Interacting with agents without a mind: the case for artificial agents

Rebecca Geiselmann[1,2], Afroditi Tsourgianni[5], Ophelia Deroy[2,3,4], and Lasana T. Harris[5,6]

[1] Graduate School of Systemic Neurosciences, Faculty for Biology, Ludwig Maximilian University of Munich, Großhadernerstr. 2, 82152 Planegg

[2] Faculty of Philosophy, Ludwig Maximilian University of Munich, Geschwister-Scholl-Platz 1, 80539 Munich

[3] Munich Center for Neuroscience, Ludwig Maximilian University of Munich, Großhadernerstr. 2, 82152 Planegg

[4] Institute of Philosophy, School of Advanced Study, University of London, Malet Street, London, WC1E

[5] Department of Experimental Psychology, University College London, 26 Bedford Way, London WC1H

[6] Alan Turing Institute for Data Science and Artificial Intelligence, 96 Euston Road, London, NW1 2DB

Corresponding author:
Rebecca Geiselmann, r.geiselmann@campus.lmu.de

Abstract

Humans may deprive each other of human qualities if the social context encourages it. But what about the opposite: do people attribute human traits to non-human entities without a mind, such as Artificial Intelligence (AI)? Perceived humanness is based on the assumption that the other can act (has agency) and has experiences (thoughts and feelings). This review shows that AI fails to fully elicit these two dimensions of mind perception. Embodied AI may trigger agency attribution but only humans trigger the attribution of experience. Importantly, people are more likely to attribute mind in general and agency specifically to AI that resembles the human form. Lastly, people's pre-dispositions and the social context affect people's tendency to attribute human traits to AI.

Introduction

In 2017, Sophia, a humanoid robot, was granted citizenship - a fundamental human right [1]. Five years later, one of Google's engineers reported that the company's Artificial Intelligence (AI) had become sentient [2]. Although peculiar, these events highlight that AI's abilities, rights, responsibilities, and societal roles remain ambiguous: do people consider AI as a machine or a human-like agent when interacting with it?

People predict the actions of humans based on the fundamental assumption that they are intentional agents that have a mind [3]. However, the mind is in the eye of the beholder, which means that it can be withdrawn from human agents (i.e. dehumanisation) but also ascribed to non-human agents (i.e. anthropomorphism) based on cognitive or motivational features associated with the perceiver, as well as physical and behavioural features of the perceived entity [4].

Since its beginning in 1956, AI has embraced the idea of simulating (i.e., imitating with the use of models) human intelligence, including scientific knowledge, common sense, and self-improvement [5,6]. Some see human intelligence as a property of internal thought processes and reasoning, while others focus on intelligent behaviour as an external characteristic [7]. Intended to be a machine that thinks or acts like humans, it has a growing impact on humans' social and individual lives: from almost undetectable algorithms that execute tasks on our behalf (e.g. setting prices in online markets) to those we are more aware of because they voice personalised advice (e.g. home assistant 'Alexa')[8]. Embodied AI, such as social robots, is even more noticeable and designed to interact closely with humans as helpers and companions in public places such as supermarkets, education, health care, and retirement homes [9]. Social robots do not only simulate humans in the way they think or act but also in their looks. The goal is for people to interact with them more intuitively and naturally [10].

While AI simulates a range of human-like features, its ontological status anchors them as machines or non-human agents. Nevertheless, Cockelberg (2011) pointed out that ontology matters less for society and ethics than how AI is anthropomorphised – the degree to which it *appears* to people (non-expert users) as human agent [11]. Here, we explore whether and when people consider – perceive, understand, predict, and manipulate – AI as non-human or human agent, utilising the socio-cognitive and interactive repertoire reserved for humans.

Anthropomorphising non-human entities

People predict the actions of other human agents based on the fundamental assumption that they have a mind. Daniel Dennett (1987) defined this strategy as the intentional stance and contrasted it with two more basic stances used for prediction: the physical and the design stance [12]. The behaviour of every physical system, e.g. the trajectory of a thrown ball, is subject to the laws of physics. Therefore, people can predict its behaviour by the physical stance, referring to causal-mechanical relationships. The design stance allows people to make predictions based on the assumption that systems, e.g. an alarm clock, work as they are meant to by design. In certain cases, these two strategies do not suffice: it may not be practical to predict how rational agents (e.g. humans) will behave based on these two stances. Therefore, people adopt the intentional stance, which relies on the attribution of mental states such as intentions, beliefs, or desires.

