

## **Beyond human-likeness: Socialness is more influential when attributing mental states to robots**

Jastrzab Binney, Laura; Chaudhury, Bishakha; Ashley, Sarah; Koldewyn, Kami; Cross, Emily

### **iScience**

DOI:  
[10.1016/j.isci.2024.110070](https://doi.org/10.1016/j.isci.2024.110070)

Published: 21/06/2024

Peer reviewed version

[Cyswllt i'r cyhoeddiad / Link to publication](#)

*Dyfyniad o'r fersiwn a gyhoeddwyd / Citation for published version (APA):*  
Jastrzab Binney, L., Chaudhury, B., Ashley, S., Koldewyn, K., & Cross, E. (2024). Beyond human-likeness: Socialness is more influential when attributing mental states to robots. *iScience*, 27(6), Article 110070. <https://doi.org/10.1016/j.isci.2024.110070>

#### **Hawliau Cyffredinol / General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

#### **Take down policy**

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

# Journal Pre-proof



Beyond human-likeness: Socialness is more influential when attributing mental states to robots

Laura E. Jastrzab, Bishakha Chaudhury, Sarah A. Ashley, Kami Koldewyn, Emily S. Cross

PII: S2589-0042(24)01295-1

DOI: <https://doi.org/10.1016/j.isci.2024.110070>

Reference: ISCI 110070

To appear in: *ISCIENCE*

Received Date: 11 September 2023

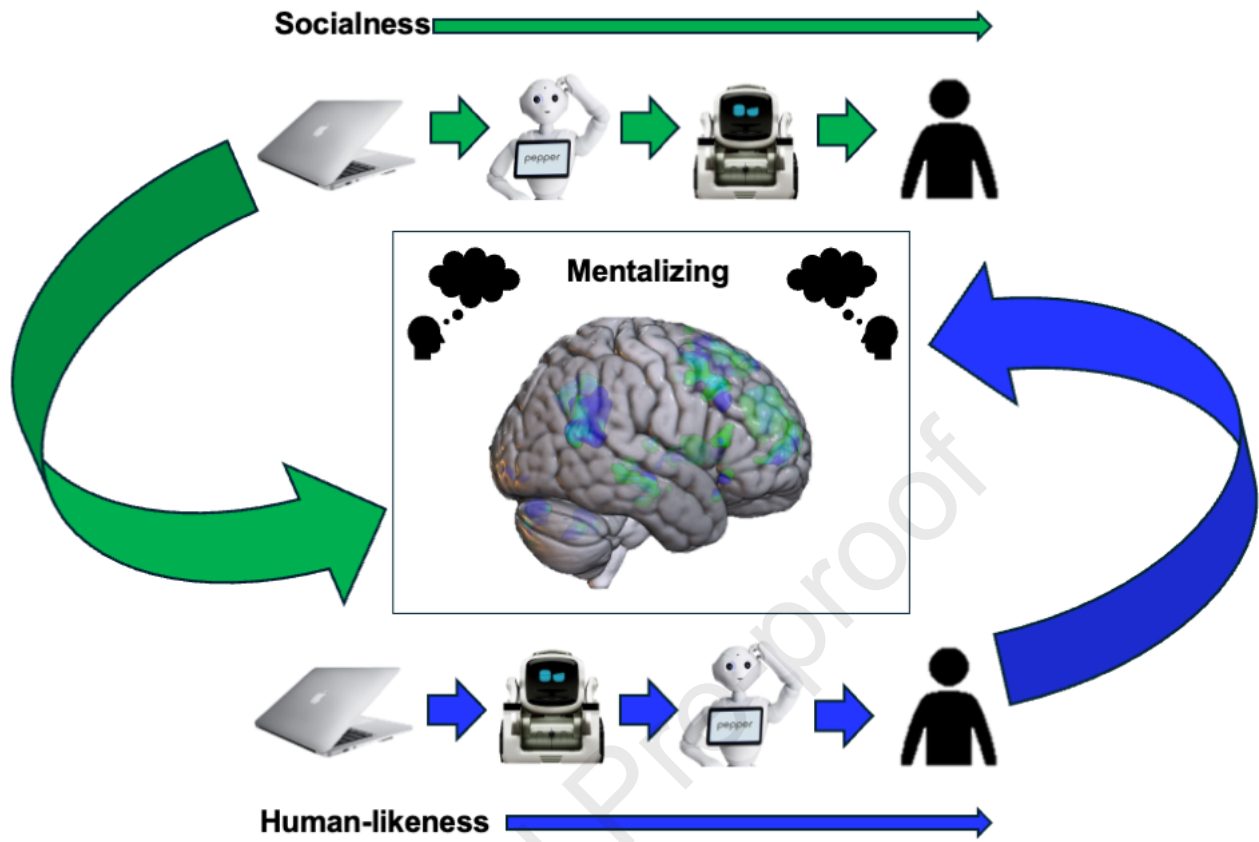
Revised Date: 8 March 2024

Accepted Date: 17 May 2024

Please cite this article as: Jastrzab, L.E., Chaudhury, B., Ashley, S.A., Koldewyn, K., Cross, E.S., Beyond human-likeness: Socialness is more influential when attributing mental states to robots, *ISCIENCE* (2024), doi: <https://doi.org/10.1016/j.isci.2024.110070>.

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2024 The Author(s). Published by Elsevier Inc.



1  
2  
3  
4  
5  
6  
7  
8  
9  
10  
11  
12  
13  
14  
15  
16  
17  
18  
19  
20  
21  
22  
23  
24  
25  
26  
27  
28  
29  
30

**Beyond human-likeness: Socialness is more influential when attributing mental states to robots**

Laura E. Jastrzab<sup>1,2</sup>, Bishakha Chaudhury<sup>2</sup>, Sarah A. Ashley<sup>1,3</sup>, Kami Koldewyn<sup>1</sup>, Emily S. Cross<sup>\*2,4</sup>

<sup>1</sup> *Institute for Cognitive Neuroscience, School of Human and Behavioural Science, Bangor University, Wales, UK*

<sup>2</sup> *Institute for Neuroscience and Psychology, School of Psychology, University of Glasgow, UK*

<sup>3</sup> *Division of Psychiatry, Institute of Mental Health, University College London, UK*

<sup>4</sup> *Chair for Social Brain Sciences, Department of Humanities, Social and Political Sciences, ETHZ, Zürich, Switzerland*

*\* Lead contact:*

Emily S. Cross: [ecross@ethz.ch](mailto:ecross@ethz.ch)

Key words: Social robotics, Second-person neuroscience, Social Cognition, & Mentalizing

**31 Summary**

32 We sought to replicate and expand previous work showing that the more human-like  
33 a robot appears, the more willing people are to attribute mind-like capabilities and  
34 socially engage with it. Forty-two participants played games against a human, a  
35 humanoid robot, a mechanoid robot, and a computer algorithm while undergoing  
36 functional neuroimaging. We confirmed that the more human-like the agent, the  
37 more participants attributed a mind to them. However, exploratory analyses revealed  
38 that the perceived *socialness* of an agent appeared to be as, if not more, important  
39 for mind attribution. Our findings suggest top-down knowledge cues may be equally  
40 or possibly more influential than bottom-up stimulus cues when exploring mind  
41 attribution in non-human agents. While further work is now required to test this  
42 hypothesis directly, these preliminary findings hold important implications for robotic  
43 design and to understand and test the flexibility of human social cognition when  
44 people engage with artificial agents.

