistance to extinction as do pictures of snakes (*19, 20*). Our study showed that associative learning successfully weakened participants' negative implicit attitude to humanoid robots and showed that amygdala-hippocampus functional connectivity was significantly increased after this associative learning. Furthermore, the implicit attitude change was negatively correlated with the functional connectivity change in the extent of amygdala-hippocampus coupling. The amygdala processes emotional information and uses it for associative learning (*34*). The hippocampus is important for the consolidation of information, including short-term and long-term memory (*35*). The connectivity of amygdala-hippocampal circuit is understood to underlie the neural basis of emotion associative learning (*35, 36*). Our results indicate that a positive humanoid robot-related memory has been established, which is competitive with the priori negative humanoid robot-related memory. However, the amygdala could still automatically and quickly detect humanoid robots, indicating that the amygdala still regard humanoid robots as a threat.

An enormous amount of literature has focused on the psychological perspective for how humanoid robots are perceived. In many science fiction accounts, humanoid robots are presented as perfect soldiers who never tire or as ideal servants who always obey (*37*). In reality, people long balk at the idea that robots have any human-like mental capacities. But research emphasizes that people's minds are matters of perception (*38*), and we perceive robots based on the ascriptions such as some ability to think, remember, and exert self-control (*39*). The more human-like a robot looks, the more people perceive it as having mental capacities, a phenomenon called anthropomorphism (*40*). Besides, the appearances of robots also convey meaning. For instance, people attribute species to some robots. Aibo is very clearly a robot dog, while Justo Cat is a robot cat. It is noteworthy that people possibly attribute race to humanoid robots (*41*). The appearances of these robots result that people perceive the robots as mechanical versions of these animals or humans. We are entering an age where there will likely be a new type of entity that combines some properties of machines with some apparently psychological capacities that were previously only evident in humans (*42*). Combined with our demonstration that humans may perceive humanoid robots as an evolutionary threat, future humanoid robots with superhuman intelligence may be perceived as a new "human race" who will fight for their own rights and enslave us. With humanoid robots, we should consider

security and ethics as early as possible, and put these considerations into the technologies we develop.

In summary, this study demonstrates that humanoid robots are perceived as an evolutionary threat, even though humanoid robots are manmade, modern objects. Our findings provide new insights into the perception of humanoid robots from an evolutionary perspective, which can help to redefine human-robot interaction and inform robot ethics.

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 and DRZ provided administrative, technical, or material support. 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. G. Timo, A. Markus, Are robots becoming unpopular? Changes in attitudes towards autonomous robotic systems in Europe. *Computers in human behavior* **93**, 53-61 (2019).
2. Z. Baobao, A. Dafoe, Artificial Intelligence: American Attitudes and Trends. *Oxford, UK: Center for the Governance of AI, Future of Humanity Institute, University of Oxford*, (2019).
3. Y. H. Liang, S. A. Lee, Fear of Autonomous Robots and Artificial Intelligence: Evidence from National Representative Data with Probability Sampling. *International Journal Of Social Robotics* **9**, 379-384 (2017); published online EpubJun (10.1007/s12369-017-0401-3).
4. E. Fast, E. Horvitz, Long-Term Trends in the Public Perception of Artificial Intelligence. *Thirty-First Aaai Conference on Artificial Intelligence*, 963-969 (2017).
5. E. Broadbent, I. H. Kuo, Y. I. Lee, J. Rabindran, N. Kerse, R. Stafford, B. A. MacDonald, Attitudes and reactions to a healthcare robot. *Telemedicine journal and e-health : the official journal of the American Telemedicine Association* **16**, 608-613 (2010); published online EpubJun (10.1089/tmj.2009.0171).
6. S. Mineka, A. Ohman, Phobias and preparedness: The selective, automatic, and encapsulated nature of fear. *Biological psychiatry* **52**, 927-937 (2002); published online EpubNov 15 (Pii S0006-3223(02)01669-4