However, intentionality attribution does not necessarily imply that the other has genuine mental states in the human sense. People may also treat *non-human* entities 'as if' they had a mind to manage social interactions with them. This may "give us the predictive power we can get by no other method" (Dennett, 1981, p. 66) [13]. Here we argue that one of the most fundamental factors that contribute to anthropomorphism is the non-human entity's ability to trigger the intentional stance.

The human tendency to anthropomorphise is so strong that people even readily perceive mind in animations of moving abstract shapes (e.g. disks and triangles), given they engage in self-propelled and goal-directed motion [14–16]. This tendency makes sense as being human is what people know [17]. When interacting with unfamiliar non-human entities, people may use their knowledge of themselves as a basis for understanding them. In other words, to understand the actions of these entities, people may automatically simulate similar actions in their cognitive system. Additionally, people are used to inferring other humans' mental states to understand and predict their actions. Consequently, people may apply the same strategy to simplify unexplainable actions from non-human entities. Lastly, anthropomorphism may give people an increased sense of belonging and control in ambiguous contexts [18].

Anthropomorphism and specifically mind attribution are automatic processes that activate socio-cognitive processes in a bottom-up way, primarily driven by human-like features and biological motion [15,19–22]. Therefore, especially embodied AI (e.g. robots, androids, or avatars) may have the potential to trigger anthropomorphism, with human-like embodied AI (i.e. humanoids) having the highest potential.

Do interactions with embodied AI trigger anthropomorphism?

Self-report studies reveal that people indeed attribute mind to embodied AI [23–25] and thus anthropomorphise them. When provided with verbal descriptions of robot and human behaviours across different contexts and instructed to explain why the agent engaged in the behaviour, people used the same conceptual toolbox of behavioural explanations for human and

robot agents [26]. People even ascribed the same level of mind to humanoids and human agents when asked to rate them from different images and verbal descriptions [27].

According to Gray, Gray, and Wegner (2007) attributing mind consists of two dimensions: the capacity for agency (covering one or several of the following capacities: self-control, morality, memory, emotion recognition, planning, communication, and thought) and the capacity for experience (covering one or several of the following capacities: hunger, fear, pain, pleasure, rage, desire, personality, consciousness, pride, embarrassment, and joy) [28]. Humans are willing to treat embodied AI (here robots) as entities with some degree or kind of agency but are reluctant to perceive them as entities that can experience mental states.

This is in line with neuroscience research. Interactions with embodied AI seem to elicit action perception and representation mechanisms [29], such as motor resonance [30], motor contagion (a behavioural manifestation of action representation) [31], and the vicarious sense of agency (dependent on action representation) [32]. Action perception and representation are based on visual cues and only elicited when the other is assumed to have the capacity to act i.e. has *agency*. In contrast, research suggests that interactions with embodied AI, including humanoids, do not activate the mentalising network in the same way as humans do. Brain areas that are less or not activated by embodied AI include the TPJ [33–36], mPFC [35], and dPFC [34]. The activation of this network reflects the inference of the other's mental states and thus the assumption that the other can *experience* mental states.

Importantly, people ascribed high agency and experience to an adult. However, they attributed high experience but no agency to a baby and high agency but no experience to god. As no one would argue that babies are not human, experiencing mental states seems to be considered more 'uniquely' human than the ability to act [28]. Even with other humans, people tend to demonstrate lower concern for others' mental states as a function of subjective perception of distance [37]. Therefore, people likely perceive embodied AI as an agent whose actions are relevant to be predicted but yet consider it as distant, outgroup member that belongs to the group of 'non-humans'[38]. Does the same hold true for disembodied AI (e.g. chatbots or algorithms) and what effect does the human-like form have on athropomorphism?

Is the human model responsible for AI's anthropomorphism?

Could it be that the human-like form is fully responsible for anthropomorphism and agency attribution - meaning that, short of a human-like face, body, or motion, other AI is perceived like machines? Brain areas responsible for motor resonance (here the mirror-neuron system) also responded to non-humanoid (industrial) robots' movements [39], indicating that action representation mechanisms are also employed when interacting with robots that move in a less human-like manner. However, the degree of action representation depends on motion kinematics [31].