## 45 **Introduction**

46 Robots have sparked curiosity and been romanticised in popular culture since von  
47 Kempelen's "Chess Turk" was introduced in 1769. In the mid-20<sup>th</sup> century, Alan  
48 Turing formalised the philosophical debate as to whether "machines think",<sup>1</sup> a  
49 question that continues to captivate many philosophical and science fiction writers.  
50 With the present study, however, we ask what might be thought of as the *opposite*  
51 question: namely, regardless of whether robots think, do *we humans* perceive robots  
52 as having minds of their own? If so, do we do so primarily based on how human-like  
53 the robot looks, or does its perceived socialness also matter?

54 Robots are already commonplace in assembly lines, factories, and dangerous  
55 jobs such as pipeline and fuel tank inspections, as well as underwater and space  
56 exploration.<sup>2,3</sup> As the deployment of robots in these contexts grows, so does their  
57 introduction to social and leisure domains, aiding people with, for example, surgeries  
58 in healthcare, serving customers in restaurants, learning in schools, and supporting  
59 adults who need help with daily living skills (for example, <sup>4-8</sup>). Robots' roles in our  
60 day-to-day lives so far, however, are typically "single-use" (e.g., robot vacuum  
61 cleaners or a robot check-in assistant at a hotel), and the ability of even the most  
62 sophisticated social robots to engage us socially is still far removed from depictions  
63 in science fiction novels and films.<sup>9,10</sup> Rapid advances in hardware and artificial  
64 intelligence are expected over the coming decades, making this a crucial time to  
65 examine human engagement with robots. This is particularly true in the social  
66 domain if we are to develop machines that can indeed engage and collaborate with  
67 humans in complex social contexts.

68 As adults, humans typically and intuitively think of other humans as having a  
69 mind, thoughts, and intentions that are different from their own, a skill known as

70 mentalizing.<sup>11,12</sup> Mentalizing is important for social interactions, allowing us to read  
71 and react to others' unspoken mental and emotional states, and their intended  
72 actions.<sup>11</sup> Neuroimaging studies have used implicit (e.g., economic games) and  
73 explicit (e.g., mind-in-the-eyes) tasks to probe human brain activity associated with  
74 mentalizing (for a review, see <sup>13</sup>). This work has identified the so-called mentalizing  
75 network, a group of brain regions thought to support thinking about others' minds.  
76 The core regions reliably included as part of the mentalizing network include bilateral  
77 temporal-parietal junction (TPJ), medial prefrontal cortex (mPFC), and Precuneus  
78 (PreC) but engagement of additional brain regions, including posterior superior  
79 temporal sulcus (pSTS), temporal poles, and posterior cingulate cortex (PCC), have  
80 also been implicated.<sup>13-18</sup> Briefly, it is thought that the mPFC is at the top of the  
81 mentalizing hierarchy and the primary source of top-down signals as well as the hub  
82 of self-referential processing. The TPJ & pSTS are intermediary in the hierarchy, with  
83 the TPJ contributing to metacognitive representations and the pSTS contributing  
84 primarily to the processing of social agents and actions (see <sup>19</sup> for a discussion). The  
85 role of the precuneus in the mentalizing system is less clear, given that other  
86 cognitive functions have been attributed to it; thus, its functional role is often  
87 described as outside of the mentalizing realm (e.g. spatial navigation). Within the  
88 mentalizing literature, however, the precuneus' is described as potentially belonging  
89 at the top of the mentalising hierarchy (along with mPFC) as a "staging post between  
90 implicit and explicit mentalizing".<sup>19</sup> For the purposes of the current study, we consider  
91 these regions collectively, focusing on engagement of the broader mentalizing  
92 system as a whole.

93         The mentalizing network is readily engaged during interactions with other  
94 humans, especially when trying to predict their future actions. Very few neuroimaging

95 studies, however, have directly addressed the extent to which mentalizing brain  
96 regions, which have ostensibly evolved to interpret other people's actions and  
97 intentions, also process non-human social partners such as robots. Understanding  
98 whether humans mentalize about robots is important for at least two reasons. First,  
99 the more we attribute a mind to robots, the more likely we are to interact with and  
100 engage with them socially.<sup>20–22</sup> Second, examining mentalizing in response to robot  
101 social partners tests the flexibility of our social cognitive system by assessing the  
102 extent to which a system that evolved to support interactions with fellow humans can  
103 be engaged during interactions with non-human agents (in this case, robots).<sup>23</sup> Prior  
104 neuroimaging studies studying the extent to which humans mentalize about robots  
105 have used empathy tasks,<sup>24,25</sup> spatial cueing tasks,<sup>22,26</sup> and economic games.<sup>27–29</sup>  
106 Several of these studies demonstrate that human—robot interactions (HRI) activate  
107 the mentalizing network, but to a lesser degree than human—human interactions  
108 (HHI).<sup>27,28,30</sup>

109         One influential theory that might help to explain the pattern of activity reported  
110 so far is the 'like-me' hypothesis,<sup>31</sup> which posits that the more human-like a non-  
111 human agent appears, the more readily social brain networks are engaged. Indeed,  
112 behavioural data generally support this idea. For example, the more human-like a  
113 robot appears, the more a human user will expect that robot to behave like a  
114 human.<sup>32</sup> Furthermore, a robot's appearance influences our assumptions about its  
115 behavioural capabilities<sup>33–35</sup> and the extent to which we attribute intentionality or a  
116 mind to them.<sup>20–22,36,37</sup> Likewise, the degree to which we anthropomorphize robots  
117 (or attribute human-like qualities to them) has also been found to depend upon a  
118 robot's human-like appearance and behaviour.<sup>38–41</sup> Given the behavioural evidence,  
119 it is perhaps not surprising that similar results are found when examining socio-



120 cognitive brain systems. For example, Krach and colleagues<sup>28</sup> reported that the  
121 increasing human-likeness of game partners' physical features was associated with  
122 increasing engagement of mentalizing network regions during an implicit mentalizing  
123 task (in this case, an iterative prisoner's dilemma game). Together, behavioural and  
124 brain imaging findings support the idea that the human-likeness of an interactive  
125 partner's appearance plays a key role in engaging socio-cognitive processes like  
126 mentalizing. However, emerging evidence raises the possibility that human-likeness  
127 alone may not fully explain which robots are seen as more desirable social partners  
128 and, thus, which robot features might be most effective at eliciting the strongest  
129 human-like social-cognition processes.<sup>42,43</sup> The influence of a robot's *social* features,  
130 per se, on human perception and engagement is an emerging area of research that  
131 will benefit from expertise from the Human Robot Interaction (HRI), social robotics,  
132 and cognitive neuroscience communities.