Doi 10.1016/S0006-3223(02)01669-4).
7. B. Subra, D. Muller, L. Fourgassie, A. Chauvin, T. Alexopoulos, Of guns and snakes: testing a modern threat superiority effect. *Cogn Emot* **32**, 81-91 (2018); published online EpubFeb (10.1080/02699931.2017.1284044).
8. D. C. Blanchard, R. J. Blanchard, Ethoexperimental approaches to the biology of emotion. *Annual review of psychology* **39**, 43-68 (1988)10.1146/annurev.ps.39.020188.000355).
9. S. Paradiso, Affective neuroscience: The foundations of human and animal emotions. *Am J Psychiat* **159**, 1805-1805 (2002); published online EpubOct (DOI 10.1176/appi.ajp.159.10.1805).
10. A. Ohman, S. Mineka, Fears, phobias, and preparedness: Toward an evolved module of fear and fear learning. *Psychological review* **108**, 483-522 (2001); published online EpubJul (10.1037//0033-295X.108.3.483).
11. M. D. Vida, M. Behrmann, Subcortical Facilitation of Behavioral Responses to Threat. *Scientific reports* **7**, 13087 (2017); published online EpubOct 12 (10.1038/s41598-017-13203-8).
12. S. Gabay, A. Nestor, E. Dundas, M. Behrmann, Monocular advantage for face perception implicates subcortical mechanisms in adult humans. *Journal of cognitive neuroscience* **26**, 927-937 (2014); published online EpubMay (10.1162/jocn_a_00528).
13. D. H. Zald, The human amygdala and the emotional evaluation of sensory stimuli. *Brain research. Brain research reviews* **41**, 88-123 (2003); published online EpubJan (
14. M. Tamietto, B. de Gelder, Neural bases of the non-conscious perception of emotional signals. *Nature Reviews Neuroscience* **11**, 697-709 (2010); published online EpubOct (10.1038/nrn2889).
15. W. A. Cunningham, M. K. Johnson, C. L. Raye, J. Chris Gatenby, J. C. Gore, M. R. Banaji, Separable neural components in the processing of black and white faces. *Psychological science* **15**, 806-813 (2004); published online EpubDec (10.1111/j.0956-7976.2004.00760.x).

16. L. M. Williams, B. J. Liddell, A. H. Kemp, R. A. Bryant, R. A. Meares, A. S. Peduto, E. Gordon, Amygdala-prefrontal dissociation of subliminal and supraliminal fear. *Human brain mapping* **27**, 652-661 (2006); published online EpubAug (10.1002/hbm.20208).

17. K. Felmingham, A. H. Kemp, L. Williams, E. Falconer, G. Olivieri, A. Peduto, R. Bryant, Dissociative responses to conscious and non-conscious fear impact underlying brain function in post-traumatic stress disorder. *Psychological medicine* **38**, 1771-1780 (2008); published online EpubDec (10.1017/S0033291708002742).

18. Z. Y. Fang, H. Li, G. Chen, J. J. Yang, Unconscious Processing of Negative Animals and Objects: Role of the Amygdala Revealed by fMRI. *Frontiers in human neuroscience* **10**, (2016); published online EpubApr 5 (Artn 146 10.3389/Fnhum.2016.00146).

19. E. W. Cook, 3rd, R. L. Hodes, P. J. Lang, Preparedness and phobia: effects of stimulus content on human visceral conditioning. *Journal of abnormal psychology* **95**, 195-207 (1986); published online EpubAug (10.1037//0021-843x.95.3.195).

20. K. Hugdahl, B. H. Johnsen, Preparedness and electrodermal fear-conditioning: ontogenetic vs phylogenetic explanations. *Behaviour research and therapy* **27**, 269-278 (1989)10.1016/0005-7967(89)90046-6).

21. J. J. Yang, P. S. F. Bellgowan, A. Martin, Threat, domain-specificity and the human amygdala. *Neuropsychologia* **50**, 2566-2572 (2012); published online EpubSep (10.1016/j.neuropsychologia.2012.07.001).

22. S. Cave, K. Dihal, Hopes and fears for intelligent machines in fiction and reality. *Nature Machine Intelligence* **1**, 74-78 (2019)10.1038/s42256-019-0020-9).

23. E. Broadbent, R. Tamagawa, A. Patience, B. Knock, N. Kerse, K. Day, B. A. MacDonald, Attitudes towards health-care robots in a retirement village. *Australasian journal on ageing* **31**, 115-120 (2012); published online EpubJun (10.1111/j.1741-6612.2011.00551.x).

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

25. S. Cave, The Problem with Intelligence: Its Value-Laden History and the Future of AI. *In AIES '20: AAAI/ACM Conference on AI, Ethics, and Society Proceedings, February 7–8, 2020, New York*, 29-35 (2020)10.1145/3375627.3375813).

26. R. C. Wolf, R. J. Herringa, Prefrontal-Amygdala Dysregulation to Threat in Pediatric Posttraumatic Stress Disorder. *Neuropsychopharmacology : official publication of the American College of Neuropsychopharmacology* **41**, 822-831 (2016); published online EpubFeb (10.1038/npp.2015.209).