The degree of the AI's human-like appearance also affects people's mind attribution toward a robot. For instance, people rated a robot with a face on its screen as more "minded" compared to robots with no or a silver display [40]. People's politeness norms were also triggered more often by a humanoid robot than a non-humanoid (mechanic) robot [41], suggesting that people ascribe

more intentionality to human-like robots. This is also reflected in the action representation system: While action representation is also elicited by the movements of industrial (non-humanoid) robots' movements, the degree of action representation depends on the robot's physical appearance [42].

Coordination in more natural scenarios (here playing a musical duet together) with artificial agents (a humanoid robot compared to a computer algorithm that either committed human-like or machine-like errors), as well as social inclusion (in a ball tossing game) - a reflection of social connection and prosociality - was also influenced by how human-like the agent appears both in terms of morphological traits and in terms of behaviour [43]. Overall, this suggests that people are more likely to attribute mind, specifically agency, to humanoid than non-humanoid AI.

What about interactions with disembodied AI that does not present a visual form? Embodied AI engages visual perception, such as processing faces [44], gestures [45], and eye-contact [46] in a similar way as humans do, which facilitates action representation and prediction [29–32]. Short of a visual form, disembodied AI does not activate visual perception and action representation, possibly making agency attribution less likely. Interestingly, similar to interactions with embodied AI, interactions with disembodied AI do not activate the mentalising network across various paradigms - only humans do [47–56]. Moreover, two reviews reveal that while interactions with embodied AI do not activate mental state inference processes, they drive more engagement of some social brain regions relative to human-human interactions, such as the vmPFC [57,58]. Interactions with disembodied AI, in turn, do not lead to increased activation of any social brain regions relative to humans.

Overall, AIs are not homogenous, the distinction between humanoid, embodied, and disembodied AI reveals that people are most likely to attribute mind in general and agency more specifically to AI that mimics human appearance and motion.

Do factors beyond the human model facilitate AI's anthropomorphism?

Besides factors related to the agent and how much it activates a human stereotype [59], anthropomorphism depends on people's individual pre-dispositions [17], i.e. different personality phenotypes have different tendencies to anthropomorphise. Concerning human-AI interactions, the higher the openness to experience and the higher the level of agreeableness, the stronger the mind attribution toward the AI (here a robot). This effect was mediated by the attitudes toward AI [60].

Importantly, also social contexts and structures affect anthropomorphism, including mind attribution toward AI. Perceived in-group membership with a robot resulted in a greater extent of implicit anthropomorphic inferences and willingness to interact with robots in general [61]. Moreover, observing someone interacting socially with a robot can enhance the adoption of an intentional stance. For instance, people who 'collaborated' with a humanoid were more likely to ascribe intentions toward it after the interaction than people who did not collaborate with it [62]. The attribution of intentional traits toward a robot was also higher after social compared to non-social priming. In the social priming condition, before evaluating different types of robots, participants were told that the robots represent types of agents that they will interact with in the coming decades [60].

Culture might also shape people's perceptions of robots. Japanese participants found a robot's intervention in a moral dilemma more morally permissible than a human's intervention, whereas there was no difference among U.S. participants [63]. Lastly, social contextualisation and embodiment may also interact: People less readily saw disembodied AI as embedded in social structures than humans and explained and justified the AI's actions when solving a moral (lethal strike) dilemma differently [64].

Altogether, an interplay between the AI's appearance and motion, people's pre-dispositions, and the social context determines the degree of anthropomorphism in human-AI interactions.

Does anthropomorphism affect human-AI interactions?

Generally, treating others as agents with a mind facilitates social connection and prosocial behaviour, such as decreased cheating and increased generosity [65–68]. This also seems to hold true for human-robot interactions. For instance, attributing a mind to a robot decreased aggressive behaviour towards it [69]. Moreover, mind attribution increases the social relevance ascribed to others' actions. For example, participants follow the eye movements of a robot more strongly when they are believed to reflect intentional compared to pre-programmed or random behaviour [70–72]. However, automatic anthropomorphism or mind attribution can also have negative consequences. When an agent is difficult to categorise as human or non-human, people's response times increase and accuracy rates decrease [73,74], suggesting cognitive conflict processing. Resolving this cognitive conflict can have negative effects on performance when interacting with these difficult to categorise agents.