133 In the current study, we sought to replicate prior findings that the mentalizing  
134 network increases in responsiveness as the appearance of robots increases in  
135 human-likeness. In additional exploratory analyses, we sought to explore the extent  
136 to which a partner's perceived *socialness* (independent from human-like physical  
137 features) might also contribute to this process. To do so, we used an established  
138 implicit mentalizing task where participants play rock-paper-scissors (RPS)<sup>43</sup> against  
139 a human and several artificial agents. We followed an experimental design like that  
140 reported by Chaminade and colleagues.<sup>27</sup> An important feature of our RPS design  
141 was that we examined how an individual's beliefs regarding the nature of the  
142 interacting agents are influenced by the human-likeness and socialness of each  
143 agent, while tightly controlling all other aspects (i.e. visual, sensorimotor, etc) of the  
144 gameplay interaction. Specifically, participants viewed the same visual stimulus

145 during game play when playing against all 4 game partners. It was only the videos  
146 before and after game play that reminded participants against whom they were  
147 playing. This design, therefore, necessitates reliance on top-down knowledge cues  
148 regarding the other player to drive neural activation during game play. The RPS  
149 game itself is familiar across cultures and age groups, and if it is unfamiliar, it is easy  
150 to learn. Also, like Chaminade and colleagues,<sup>27</sup> we used videos of game partners to  
151 increase the sense of live interactions during game play. We controlled wins and  
152 losses across all game partners, and explicitly told participants that the robot  
153 competitors had been endowed with artificial intelligence and would play  
154 strategically. Similar to Krach and colleagues,<sup>28</sup> we included two robotic partners that  
155 differed in their human-like appearance. One robot appeared humanoid, with clear  
156 human-like features including a body, torso, arms, hands, fingers, head and eyes.  
157 The other was a mechanoid robot, which had expressive eyes but no other human-  
158 like physical features (refer to Figure 1). Importantly, both the humanoid and  
159 mechanoid robots in our study are designed to engage people with socially  
160 interactive behaviours.

161 From prior data, we expected that both robots would engage the mentalizing  
162 network, though to a lesser extent than the human game-partner. Indeed, we  
163 preregistered a prediction that the magnitude of response of core brain regions  
164 within the mentalizing network (specifically TPJ, mPFC, and Precuneus) would  
165 linearly increase as game partners increased in human-like appearance. We further  
166 explored the extent to which participants found each robotic game partner fun,  
167 sympathetic, competitive, successful, strategic, intelligent, and competitive. Here we  
168 hypothesized, again based on previous findings<sup>27,28</sup> that these factors would  
169 increase with increasing human-likeness. Finally, in an exploratory analysis, to

170 address questions related to participants' perceptions of the socialness of the  
171 different game partners, we reversed the order of the robots (by changing the rank  
172 order) in our linear contrast models, allowing us to test the extent to which this  
173 "perceived socialness" might explain differences in the engagement of the  
174 mentalizing network across game partners *better* than simply the agents' physical  
175 appearance.

Journal Pre-proof

176 **Results**177 **Neuroimaging Results**178 **Socialness and human-likeness influence mentalizing but socialness is more**  
179 **robust**180 *Pre-registered*

181 Repeated-measures ANOVAs with game partner as a within-subjects factor was  
182 significant in several key mentalizing ROIs during game play (bilateral TPJ and left  
183 middle frontal gyrus (lmFG)), as well as bilateral pSTS. All pairwise comparisons in  
184 this section were corrected for multiple comparisons (Bonferroni). Follow-up paired  
185 sample t-tests in bilateral pSTS and ITPJ revealed that this was largely driven by  
186 higher activity in response to the human compared to all other conditions, suggesting  
187 that these regions are more reliably engaged by human than artificial stimuli (see  
188 Supplementary Table 5). Right TPJ was an exception, in that, while the human  
189 significantly differed from both robots, no significant difference between the human  
190 and computer was found. No other significant comparisons during gameplay and  
191 within these ROIs remained after correcting for multiple comparisons.

192 Results from the pSTS revealed significant differences between game players while  
193 playing the game (rpSTS:  $F(3, 123) = 12.39, p < 0.001, \eta^2 = 0.23$ ; lpSTS:  $F(3, 123) =$   
194  $6.96, p < 0.001, \eta^2 = 0.15$ ), which was unexpected as there were no visual  
195 differences during game play across the 4 conditions.

196 Contrary to our expectations, mentalizing regions were not activated above baseline  
197 during the RPS games. Average activity across the group was close to zero or,  
198 indeed, slightly negative across nearly all conditions (refer to Figures 1, S2 & Table  
199 S5).

200 *Exploratory*

201 Additionally, the pSTS revealed strong significant differences across game partners  
202 while participants watched the introductory video (video 1) of each game partner  
203 before playing commencing each game series (rpSTS:  $F(3, 123) = 29.40, p < 0.001,$   
204  $np = 0.42;$  lpSTS:  $F(3, 123) = 13.26, p < 0.001, np = 0.24$ ). While none of the other  
205 ROIs revealed significant pairwise differences between either robot and the  
206 computer, there was a significant difference between MR and CP in rpSTS (and  
207 approached significance in lpSTS) during the video preceding gameplay (rpSTS:  $p <$   
208  $0.001, d = -0.73;$  lpSTS:  $p = 0.056, d = -0.32;$  See Supplementary Table S5).

209

### 210 **Linear effect of human-likeness in mentalizing ROIs during gameplay**

#### 211 *Pre-Registered*

212 All mentalizing ROIs which revealed a significant within subject effect of partner  
213 (Bilateral TPJ, ImFG, and bilateral pSTS) also revealed a significant linear within-  
214 subjects contrast effect of human-likeness (HP > HR > MR > CP), as predicted (refer  
215 to Table S5).

#### 216 *Exploratory*

217 We explored whether changing the rank order of the robots (in the 4-element  
218 hierarchy) in the within-subject contrasts according to socialness ratings further  
219 bolstered the linear effect (HP > MR > HR > CP, refer to Table S5). Results from  
220 behavioural ratings suggested that socialness (as assessed by perceived fun,  
221 competitiveness, and sympathy, see below) models were improved by reversing the  
222 order of the robots. Indeed, across ROIs, the mechanoid robot evoked numerically  
223 higher, though often not significantly so, responses than the humanoid robot. Despite  
224 the lack of statistically significant differences between the robots in pairwise

225 comparisons, the linear effect of 'socialness' resulted in a larger effect size than the  
226 'humanness' model, suggesting socialness may be even more important than  
227 humanness in mind attribution toward robots, as measured by engagement of brain  
228 regions associated with mentalizing.

### 229 **The mechanoid is more similar to the human than the humanoid or computer**

#### 230 *Pre-Registered*

231 No FWE ( $p < .05$ ) or uncorrected ( $p < .001$ ) clusters survived simple whole brain  
232 contrasts between the humanoid or mechanoid and the computer (refer to Table S4).  
233 There were no significant clusters during the [Humanoid (HR) > Mechanoid (MR)] but  
234 the inverse contrast revealed a significant cluster ( $k = 313$ ) in nucleus accumbens  
235 (MNI: -4 10 -10). The [Human Partner (HP) > Computer Partner (CP)] contrast  
236 resulted in significant mentalizing clusters in bilateral TPJ, mFG, mPFC, precuneus,  
237 rpSTS, IFG, nucleus accumbens, and cerebellum.

238 To assess whether regions outside our pre-selected ROIs might be sensitive to  
239 Human-likeness, we tested whether any brain regions showed a pattern of activity  
240 such that Human Partner (HP) > Humanoid Robot (HR) > Mechanoid Robot (MR) >  
241 Computer Partner (CP). This analysis revealed that rTPJ, precuneus, mPFC,  
242 bilateral mFG, and nucleus accumbens all survived the FWE-corrected peak-level  
243 threshold.