27. R. S. Menon, S. Ogawa, J. P. Strupp, K. Ugurbil, Ocular dominance in human V1 demonstrated by functional magnetic resonance imaging. *Journal of neurophysiology* **77**, 2780-2787 (1997); published online EpubMay (10.1152/jn.1997.77.5.2780).

28. H. Bi, B. Zhang, X. Tao, R. S. Harwerth, E. L. Smith, 3rd, Y. M. Chino, Neuronal responses in visual area V2 (V2) of macaque monkeys with strabismic amblyopia. *Cerebral cortex* **21**, 2033-2045 (2011); published online EpubSep (10.1093/cercor/bhq272).

29. F. Mormann, J. Dubois, S. Kornblith, M. Milosavljevic, M. Cerf, M. Ison, N. Tsuchiya, A. Kraskov, R. Q. Quiroga, R. Adolphs, I. Fried, C. Koch, A category-specific response to animals in the right human amygdala. *Nature neuroscience* **14**, 1247-1249 (2011); published online EpubOct (10.1038/nn.2899).

30. U. Rutishauser, O. Tudusciuc, D. Neumann, A. N. Mamelak, A. C. Heller, I. B. Ross, L. Philpott, W. W. Sutherling, R. Adolphs, Single-Unit Responses Selective for Whole Faces in the Human Amygdala. *Current Biology* **21**, 1654-1660 (2011); published online EpubOct 11 (10.1016/j.cub.2011.08.035).
31. H. C. Breiter, N. L. Etcoff, P. J. Whalen, W. A. Kennedy, S. L. Rauch, R. L. Buckner, M. M. Strauss, S. E. Hyman, B. R. Rosen, Response and habituation of the human amygdala during visual processing of facial expression. *Neuron* **17**, 875-887 (1996); published online EpubNov (Doi 10.1016/S0896-6273(00)80219-6).
32. T. Wheatley, S. C. Milleville, A. Martin, Understanding animate agents - Distinct roles for the social network and mirror system. *Psychological science* **18**, 469-474 (2007); published online EpubJun (DOI 10.1111/j.1467-9280.2007.01923.x).
33. E. Bonda, M. Petrides, D. Ostry, A. Evans, Specific involvement of human parietal systems and the amygdala in the perception of biological motion. *Journal Of Neuroscience* **16**, 3737-3744 (1996); published online EpubJun 1 (
34. S. Maren, Long-term potentiation in the amygdala: a mechanism for emotional learning and memory. *Trends in neurosciences* **22**, 561-567 (1999); published online EpubDec (10.1016/s0166-2236(99)01465-4).
35. H. Eichenbaum, How does the hippocampus contribute to memory? *Trends in cognitive sciences* **7**, 427-429 (2003); published online EpubOct (10.1016/j.tics.2003.08.008).
36. P. J. Brasted, T. J. Bussey, E. A. Murray, S. P. Wise, Role of the hippocampal system in associative learning beyond the spatial domain. *Brain : a journal of neurology* **126**, 1202-1223 (2003); published online EpubMay (10.1093/brain/awg103).
37. C. Stephen, D. Kanta, Ancient dreams of intelligent machines: 3,000 years of robots. *Nature* **559**, 473-475 (2018)doi: 10.1038/d41586-018-05773-y).
38. K. Weisman, C. S. Dweck, E. M. Markman, Rethinking people's conceptions of mental life. *Proceedings of the National Academy of Sciences of the United States of America* **114**, 11374-11379 (2017); published online EpubOct 24 (10.1073/pnas.1704347114).
39. Y. E. Bigman, A. Waytz, R. Alterovitz, K. Gray, Holding Robots Responsible: The Elements of Machine Morality. *Trends in cognitive sciences* **23**, 365-368 (2019); published online EpubMay (10.1016/j.tics.2019.02.008).
40. E. J. de Visser, S. S. Monfort, R. McKendrick, M. A. B. Smith, P. E. McKnight, F. Krueger, R. Parasuraman, Almost Human: Anthropomorphism Increases Trust Resilience in Cognitive Agents. *J Exp Psychol-Appl* **22**, 331-349 (2016); published online EpubSep (10.1037/xap0000092).
41. C. Bartneck, Robots And Racism. *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction - HRI '18*, 196-204 (2018).
42. T. J. Prescott, Robots are not just tools. *Connection Science*, 142-149 (2017).