AI does not represent both mind perception dimensions as humans do. This is also reflected in human-AI compared to human-human interactions. Agents that display a high degree of agency, but only a low degree of experience are labeled as 'moral agents' with full moral responsibilities and the ability to show intentional behavior, in particular when it is harmful [28]. As a consequence, they are more likely to be harmed by others [75,76], denied moral rights, and judged more harshly for behaviors that lead to negative consequences [28].

For instance, people provide different moral judgments of robots and humans in the case of moral dilemmas. Robots were blamed more than human agents when they failed to intervene in a situation in which they could save multiple lives but had to sacrifice one person's life [77]. These findings seem to be independent of the participant's cultural background, as similar responses were observed in the US and Japan [63]. Harmful behaviour is also expressed towards disembodied AI: In line with prior research [78–81], across five different economic games, people were less inclined to cooperate with AI agents than with anonymous humans when it was individually but not mutually advantageous to defect. Algorithm exploitation proved to be the main driver: the effect was not driven by a competitive wish to end up better off than the AI but came from accepting the decision to act selfishly and leave the AI agent less well-off [82].

In sum, anthropomorphism can improve human-AI interactions, unless the AI is difficult to categorise as either human or machine, which takes up cognitive resources. Moreover, as people tend to ascribe agency to AI, AI is treated as a moral agent with full moral responsibilities.

Conclusion

The literature review highlights differences in mind attribution towards AI compared to humans. Firstly, it seems that AI does not fully represent both mind perception dimensions. At the current state, only interactions with human agents (adults) are attributed both agency and experience. This is also reflected in human-AI interactions with AI being treated as moral agents. Secondly, AIs are not homogenous, the distinction between humanoid, embodied, and disembodied AI reveals that people are more likely to attribute mind in general and agency more specifically to AI that resembles the human form. This also provides insights into what may increase the attribution of experience, namely an embodied form. Thirdly, beyond the human form, people's pre-dispositions and the social context affect AI's anthropomorphism. This being said it remains to be investigated whether considering AI as human will involve a shifting of the human stereotype or an error due to a cognitive limitation of how the brain enables interaction with AI.

If allowed/required:

BOX:

People understand others as entities with intentionality when attributing to them beliefs, desires, or intentions to make sense of their behaviour (i.e., intentional stance). They also perceive them as entities with phenomenal experiences attributing to them emotions, moods, or pain, which is also referred to as the phenomenal stance. Some authors [83] posit that moral concern for others emerges when we consider others as a subject of phenomenal experience. Adoption of a phenomenal stance would further promote affiliation, social interaction, and cooperation. Do people take a phenomenal stance when interacting with embodied AI?

Behavioural research suggests that people empathise with embodied AI. For example, watching a robot express fear of losing its memory and then observing it lose its memory induced more self-reported empathy than a control condition in which memory was not lost [84]. Neuroscience supports these findings [85,86]. When participants observed painful actions towards robots, the same patterns of neural activation in the ACC were found [87] that were observed when viewing pictures of painful vs. non-painful human stimuli [88–90]. Hence, people may take the phenomenal stance beyond the intentional stance when interacting with embodied AI.

Acknowledgments

We thank Jessica Brough, Jurgis Karpus, and Louis Longin for their input.

Funding: This work was supported by the European Innovation Council project 'EMERGE' 101070918, the Volkswagen Foundation project 'Co-Sense' and Bayrisches Forschungsinstitut für Digitale Transformation (bidt) project 'Co-Learn' (to OD) and the Alan Turing Institute, the UK's national institute for data science and artificial intelligence (to LH).

Conflict of interest or competing interest: none.

Declaration of interest: none.