#### 244 *Exploratory*

245 When the human was compared to the humanoid and mechanoid robots, several  
246 regions associated with mentalizing were significant at the cluster level after FWE  
247 correction (refer to Figure 2). The [HP > HR] contrast resulted in significant clusters  
248 in bilateral TPJ, precuneus, rmFG, rIFG, rpSTS after FWE corrections. The [HP >

249 MR] contrast yielded significant engagement of rTPJ, precuneus, rpSTS, and  
250 cerebellum after FWE corrections.

251 In line with our socialness questions, we also tested whether any brain regions  
252 showed a pattern of activity if we reversed the order of the robots in our parametric  
253 analysis; i.e., so that the order was now: Human Partner (HP) > Mechanoid Robot  
254 (MR) > Humanoid Robot (HR) > Computer Partner (CP). Results revealed a similar  
255 pattern to both HP>CP and the HP>HR>MR>CP model above but now also included  
256 significant clusters in: bilateral pSTS, supplementary motor area, rIFG,& ITPJ. Refer  
257 to Figure S1 and Table S4.

258

## 259 **Behavioural Results**

### 260 **Manipulation Check**

261 During verbal debriefing with participants, six out of 42 neuroimaging participants  
262 questioned whether the videos were live during our verbal debriefing. Given this, we  
263 re-ran all behavioural and neuroimaging analyses with only the “true believers” (see  
264 OSF project page for details). Doing so did not change the findings in either degree  
265 or direction of significance. Therefore, the analyses are reported with the full sample,  
266 including the non-believers.

### 267 **Debrief Questions: Mechanoid perceived as more social, but not intelligent,** 268 **than the humanoid**

#### 269 *Pre-Registered*

270 All pairwise comparisons in this section were corrected for multiple comparisons  
271 (Bonferroni). Greenhouse-Geisser corrections were made if any rmANOVA was

272 found to violate Mauchley's tests of sphericity (refer to Figures 3, S3, & Table S6 for  
273 details from this section).

274 We found no effect of perceived *success* in winning ( $F(3, 123) = 0.50, p = 0.685, \eta^2$   
275  $= .012$ ) or *strategy* employed ( $F(3, 123) = 0.32, p = 0.811, \eta^2 = .008$ ) against each  
276 game partner, despite stressing to participants that the computer was using a  
277 random algorithm, while the other partners were all trying to win.

278 *Fun* ( $F(3, 123) = 33.90, p < 0.001, \eta^2 = .453$ ), *Competitiveness* ( $F(3, 123) = 17.24,$   
279  $p < 0.001, \eta^2 = .296$ ), *Sympathy* ( $F(2.50, 102.58) = 58.59, p < 0.001, \eta^2 = .588;$   
280 Greenhouse-Geisser corrected) and *Intelligence* ( $F(2.51, 102.91) = 12.16, p < 0.001,$   
281  $\eta^2 = .229;$  Greenhouse-Geisser correction) were all significantly different amongst  
282 the four conditions and followed a significant linear pattern based on human-  
283 likeness.

#### 284 *Exploratory*

285 However, *Fun*, *Competitiveness*, and *Sympathy*, revealed a stronger linear pattern  
286 based on socialness, wherein we changed the rank order of the robots in the 4-  
287 element hierarchy. However, only post-hoc tests on ratings of *Fun* and  
288 *Competitiveness* showed differences between robots, where mean ratings for the  
289 mechanoid robot were higher than for the humanoid robot ( $p=0.006$  &  $p=0.049,$   
290 respectively).

291

#### 292 **Inclusion of Others and Self (IOS): No difference in perception of closeness** 293 **between the robots or a human stranger**

294 IOS scores varied significantly between the 6 agents ( $F(3.70, 148.02) = 122.40, p <$   
295  $0.001, \eta^2 = 0.754$ ). Pairwise comparisons of the computer, human game partner, and



296 close friend significantly differed from all other agents and each other on the IoS,  
297 even after correcting for multiple comparisons (Bonferroni). Pairwise comparisons of  
298 the mechanoid robot, humanoid robot, and human stranger did not significantly differ  
299 from each other (Please see our OSF page for details).

300

## 301 **DISCUSSION**

302 With the present study, we have replicated and extended previous findings,  
303 demonstrating that both human-likeness and perceived 'socialness' shape the extent  
304 to which participants engage mentalizing regions while playing games against  
305 robotic partners. We found that although human-likeness models showed increased  
306 theory-of-mind network engagement (as predicted and pre-registered), the  
307 socialness model was even more robust. While this analysis was exploratory and will  
308 require replication via hypothesis-confirming follow-up work, it is important for two  
309 reasons. First, it suggests that mentalizing processes during interactive exchanges  
310 (in this case, a game) are better predicted by how *social* we find our interaction  
311 partner, rather than being solely based on how human-like they look. This finding  
312 has the potential to update our models of how mentalizing systems can be engaged,  
313 particularly by non-human interactants. Secondly, the extent to which humans will  
314 ascribe mental states to robots is likely to become increasingly relevant as roboticists  
315 develop increasingly sophisticated embodied artificial agents designed to engage  
316 human users on a social level. Successful social interactions with such social robots  
317 will require people to think about how the robot "thinks". A better understanding of  
318 the factors that influence mentalizing towards and about robots should lead to higher  
319 quality and more sustained long-term interactions with robots in social domains (e.g.,  
320 <sup>43</sup>).

321 As with the two previous neuroimaging studies on which we based our current study,  
322 we found increasing activation in mentalizing regions with increasing human-  
323 likeness.<sup>27,28</sup> We also found similar behavioural ratings, showing that while  
324 participants did not perceive strategy and success differently across game partners  
325 (suggesting participants did not feel that they won or lost more against any one  
326 game partners), participants did perceive the game partners differently based on  
327 social factors like perceived intelligence, fun, competitiveness, and sympathy.  
328 However, unlike previous studies, we explored how these social factors might  
329 contribute to mind attribution and found that changing the rank order of robots in the  
330 4-element hierarchy in our linear contrast models to reflect participants' evaluations  
331 of socialness resulted in numerically stronger models than those based on the  
332 human-likeness of physical features alone.

### 333 *Quantifying and exploring human-likeness vs. socialness*

334 While the human and social models were both significant and strong, one possibility  
335 for the numerically stronger social model is that the mechanoid robot was perceived  
336 as more social because it exhibited higher levels of hedonic factors (as rated by fun,  
337 competitiveness, and sympathy) than did the humanoid robot. This finding is  
338 consistent with participant qualitative perceptions and behavioural ratings of this  
339 same robot in recently published work.<sup>44,45</sup> For example, in one scenario from our  
340 study, when the mechanoid robot lost the RPS series, it pouted and slammed its  
341 forklift on the table while moving around in circles in protest. Whereas, when the  
342 humanoid robot lost, it responded similarly to the human in a more measured  
343 manner, by lowering its arms and shaking its head and/or looking down in defeat.  
344 While these differences in personality and behaviour were not objectively measured  
345 in our study, others report that manipulating social features of robots such as

346 personality,<sup>46,47</sup> emotional arousal,<sup>48</sup> and other hedonic features such as enjoyment  
347 and sociability<sup>49</sup> can increase user engagement, acceptance, and/or satisfaction.