Figures

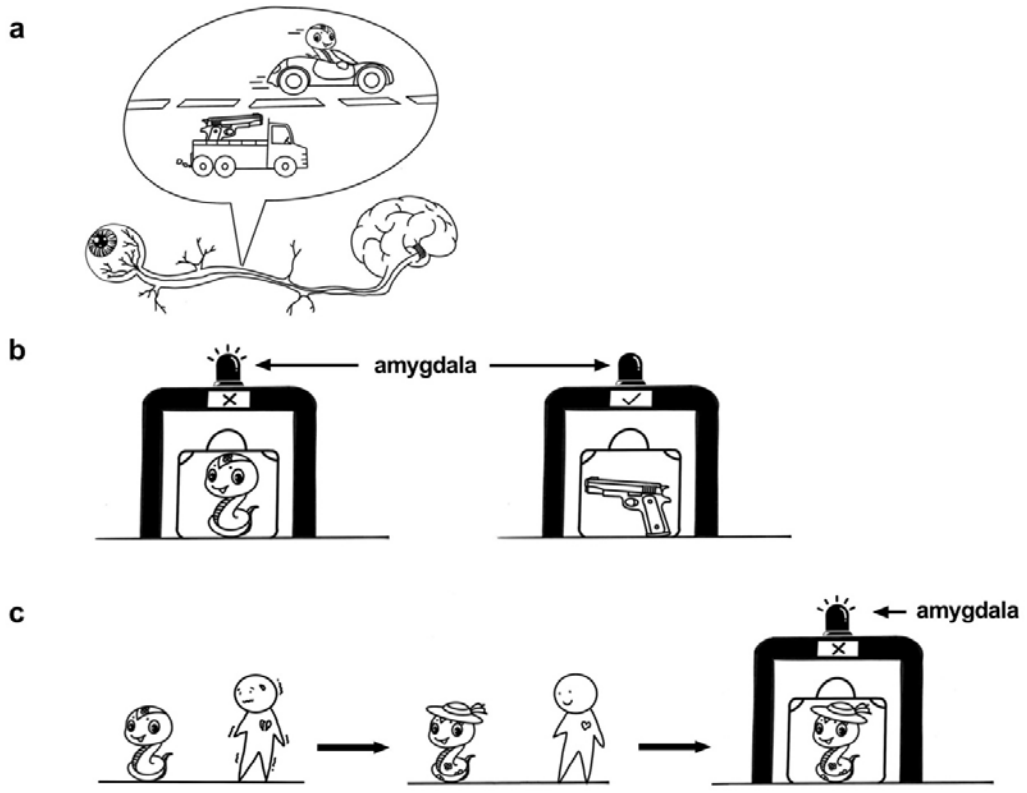

Figure 1. Automaticity for evolutionary threats as compared to modern threats. (a) Monocular advantage. The monocular input information of evolutionary threats (*e.g.*, snake) is faster than information of modern threats (*e.g.*, gun) to the brain for perception. (b) As the security devices can detect 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. (c) Although we can turn our fear for evolutionary threats into explicit liking by some way (such as associative learning), the amygdala can still respond to unconsciously presented images of evolutionary threats.