References

1. Sini R: **Does Saudi robot citizen have more rights than women? BBC. Retrieved October 13, 2021**. 2017,
2. Tiku N: **The Google engineer who thinks the company's AI has come to life**. *Wash Post* 2022, **11**.
3. Dennett DC: *The intentional stance*. MIT press; 1987.
4. **Waytz: Causes and consequences of mind perception - Google Scholar**. [date unknown],
5. McCarthy J: **Artificial intelligence, logic and formalizing common sense**. In *Philosophical logic and artificial intelligence*. . Springer; 1989:161–190.
6. Haenlein M, Kaplan A: **A brief history of artificial intelligence: On the past, present, and future of artificial intelligence**. *Calif Manage Rev* 2019, **61**:5–14.
7. Russell S, Norvig P: **Artificial intelligence: a modern approach, global edition 4th**. *Foundations* 2021, **19**:23.
8. Köbis N, Bonnefon J-F, Rahwan I: **Bad machines corrupt good morals**. *Nat Hum Behav* 2021, **5**:679–685.
9. Fong T, Nourbakhsh I, Dautenhahn K: **A survey of socially interactive robots**. *Robot Auton Syst* 2003, **42**:143–166.
10. Broadbent E: **Interactions with robots: The truths we reveal about ourselves**. *Annu Rev Psychol* 2017, **68**:627–652.
11. Coeckelbergh M: **Humans, animals, and robots: A phenomenological approach to human-robot relations**. *Int J Soc Robot* 2011, **3**:197–204.
12. Dennett D: **Intentional systems theory**. 2009,
13. Dennett DC: **True believers: The intentional strategy and why it works**. 1981,
14. Heider F, Simmel M: **An experimental study of apparent behavior**. *Am J Psychol* 1944, **57**:243–259.
15. Opfer JE: **Identifying living and sentient kinds from dynamic information: The case of goal-directed versus aimless autonomous movement in conceptual change**. *Cognition* 2002, **86**:97–122.
16. Schultz J, Friston KJ, Wolpert DM, Frith CD: **Activation in superior temporal sulcus parallels a parameter inducing the percept of animacy**. In *28th European Conference on Visual Perception (ECVP 2005)*. . Pion Ltd.; 2005:62.
17. Epley N, Waytz A, Cacioppo JT: **On seeing human: a three-factor theory of anthropomorphism.** *Psychol Rev* 2007, **114**:864.
18. Harris LT, van Etten N, Gimenez-Fernandez T: **Exploring how harming and helping behaviors drive prediction and explanation during anthropomorphism**. *Soc*

Neurosci 2021, **16**:39–56.
19. Castelli F, Happé F, Frith U, Frith C: **Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns**. *Neuroimage* 2000, **12**:314–325.
20. Gao T, McCarthy G, Scholl BJ: **The wolfpack effect: Perception of animacy irresistibly influences interactive behavior**. *Psychol Sci* 2010, **21**:1845–1853.
21. Wheatley T, Weinberg A, Looser C, Moran T, Hajcak G: **Mind perception: Real but not artificial faces sustain neural activity beyond the N170/VPP**. *PloS One* 2011, **6**:e17960.
22. Schein C, Gray K: **The unifying moral dyad: Liberals and conservatives share the same harm-based moral template**. *Pers Soc Psychol Bull* 2015, **41**:1147–1163.
23. Wiese E, Metta G, Wykowska A: **Robots as intentional agents: using neuroscientific methods to make robots appear more social**. *Front Psychol* 2017, **8**:1663.
24. Marchesi S, Ghiglino D, Ciardo F, Perez-Osorio J, Baykara E, Wykowska A: **Do we adopt the intentional stance toward humanoid robots?** *Front Psychol* 2019, **10**:450.
25. Marchesi S, Spatola N, Perez-Osorio J, Wykowska A: **Human vs Humanoid. A behavioral investigation of the individual tendency to adopt the intentional stance**. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. . 2021:332–340.
26. De Graaf MM, Malle BF: **People's explanations of robot behavior subtly reveal mental state inferences**. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. . IEEE; 2019:239–248.
27. Thellman S, Silvervarg A, Ziemke T: **Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots**. *Front Psychol* 2017, **8**:1962.
28. Gray HM, Gray K, Wegner DM: **Dimensions of mind perception**. *science* 2007, **315**:619–619.
29. Wykowska A, Chellali R, Al-Amin M, Müller HJ: **Implications of robot actions for human perception. How do we represent actions of the observed robots?** *Int J Soc Robot* 2014, **6**:357–366.
30. Chaminade T, Rosset D, Da Fonseca D, Nazarian B, Lutcher E, Cheng G, Deruelle C: **How do we think machines think? An fMRI study of alleged competition with an artificial intelligence**. *Front Hum Neurosci* 2012, **6**:103.
31. Bisio A, Sciutti A, Nori F, Metta G, Fadiga L, Sandini G, Pozzo T: **Motor contagion during human-human and human-robot interaction**. *PloS One* 2014, **9**:e106172.
32. Roselli C, Ciardo F, De Tommaso D, Wykowska A: **Human-likeness and attribution of intentionality predict vicarious sense of agency over humanoid robot actions**. *Sci Rep* 2022, **12**:1–7.
33. Wang Y, Quadflieg S: **In our own image? Emotional and neural processing differences when observing human–human vs human–robot interactions**. *Soc Cogn Affect Neurosci* 2015, **10**:1515–1524.
34. Rauchbauer B, Nazarian B, Bourhis M, Ochs M, Prévot L, Chaminade T: **Brain activity during reciprocal social interaction investigated using conversational robots as control condition**. *Philos Trans R Soc B* 2019, **374**:20180033.
35. Hmamouche Y, Ochs M, Prévot L, Chaminade T: **Neuroscience to Investigate Social Mechanisms Involved in Human-Robot Interactions**. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*. . 2020:52–56.
36. Kelley MS, Noah JA, Zhang X, Scassellati B, Hirsch J: **Comparison of human social brain activity during eye-contact with another human and a humanoid robot**.