348 The neuroimaging evidence from this study supports both human-likeness and  
349 socialness models when attributing mental states. Bilateral TPJ, bilateral pSTS, and  
350 ImFG showed significant increases with human-likeness and a numerically stronger  
351 linear increase with socialness. While we expected the whole mentalizing network  
352 and pSTS to show a similar response pattern, the exceptions were in mPFC,  
353 precuneus, and rmFG.

354 We were unable to clearly assess the role played by our mPFC, Precuneus, and  
355 rmFG ROIs in this study, as we found no significant differences to emerge between  
356 the agents during game play. However, a wealth of research has proposed that  
357 these regions are central to mentalising and animacy (e.g.<sup>13,16,18</sup>). As our localisers  
358 did not reliably elicit mPFC or rmFG response in this participant cohort, we created  
359 ROI from coordinates in the original localiser paper.<sup>50</sup> It is possible that our “generic”  
360 ROIs failed to capture individuals’ peak mentalizing voxels across these regions.  
361 However, mPFC and rmFG activation clearly emerges in many of our group whole  
362 brain contrasts. Precuneus clusters in our localizer and main experimental task were  
363 large and the peak cluster from the localiser was more inferior and lateral than the  
364 peak clusters in the main experimental task. Last, it is also possible that our  
365 localisers produced coordinates for offline social cognition or mentalizing and not for  
366 online social cognition.<sup>51</sup> Thus, mPFC, rmFG, and Precuneus may play a role in  
367 mentalizing in our study but were perhaps not well captured by our choice of ToM  
368 localiser and, thus, the resulting ROI coordinates. Future studies may consider  
369 creating simple spheres from t-value peaks reported from our main task  
370 experimental data or from peaks reported in other similar papers.

371 We also explored the response profile of a region in the pSTS that is sensitive to  
372 interactive information in observed dyadic social interactions.<sup>52</sup> This region is nearby,  
373 but distinct from the TPJ, and might plausibly discriminate between game partners.  
374 Response in the pSTS discriminated between game partners both during game play  
375 and during the video preceding each game series. This was somewhat surprising as  
376 the pSTS is largely responsive to the perceptual features of interactions, particularly  
377 biological motion.<sup>53,54</sup> In our design, there were no social perceptual features to  
378 process during game play as players observed the same visual stimuli during game  
379 play across all four conditions. This suggests that perhaps top-down knowledge cues  
380 may be more influential in this region than previously thought. We further explored  
381 this data by testing our linear human-likeness and socialness models on the pSTS  
382 data from video 1 and gameplay. Both models were significant, but in this case the  
383 social model was numerically stronger during both gameplay (rpSTS only) and video  
384 1 (bilateral pSTS). The pSTS has been implicated as a part of the social cognition  
385 and mentalising networks and has previously been shown to integrate both  
386 perceptual and social features.<sup>52,55-57</sup> The pSTS also responds strongly to social  
387 interactions between non-human agents such as moving shapes and dots of light  
388 that mimic social scenarios (e.g. <sup>55,57,58</sup>), and does so even more strongly when  
389 participants are led to believe an object is animate versus inanimate.<sup>37,59</sup> One  
390 possibility is that because participants were engaging in a real-time interaction in our  
391 study, the pSTS was more strongly driven by the social features of game partners  
392 rather than their visual features. When motion and visual cues to humanness conflict  
393 or are not reliably aligned with more top-down attributions of socialness, the more  
394 superior regions in the pSTS may prioritise top-down knowledge cues to humanness  
395 in social interactions.

396 While our neuroimaging and behavioural results indicate a linear effect of human-  
397 likeness and socialness across conditions, pairwise comparisons from our ROIs also  
398 show that the human partner is perceived significantly differently from all others  
399 game partners. While this result is perhaps unsurprising, it suggests that a uniquely  
400 human factor still differentiates people from animate non-human entities, even when  
401 they are quite 'human-like' in appearance or behaviour. This result has been  
402 reported previously,<sup>37,43</sup> and is consistent with the idea that the mentalizing system  
403 may be best tuned to human actors and human social cues. It is possible with  
404 advancing technology and design that the line between robots and humans may blur,  
405 and mentalising regions will become increasingly recruited.

406 One surprise in our results is that game play did not drive responses in mentalizing  
407 regions above baseline. Our expectation, based on prior research,<sup>27,60</sup> was that this  
408 task would indeed drive engagement of the mentalizing network, at least for the  
409 human partner, above baseline. Previous studies<sup>27,28</sup> found negative activation to the  
410 computer condition and to the non-android robots in mentalizing ROIs, but above-  
411 baseline response to the human partner. One possibility here is that our task was  
412 particularly demanding, requiring not only mentalizing but also analysing and  
413 remembering strategies for each opponent. It is possible that the negative responses  
414 seen in our results are a result of most mentalizing regions being part of, or close to,  
415 the default mode network, which tends to deactivate during difficult or demanding  
416 tasks.<sup>61</sup> Additionally, the DMN is also thought to reflect involvement in perceptually  
417 decoupled thought processes.<sup>62</sup> More specifically, our use of a passive rest  
418 condition as a baseline could have obscured important changes in activation in  
419 response to the task. For example, during minimal baseline tasks (such as passive  
420 rest), mind-wandering and other internally generated thoughts (as opposed to those

421 evoked by external stimuli) are likely to occur, and this could comprise similar social  
422 cognitive processes at equal or greater magnitude to those required by the more  
423 focused task-related processing.<sup>61,63</sup> If social processes (e.g. mentalising) are higher  
424 at rest than in the task, then we should see what looks like deactivation when  
425 comparing the experimental conditions to the passive baseline. Future studies might  
426 consider using an active, rather than passive, baseline<sup>64</sup> for teasing out the  
427 difference in social responses to different partners and computing response  
428 differences across those experimental conditions.

429           It is also possible that the activation in the mentalising network was  
430 attenuated because participants were not actively viewing their game partners during  
431 game play, and therefore were not receiving a constant stream of visual and social  
432 feedback in real-time as they would have in 'real life'. Instead, perhaps they were  
433 relying on memory or impressions of their game partners when playing. Future  
434 studies might more robustly activate ToM regions during the game with real-time  
435 feedback and/or actual live gameplay. Overall, however, the results are consistent  
436 with our pre-registered hypothesis as higher activation levels (or less deactivation)  
437 for humans emerged as compared to robots and for robots as compared to the  
438 computer condition.

439 As with previous studies, and unbeknownst to the participants, we controlled wins  
440 and losses amongst game partners so our findings could not be explained by  
441 winning or losing more to any one partner. Participants' ratings of success and  
442 strategy against each of the 4 game partners did not significantly differ, suggesting  
443 that they accurately perceived their own performance, including that their strategy  
444 did not work any more efficiently for one partner than another, like previous  
445 findings.<sup>27</sup> Therefore, it is unlikely that our findings are due to perceived differences

446 in difficulty in playing each partner. Employing a strategic approach to the game  
447 likely relates to thinking about the mind of the other player, and thus to activity in the  
448 mentalizing network. As a result, participants in this study may have reduced their  
449 mentalizing about game partners as they found that their strategies were not  
450 working. Future studies might look at manipulating wins and losses or alter initial  
451 briefing instructions to create different impressions of each game partner's fun and  
452 competitiveness to explore the extent to which socialness can be manipulated to  
453 influence mind attribution toward robots.