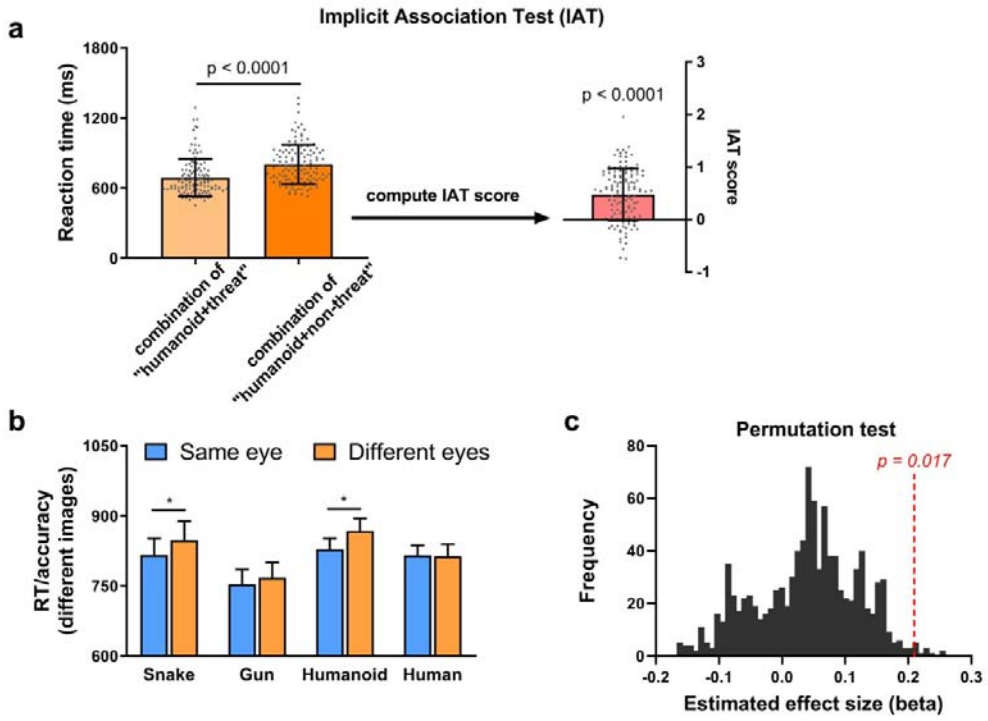

Figure 2. Implicit Association Test and Monocular Advantage Task. (a) Negative implicit attitude to humanoid robots. In the Implicit Association Test, participants responded faster in combination of "humanoid + threat" than in combination of "humanoid + non-threat", and the computed IAT scores (effect 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 different condition of Image Match, 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 different condition of Image Match 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 ($p = 0.017$). * $p < 0.05$.

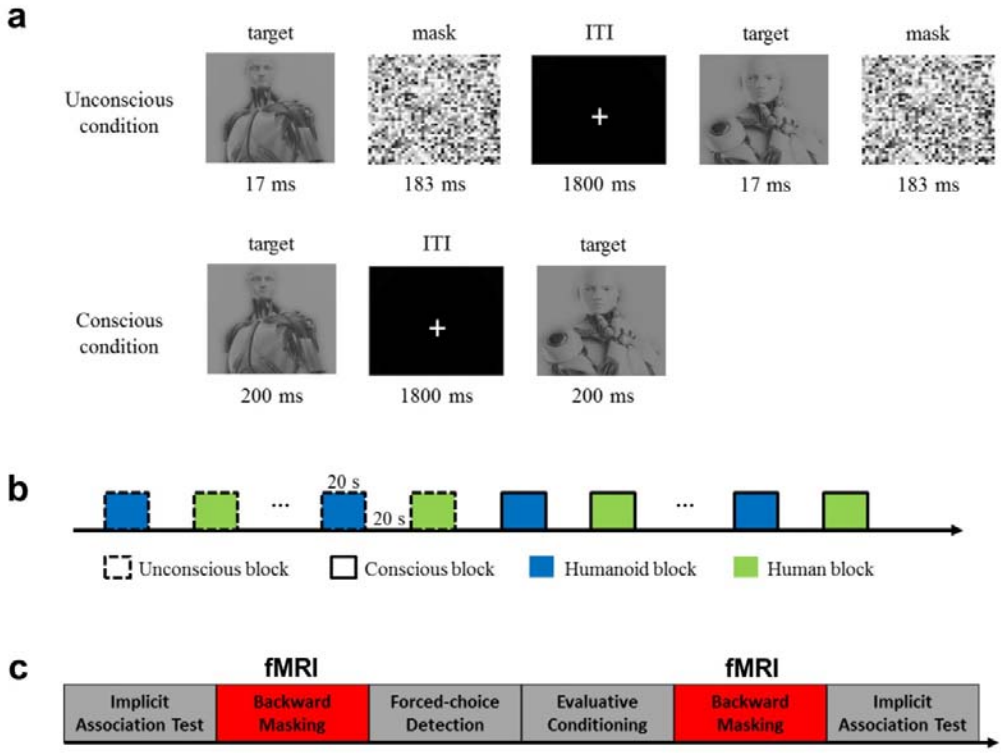

Figure 3. Description of the Backward Masking Task and the procedure for the fMRI experiment. (a) Time setting of the Backward Masking Task. In the unconscious condition, the target image was presented for 17 ms followed by a mask for 183 ms and a fixation for 1800 ms. In the conscious condition, the 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.