Front Robot AI 2021, **7**:599581.
37. Kteily N, Hodson G, Bruneau E: **They see us as less than human: Metadehumanization predicts intergroup conflict via reciprocal dehumanization.** *J Pers Soc Psychol* 2016, **110**:343.
38. Spatola N, Urbanska K: **God-like robots: the semantic overlap between representation of divine and artificial entities**. *Ai Soc* 2020, **35**:329–341.
39. Gazzola V, Rizzolatti G, Wicker B, Keysers C: **The anthropomorphic brain: the mirror neuron system responds to human and robotic actions**. *Neuroimage* 2007, **35**:1674–1684.
40. Broadbent E, Kumar V, Li X, Sollers 3rd J, Stafford RQ, MacDonald BA, Wegner DM: **Robots with display screens: a robot with a more humanlike face display is perceived to have more mind and a better personality**. *PloS One* 2013, **8**:e72589.
41. Babel F, Hock P, Kraus J, Baumann M: **Human-Robot Conflict Resolution at an Elevator-The Effect of Robot Type, Request Politeness and Modality**. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. . IEEE; 2022:693–697.
42. Kupferberg A, Huber M, Helfer B, Lenz C, Knoll A, Glasauer S: **Moving just like you: motor interference depends on similar motility of agent and observer**. *PloS One* 2012, **7**:e39637.
43. Ciardo F, De Tommaso D, Wykowska A: **Joint action with artificial agents: Human-likeness in behaviour and morphology affects sensorimotor signaling and social inclusion**. *Comput Hum Behav* 2022, **132**:107237.
44. Sacino A, Cocchella F, De Vita G, Bracco F, Rea F, Sciutti A, Andrighetto L: **Human-or object-like? Cognitive anthropomorphism of humanoid robots**. *Plos One* 2022, **17**:e0270787.
45. Chaminade T, Okka MM: **Comparing the effect of humanoid and human face for the spatial orientation of attention**. *Front Neurorobotics* 2013, **7**:12.
46. Kompatsiari K, Bossi F, Wykowska A: **Eye contact during joint attention with a humanoid robot modulates oscillatory brain activity**. *Soc Cogn Affect Neurosci* 2021, **16**:383–392.
47. McCabe K, Houser D, Ryan L, Smith V, Trouard T: **A functional imaging study of cooperation in two-person reciprocal exchange**. *Proc Natl Acad Sci* 2001, **98**:11832–11835.
48. Krach S, Hegel F, Wrede B, Sagerer G, Binkofski F, Kircher T: **Can machines think? Interaction and perspective taking with robots investigated via fMRI**. *PloS One* 2008, **3**:e2597.
49. Assaf M, Kahn I, Pearlson GD, Johnson MR, Yeshurun Y, Calhoun VD, Hendler T: **Brain activity dissociates mentalization from motivation during an interpersonal competitive game**. *Brain Imaging Behav* 2009, **3**:24–37.
50. Coricelli G, Nagel R: **Neural correlates of depth of strategic reasoning in medial prefrontal cortex**. *Proc Natl Acad Sci* 2009, **106**:9163–9168.
51. Kätsyri J, Hari R, Ravaja N, Nummenmaa L: **The opponent matters: elevated fMRI reward responses to winning against a human versus a computer opponent during interactive video game playing**. *Cereb Cortex* 2013, **23**:2829–2839.
52. Anders S, Heussen Y, Sprenger A, Haynes J-D, Ethofer T: **Social gating of sensory information during ongoing communication**. *NeuroImage* 2015, **104**:189–198.
53. Schindler S, Kruse O, Stark R, Kissler J: **Attributed social context and emotional content recruit frontal and limbic brain regions during virtual feedback processing**. *Cogn Affect Behav Neurosci* 2019, **19**:239–252.
54. McDonald KR, Pearson JM, Huettel SA: **Dorsolateral and dorsomedial prefrontal cortex track distinct properties of dynamic social behavior**. *Soc Cogn Affect*