#### 454 *Theoretical implications*

455 Our results support growing evidence emerging from the intersection of social  
456 robotics and social neuroscience that multiple routes exist to non-human agents  
457 being perceived as "like-me",<sup>37,43</sup> including not only a human-like appearance or  
458 motion profile, but also being perceived as 'social' based on behaviours or  
459 background knowledge about a robot. Significant R&D investment continues to fuel  
460 the development of socially interactive robots with whom human users can intuitively  
461 and effectively collaborate, which often attempt to capture as much human-likeness  
462 as possible while also avoiding the uncanny valley.<sup>65-70</sup> However, the extent to which  
463 an agent is perceived as "like-me" extends beyond physical form, capabilities, and  
464 movement, and growing evidence supports that prior knowledge about and the  
465 perceived socialness of a robot may more strongly influence their reception (and  
466 people's ability to collaborate or cooperate with them in an intuitive manner) in social  
467 settings.<sup>42,44,71-76</sup>

468 A few neuroimaging studies have investigated how these top-down knowledge cues  
469 and bottom-up stimulus cues influence perceptions of animacy and the flexibility of

470 our social cognitive system. One study found that stimulus cues overrode knowledge  
471 cues to animacy<sup>77</sup>; whereas, others found the inverse, knowledge, not stimulus, cues  
472 more strongly influenced animacy perception.<sup>43,78</sup> Yet, a key mentalizing region  
473 (rTPJ) was most sensitive when *both* stimulus and knowledge cues to animacy were  
474 presented compared to when only one (or none) of those cues were present.<sup>37</sup>  
475 These various findings are likely influenced by the type of task and cues used, and  
476 our study adds to the narrative that top-down knowledge-based cues of socialness  
477 can be just as, if not more, powerful for driving mind attribution during social  
478 interactions with artificial agents than bottom-up visual cues to human-likeness  
479 alone.

480 Therefore, physical features denoting human-likeness may not be the most important  
481 consideration for those designing socially engaging robots, and instead a  
482 reorientation toward an emphasis on socialness may be more fruitful for fostering  
483 social behaviours and attitudes toward robots. Ultimately, our findings set the stage  
484 for future work to disentangle not only which physical and social features play the  
485 most important roles in mind attribution to artificial agents, but also how ongoing  
486 experience with such agents changes and develops such perceptions.

#### 487 *Limitations & Future Directions*

488 Throughout our study we examined human-likeness and socialness using linear  
489 models. However, it is noteworthy that these concepts are frequently regarded as  
490 non-linear, especially when applied to social robotics.<sup>79,80</sup> Even in our study, results in  
491 one of the ROIs (the rTPJ) may have been better explained using a non-linear  
492 function. One possibility to explain why our linear models for human-likeness and  
493 socialness were still robust in most ROIs is that, while our humanoid had a human



494 shape (with a torso, arms, and head), neither our humanoid nor mechanoid robots  
495 approached realistic human-likeness. If we had included more realistic human-  
496 looking robots (androids) in the design, non-linear models may have offered a better  
497 model fit.<sup>80</sup> During the experimental design phase in future studies, consideration of  
498 which conditions might best test whether linear or curvilinear functions most  
499 parsimoniously account for neural activity, and whether which function best fits the  
500 data could vary across regions of interest, should be driven by several factors,  
501 including robot physical and social features.

502         Next, while we designed our video stimuli to be as believable as possible,  
503 ultimately 6 of our 42 participants did not believe that they were playing a live game.  
504 Removing the non-believers from analyses, however, did not change our overall  
505 findings (see OSF for more details). Thus, we consider our results to reflect brain  
506 response when participants are engaged in true real-time interactions with their  
507 game partners. In the past decade, a discussion has emerged around designing  
508 real-time social interactions in a genuinely interactive context. This movement is  
509 grounded in the understanding that social cognition may be fundamentally different  
510 during active versus passive social interactions, termed 'second-person  
511 neuroscience.'<sup>81</sup> A growing but comparatively small proportion of fMRI studies have  
512 attempted second-person neuroscience in human interactions; even fewer, to date,  
513 have attempted work at the intersection of social neuroscience and social robotics.  
514 However, in one fMRI study, participants engaged in real-time discussion via a live-  
515 feed interface with either a human or a conversational robot.<sup>82</sup> Their findings  
516 revealed increased neural activity during HHI compared to HRI in specific  
517 mentalising regions, most notably the TPJ (but not mPFC) and social motivation  
518 regions, including hypothalamus and amygdala. More neuroimaging work to date

519 has deployed technologies such as EEG or fNIRS to examine direct, embodied  
520 Human—Robot interactions (see <sup>9,23,83</sup> for a discussion). For example, in live-  
521 interactive paradigms with robots, most people used mechanistic terms to describe  
522 robots.<sup>84</sup> Further, whether someone tends to favor mentalistic or mechanistic  
523 explanations for robot behavior can be predicted from resting-state EEG signals  
524 before participants engage in describing robot behavior.<sup>85</sup> These studies highlight the  
525 value of a number of different neuroimaging techniques for exploring second-person  
526 neuroscience perspectives in the context of HRI. Our comprehension of real-time  
527 mechanisms and the outcomes of social engagement with robots hinges on  
528 combining these approaches with rigorous and theoretically driven experimental  
529 designs.

530 A further possible limitation in our study is that we used only people who identified as  
531 male and, therefore, we were not able to comment on gendered effects of  
532 mentalizing in the context of social interactions with robots. We chose male  
533 participants because one aim of the study was to replicate previous designs,<sup>27,28</sup>  
534 which also used only male samples. However, the influence of participant's gender  
535 on mentalising in the context of social robotics is an area of much needed  
536 investigation. The broader literature on gendered effects in mentalizing is mixed<sup>86,87</sup>  
537 but the prevailing narrative suggests that females have a “female advantage”, across  
538 cultures, on many social cognition measures, outperforming males on mentalizing  
539 tasks.<sup>88–91</sup> Indeed, one study found variations in mPFC activation during a ToM task  
540 to be more pronounced in women compared to men.<sup>92</sup> To further complicate matters,  
541 the gender of the human player may also be important. Both previous studies that  
542 informed our study design used a male human player; in our study, the human player  
543 was female. It is possible that this difference in study design could have influenced

544 participant strategy and possibly neural activation. Indeed, prior work suggests that  
545 participants play differently depending on the gender of their game partner.<sup>93,94</sup> It will  
546 be important to thoughtfully consider gender effects of the participants, human  
547 confederates, and perhaps even the perceived gender of the robots when planning  
548 future research studies using similar designs.

549 *Concluding thoughts*

550 Our primary findings confirm previous research that human-likeness plays an  
551 important role in the attribution of mind to robots. However, our exploratory analyses  
552 suggest that the perceived socialness of a robot also plays an equally, if not more  
553 important role than physical features denoting human-likeness in mind attribution.  
554 Incorporating knowledge- or experience-based social cues and features into robots  
555 who are designed to engage human users on a social level has the potential to  
556 increase user engagement and interest for more lasting and higher quality  
557 relationships with our robotic partners.

**558 Acknowledgements**

559 This work has received funding from the European Research Council under the  
560 European Union's Horizon 2020 research and innovation programme (Grant  
561 Agreement numbers: 677270 (Social Robots) & 716974: Becoming Social)).