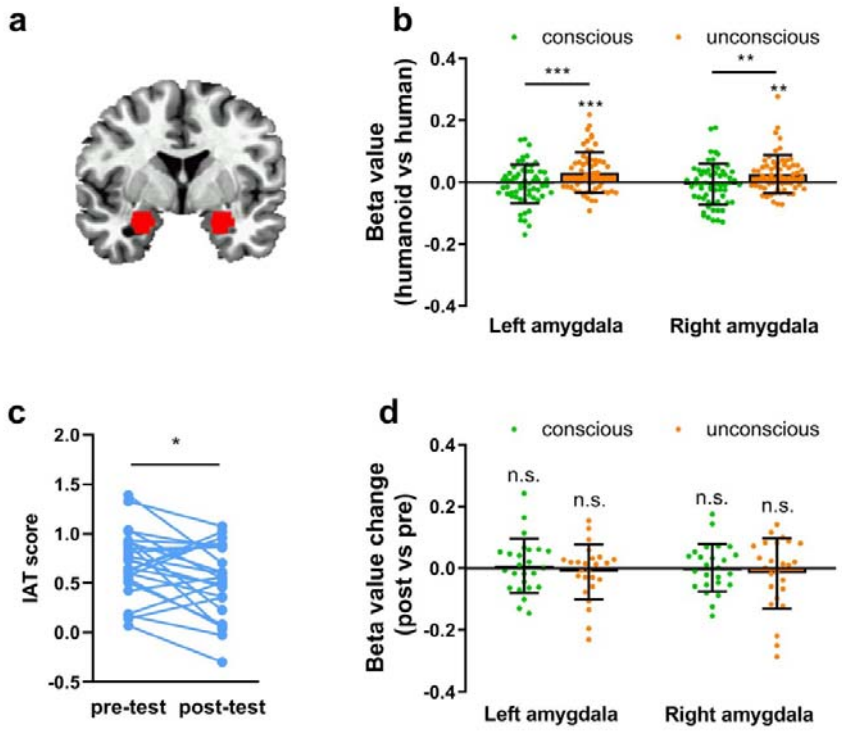

Figure 4. Humanoid robot image related amygdala activity and amygdala activity change. (a) The region of interest for bilateral amygdala. (b) Although no amygdala activity differences were detected for consciously presented humanoid robot vs. human images, greater amygdala activity was induced by humanoid robot images compared to images of humans under unconscious presentation. (c) Significantly smaller IAT scores were found in post-test compared to pre-test, indicating that participants' negative implicit attitude to humanoid robots was successfully weakened. (d) Despite successfully weakening the negative implicit attitudes to humanoid robots, the amygdala activity differences of humanoid robot vs. human images did not change under conscious or unconscious presentations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$, n.s. = not significant. For a and c, plotted data represent the mean ± SD. across participants. IAT = implicit association test.

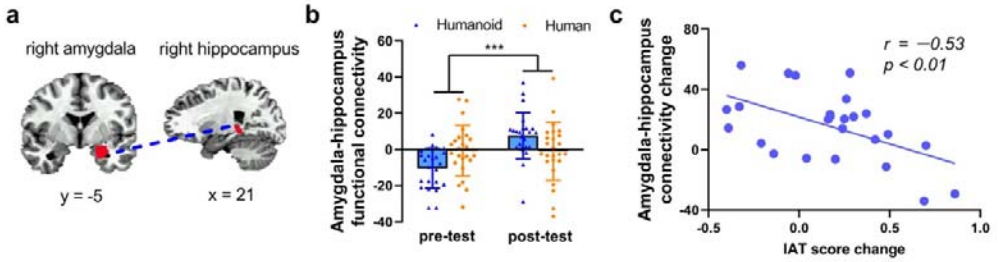

Figure 5. Amygdala-hippocampus functional connectivity was enhanced after attitude modulation. (a) A whole brain connectivity analysis demonstrated a significant time main effect in functional connectivity of right amygdala and right hippocampus. (b) Amygdala-hippocampus functional connectivity was enhanced in the post-test compared to the pre-test. We conducted a two factor (time factor: pre-test, post-test; image factor: humanoid, human) repeated measure ANOVA on amygdala-hippocampus functional connectivity under unconscious presentation. There was a significant time × image interaction effect ($F_{(1, 24)} = 11.45$, $p < 0.001$) and a significant time main effect ($F_{(1, 24)} = 10.47$, $p < 0.01$), but no significant image main effect ($F_{(1, 24)} = 0.51$, $p = 0.82$). (c) IAT score change was negatively correlated with the change of amygdala-hippocampus functional connectivity ($r = -0.53$, $p < 0.01$). For b, plotted data represent the mean ± SD. across participants. IAT = Implicit Association Test.