Neurosci 2020, **15**:383–393.
55. Koban L, Gianaros PJ, Kober H, Wager TD: **The self in context: brain systems linking mental and physical health**. *Nat Rev Neurosci* 2021, **22**:309–322.
56. Fareri DS, Hackett K, Tepfer LJ, Kelly V, Henninger N, Reeck C, Giovannetti T, Smith DV: **Age-related differences in ventral striatal and default mode network function during reciprocated trust**. *NeuroImage* 2022, **256**:119267.
57. Lee VK, Harris LT: **Sticking with the nice guy: Trait warmth information impairs learning and modulates person perception brain network activity**. *Cogn Affect Behav Neurosci* 2014, **14**:1420–1437.
58. Vaitonyte G, Valiene E, Senvaityte D: **Signs of Culture in Computer Games: Assumption for Education**. In *Proceedings TEEM 2022: Tenth International Conference on Technological Ecosystems for Enhancing Multiculturality: Salamanca, Spain, October 19–21, 2022*. . Springer; 2023:738–746.
59. Lasana T. Harris: **The Neuroscience of Human and Artificial Intelligence Presence**. *Annual Review of Psychology* 2023, **75**.
60. Spatola N, Marchesi S, Wykowska A: **Intentional and Phenomenal attributions in the light of the influence of personality traits, and Attitudes towards robots on pro-social behaviour in human-robot interactio**. 2021, doi:10.31234/osf.io/qaw3t.
61. Kuchenbrandt D, Eyssel F, Bobinger S, Neufeld M: **When a robot's group membership matters**. *Int J Soc Robot* 2013, **5**:409–417.
62. Abubshait A, Pérez-Osorio J, De Tommaso D, Wykowska A: **Collaboratively framed interactions increase the adoption of intentional stance towards robots**. In *2021 30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*. . IEEE; 2021:886–891.
63. Komatsu T, Malle BF, Scheutz M: **Blaming the Reluctant Robot: Parallel Blame Judgments for Robots in Moral Dilemmas across U.S. and Japan**. In *Proceedings of the 2021 ACM/IEEE International Conference on Human-Robot Interaction*. . ACM; 2021:63–72.
64. Malle BF, Magar ST, Scheutz M: **AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma**. In *Robotics and well-being*. . Springer; 2019:111–133.
65. Bering J, Johnson D: **" O Lord… You Perceive my Thoughts from Afar": Recursiveness and the Evolution of Supernatural Agency**. *J Cogn Cult* 2005, **5**:118–142.
66. Haley KJ, Fessler DM: **Nobody's watching?: Subtle cues affect generosity in an anonymous economic game**. *Evol Hum Behav* 2005, **26**:245–256.
67. Shariff AF, Norenzayan A: **God is watching you: Priming God concepts increases prosocial behavior in an anonymous economic game**. *Psychol Sci* 2007, **18**:803–809.
68. Epley N, Waytz A, Akalis S, Cacioppo JT: **When we need a human: Motivational determinants of anthropomorphism**. *Soc Cogn* 2008, **26**:143–155.
69. Keijsers M, Kazmi H, Eyssel F, Bartneck C: **Teaching robots a lesson: determinants of robot punishment**. *Int J Soc Robot* 2021, **13**:41–54.
70. Wiese E, Wykowska A, Zwickel J, Müller HJ: **I see what you mean**. *PLOS ONE* 2012, **7**.
71. Wykowska A, Wiese E, Prosser A, Müller HJ: **Beliefs about the minds of others influence how we process sensory information**. *PloS One* 2014, **9**:e94339.
72. Özdem C, Wiese E, Wykowska A, Müller H, Brass M, Van Overwalle F: **Believing androids–fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents**. *Soc Neurosci* 2017, **12**:582–593.