562 The authors thank the reviewers for their constructive and helpful feedback.  
563 Additionally, the authors thank Nikolas Vitsakis, Kiara Jackson, and Jacynth Grundy  
564 for assistance with data collection and Dr. Julia Landsiedel for fMRI programming  
565 advice.

**566 Author Contributions**

567 Conceptualisation: LEJ, ESC, KK; Methodology: LEJ, ESC, KK, BC; Formal  
568 Analysis: LEJ; Investigation: LEJ, BC, SAA; Writing - Original Draft: LEJ, ESC, KK;  
569 Writing - Review & Editing: LEJ, ESC, KK, BC, SAA; Supervision: ESC & KK;  
570 Funding Acquisition: ESC & KK.

**571 Declaration of Interests**

572 The authors declare no competing interests.

573

**574 Supplementary Data**

575 See Supplementary Data section for more detail.

576

577 **Main Figure Titles and Legends**

578 **Figure 1.** Average percent signal change (PSC) during gameplay in mentalizing  
579 ROIs and pSTS with significant within subject rmANOVA (Error bars are SEM). See  
580 also Table S5 & Figure S2.

581 **Figure 2.** Whole brain T-map overlap analysis (Human > Computer (Red); Human >  
582 Humanoid (Blue); Human > Mechanoid (Green)). See also Table S4 & Figure S1.

583 **Figure 3.** Average Likert (0-10) scale ratings of Debrief questions (Error bars are  $\pm$   
584 SEM). See also Table S6 & Figure S3.

**585 STAR Methods****586 Resource Availability****587 Lead contact**

588 Further information for resources should be directed to Emily Cross

589 ([emily.cross@gess.ethz.ch](mailto:emily.cross@gess.ethz.ch))

**590 Materials availability**

591 Robot videos and example human and computer videos are provided on our OSF  
592 page (<https://osf.io/t4apv/>). See the key resources table for details.

593

**594 Data and code availability**

- 595 • The de-identified fMRI data have been compressed and deposited across 3 sites  
596 at Mendeley data and are publicly available as of the date of publication. See the  
597 key resources table for details.
- 598 • Code for the robot introduction and main experiment have been deposited on  
599 Github and are publicly available as of the date of publication. See the key  
600 resources table for details.
- 601 • Any additional information required to re-analyze the data reported in this paper  
602 is available from the lead contact upon request.

**603 EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS****604 Human Participants**

605 Due to the availability of scanning resources, participants were recruited from 2 sites:  
606 (i) the greater Glasgow area (Scotland, UK); and (ii) the greater Bangor area (Wales,  
607 UK). Glasgow participants completed the study at the Centre for Cognitive

608 Neuroimaging (CCNi) at the University of Glasgow, while Bangor participants  
609 completed the study at the Bangor Imaging Unit (BIU) at Bangor University.

610 Twenty right-handed males (mean age = 20.95 years; SD = 1.82; range = 19-26)  
611 participated from the greater Bangor area and 24 right-handed males (mean age =  
612 22.45 years; SD = 3.63; range = 18-32) participated from Glasgow. There were no  
613 significant differences in either age ( $t(40) = 1.76, p = .087$ ) or education ( $U = 206.50,$   
614  $Z = -.391, p = .696$ ) between data collection sites. Only males were recruited,  
615 consistent with previous studies in this area,<sup>27,28</sup> in order to control for any potential  
616 effects of gender on mentalizing.<sup>92</sup> Two participants withdrew from the study due to  
617 claustrophobia (1 subject from each site). The final fMRI participant sample included  
618 a total of 42 participants (mean age = 21.74 years; SD = 3.03; range = 18-32).

619 All participants reported normal or corrected-to-normal vision, no history of  
620 neurological or psychiatric disorders, and were right-handed as confirmed on the  
621 Edinburgh Handedness Questionnaire<sup>95</sup>; mean = 1.48, sd = .34).

622 All participants reported low familiarity with robots. Median engagement with robots  
623 in daily life (measured from 1 (never) to 7 (daily)) was 2 (IQR 1). Median number of  
624 robot-themed movies or TV shows seen was 4 (IQR 1) out of the 14 listed (Riek et  
625 al, 2011).<sup>96</sup>

626 All participants provided written informed consent prior to their involvement and  
627 received monetary compensation for study participation (£12/hour). All study  
628 procedures were approved by the respective university ethics boards: (i) Bangor  
629 University (Approval no. 2019-16639) and (ii) Glasgow University (Approval no.  
630 300180110).

631 Study site was a significantly different between subjects factor in rmANOVA for rTPJ,  
632 rmFG, & Precuneus but when we ran site separately for each of those ROIs, the  
633 results did not differ from the combined group or change the outcome; therefore,  
634 both sites were kept together in the results reported in this paper. Please see our  
635 OSF for details on the results from the separate groups. Further, we ran site as a  
636 covariate of no-interest in our model estimation and did not find differences in our  
637 whole brain data; therefore, the sites were subsequently analysed and reported  
638 together (please see our OSF for more details).

## 639 **METHOD DETAILS**

### 640 **Experimental Design**

641 We designed a Rock-Paper-Scissors (RPS) task similar to a previous study,<sup>26</sup> and  
642 followed a similar briefing procedure.<sup>27,28</sup> RPS was chosen for its familiarity across  
643 ages and cultures, and ease of rule learning. Previous studies have shown this game  
644 to engage mentalizing regions when played against human and non-human  
645 partners.<sup>27,30,97</sup>

646 Participants saw videos of their respective game partners before and after each 3-  
647 game series (refer to Figure 2). Each video was unique and all participants saw the  
648 same set of videos. During the pre-recorded videos, the human and robots reacted  
649 emotively to winning and losing a round. For example, the human and humanoid  
650 often put their hands up (or the forklift for the mechanoid) in exasperation when  
651 losing or happiness when winning. The mechanoid had expressive, pixelated eyes  
652 and was capable of moving within a restricted space on the table. Whereas, the  
653 humanoid had two lights for eyes that could flash but were not expressive and while  
654 the humanoid's arms, head, and torso could move, it did not move its position on the



655 floor during any interactions with participants. All robot videos, and example human  
656 videos, are available on our OSF and Mendeley data (see link in STAR table). The  
657 computer condition, which participants were told did not have an algorithm to win,  
658 involved a screensaver (Apple iMac 'Flurry') for both the pre- and post-game videos.  
659 During the game play, participants saw the same visual input across all 4 conditions,  
660 namely a score card across the top of the screen (for win/loss/tie in each series),  
661 pictures of the rock paper and scissors, and a countdown from 2 to 0 (refer to Figure  
662 S4 for an example). To minimize movement in the scanner, we did not utilize  
663 synchrony through a "fist-swing" as players might in real-life, rather participants were  
664 instructed to select their RPS choice from a button box on '0' in the countdown. The  
665 button press for rock, paper, and scissors and order of the items on the screen  
666 during gameplay were assigned randomly across participants.

667 In-line with previous designs,<sup>27,28,97</sup> participants were told that they were playing a  
668 live game and viewing their game-partners through a live video feed, but in reality,  
669 neither the remote practice nor the in-scanner games (described below) were live. All  
670 videos were pre-recorded and designed to give the impression of a live game. Wins  
671 and losses were controlled across the four conditions so that each participant won  
672 10 rounds and lost 10 rounds against each partner. The order in which participants  
673 played partners was pseudo-randomized across four 8-minute functional runs.