73. Cheetham M, Pedroni A, Antley A, Slater M, Jäncke L: **Virtual milgram: empathic concern or personal distress? Evidence from functional MRI and dispositional measures**. *Front Hum Neurosci* 2009,
74. Cheetham M, Suter P, Jancke L: **Perceptual discrimination difficulty and familiarity in the uncanny valley: more like a "Happy Valley."** *Front Psychol* 2014, **5**:1219.
75. Fiske ST, Cuddy AJC, Glick P, Xu J: **A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition.** *J Pers Soc Psychol* 2002, **82**:878–902.
76. Fiske ST, Cuddy AJC, Glick P: **Universal dimensions of social cognition: warmth and competence**. *Trends Cogn Sci* 2007, **11**:77–83.
77. Malle BF, Scheutz M, Arnold T, Voiklis J, Cusimano C: **Sacrifice One For the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents**. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction*. . ACM; 2015:117–124.
78. Torta E, Oberzaucher J, Werner F, Cuijpers RH, Juola JF: **Attitudes towards socially assistive robots in intelligent homes: results from laboratory studies and field trials**. *J Hum-Robot Interact* 2013, **1**:76–99.
79. Sandoval EB, Brandstetter J, Obaid M, Bartneck C: **Reciprocity in human-robot interaction: a quantitative approach through the prisoner's dilemma and the ultimatum game**. *Int J Soc Robot* 2016, **8**:303–317.
80. Maggioni MA, Rossignoli D: **If it Looks like a Human and Speaks like a Human... Dialogue and cooperation in human-robot interactions**. *ArXiv Prepr ArXiv210411652* 2021,
81. Whiting T, Gautam A, Tye J, Simmons M, Henstrom J, Oudah M, Crandall JW: **Confronting barriers to human-robot cooperation: balancing efficiency and risk in machine behavior**. *Iscience* 2021, **24**:101963.
82. Karpus J, Krüger A, Verba JT, Bahrami B, Deroy O: **Algorithm exploitation: Humans are keen to exploit benevolent AI**. *iScience* 2021, **24**:102679.
83. Jack AI, Robbins P: **The phenomenal stance revisited**. *Rev Philos Psychol* 2012, **3**:383–403.
84. Seo SH, Geiskkovitch D, Nakane M, King C, Young JE: **Poor thing! Would you feel sorry for a simulated robot? A comparison of empathy toward a physical and a simulated robot**. In *2015 10th ACM/IEEE international conference on human-robot interaction (HRI)*. . IEEE; 2015:125–132.
85. Suzuki Y, Galli L, Ikeda A, Itakura S, Kitazaki M: **Measuring empathy for human and robot hand pain using electroencephalography**. *Sci Rep* 2015, **5**:1–9.
86. Chang W, Wang H, Yan G, Lu Z, Liu C, Hua C: **EEG based functional connectivity analysis of human pain empathy towards humans and robots**. *Neuropsychologia* 2021, **151**:107695.
87. Rosenthal-Von Der Pütten AM, Schulte FP, Eimler SC, Sobieraj S, Hoffmann L, Maderwald S, Brand M, Krämer NC: **Investigations on empathy towards humans and robots using fMRI**. *Comput Hum Behav* 2014, **33**:201–212.
88. Jackson PL, Rainville P, Decety J: **To what extent do we share the pain of others? Insight from the neural bases of pain empathy**. *Pain* 2006, **125**:5–9.
89. Morrison I, Peelen MV, Downing PE: **The sight of others' pain modulates motor processing in human cingulate cortex**. *Cereb Cortex* 2007, **17**:2214–2222.
90. Saarela MV, Hlushchuk Y, Williams AC de C, Schürmann M, Kalso E, Hari R: **The compassionate brain: humans detect intensity of pain from another's face**. *Cereb Cortex* 2007, **17**:230–237.