674 To give the impression of a live game, participants met all game partners in person  
675 in the "game room" and played one truly live, in-person, round of rock-paper-scissors  
676 with each partner. They went to the imaging suite to play a "live" practice round of  
677 RPS with their partners via the "video feed". This practice round served to familiarise  
678 participants with the game and practice pushing the buttons to register their answer

679 with the correct timing. Participants played each partner twice in each practice round  
680 and could complete up to 3 practice rounds (24 total games) to ensure they  
681 understood the game before entering the scanner. All participants demonstrated  
682 understanding of the game and button presses by the 3<sup>rd</sup> practice round.

683 Participants then completed the fMRI task, playing the same RPS game. Each fMRI  
684 run contained 5 rounds with each of the 4 partners (20 rounds per partner across all  
685 4 runs), pseudorandomized across participants. In total, each participant completed  
686 four, 8-minute RPS runs. After the scan, participants completed several  
687 questionnaires (listed below) on a laptop and were then debriefed. The debriefing  
688 unveiled the study deception (that the various game partners were pre-recorded, not  
689 live, and that all partners used the same random algorithm and were not  
690 independently controlled). Both the practice round and game in the scanner were  
691 programmed in Python 3.7 and run from the command line (see STAR Methods  
692 Table).

### 693 **MRI Parameters, Pre-processing, & GLM Estimation**

694 At both data collection sites (CCNI and BIU), stimuli were projected onto a mirror  
695 from a projector located behind the scanner. Responses were recorded with an MRI-  
696 compatible keypad.

697 A dual-echo EPI sequence was used to improve signal-to-noise ratio (SNR) in frontal  
698 and temporal regions.<sup>98</sup> All structural and functional sequence parameters are  
699 detailed in Tables S1 & S2.

700 Data pre-processing was carried out in SPM12 (Wellcome Trust Centre for  
701 Neuroimaging, London) implemented in Matlab 2018a (Mathworks, Natick, MA,  
702 USA). Pre-processing consisted of standard SPM12 defaults for slice time

703 correction, realignment and re-slicing, co-registration, unified segmentation &  
704 normalisation, and smoothing; except for a 6mm FWHM Gaussian smoothing kernel.  
705 All analyses were performed in normalized MNI space. Block durations and onsets  
706 for each of the 4 experimental conditions during Video 1, the RPS game, and Video  
707 2 were modelled by convolving the hemodynamic response function and with a high  
708 pass filter of 128s. Head motion parameters were modelled as nuisance regressors.  
709 Functional scans provided whole brain coverage.

### 710 **ROI Creation & Analyses**

711 Our choice of ROIs was informed by previous studies<sup>27,28</sup>; however, ROI placement  
712 was based on peak activation from the independent localizers (refer to Figure S2 &  
713 Table S3). Participants undertook two passive-viewing tasks to help identify brain  
714 regions of interest after playing the RPS game.

715 *Mentalizing.* Localizer 1 was a short-animated film ('Partly Cloudy'; Pixar Animation  
716 Studios, 2009) coded for event type (mentalizing, pain, social, and control). We used  
717 the mentalizing > pain contrast to identify ROI coordinates in bilateral TPJ, bilateral  
718 mFG, and Precuneus independently from our main experimental task. Neither Medial  
719 Prefrontal Cortex (mPFC) nor rmFG activation appeared as expected in Localiser 1,  
720 therefore, we used mPFC & rmFG coordinates from the original localiser paper<sup>50</sup> and  
721 created 6mm spheres around those coordinates.

722 *Social Interaction.* To localize pSTS, we employed an established localizer which  
723 involves passive viewing of 3 conditions: (i) interacting, (ii) non-interacting, and (iii)  
724 scrambled point-light figures.<sup>52,57</sup> We used the interaction > scrambled contrast (i.e.,  
725 two human point light figures interacting vs. scrambled dot motion) to derive our  
726 pSTS coordinates independently from our experimental task.

727 We used a control ROI (V1/BA17) from the WFU PickAtlas<sup>99</sup> as a form of verification  
728 that activity differences seen between conditions during game play was not  
729 attributable to non-specific whole brain activation differences. In other words, we  
730 would not expect differences between conditions in V1 activity during game play, as  
731 participants saw the same set-up across all conditions, and this control ROI allowed  
732 us to evaluate this possibility.

733 Group-constrained, subject specific ROIs were created like the methods described  
734 elsewhere<sup>52</sup> using an uncorrected height threshold of  $p < .0001$ . This protocol  
735 creates subject-specific ROIs based on independent data (i.e. localizers). Briefly, we  
736 established an initial 6mm bounding sphere centred around the peak T-value from  
737 group activation in our pre-registered localizer contrasts (i.e. interacting vs non-  
738 Interacting, mentalizing vs pain). Within this initial bounding sphere, we employed a  
739 leave one subject out (LOSO) iterative process based on group level analyses,  
740 resulting in a more refined search sphere. Finally, we generated subject specific  
741 regions of interest (ROIs) within this constrained search space by selecting the top  
742 100 contiguous voxels for each subject, thereby accounting for inter-subject  
743 variability within these restricted search spaces. Percent signal change was then  
744 extracted from ROIs using in-house scripts in Matlab 2018a and the MarsBar  
745 toolbox.

746 None of the ROIs overlapped. Both right and left TPJ were slightly shifted so the  
747 entire sphere was within the boundaries of the brain; all other ROIs created from the  
748 localisers remained true to the peak activation. Please refer to Supplementary  
749 Figures for all ROI coordinates.

## 750 **Behavioural Measures**

## 751 **Debrief Questions**

752 *Pre-Registered*. After scanning, participants answered questions about their  
753 experience of the study using FormR.<sup>100</sup> Participants rated responses to the  
754 following questions on a scale from 0-10: (i) how well they were able to adopt an  
755 efficient *strategy* against each partner, (ii) how *successful* they were against each  
756 partner, (iii) how much *fun* it was to play each partner, (iv) how much *sympathy* they  
757 had for each partner when they lost, and then each partner's (v) *competitiveness*,  
758 and (v) *intelligence*.

### 759 **Inclusion of Others and Self (IOS)**

760 The Inclusion of Others and Self (IOS) is a measure of closeness and  
761 interconnectedness between two individuals.<sup>101</sup> A series of 7 increasingly  
762 overlapping circles are presented to the participant on paper. Each pair of circles  
763 contains the word "self" in one circle and "other" in the other circle. Participants are  
764 then asked to choose which circle represents their relationship to the agent in  
765 question. We asked participants to show which set of overlapping circles best  
766 describes the following agents: (1) computer, (2) mechanoid robot, (3) humanoid  
767 robot, (4) a human stranger, (5) the human from the experiment (LEJ), and (6) a  
768 close friend. Non-robot items were included for comparison to determine where the  
769 robot stood relative to other people in the participant's lives. The IOS provides  
770 another way to address the participant's view of their relationship to various humans  
771 and robots. Responses from the paper and pencil format of the IOS were recorded  
772 onto a 7-point scale from 1 (no overlap) to 7 (nearly complete overlap).

## 773 **QUANTIFICATION AND STATISTICAL ANALYSES**

### 774 **fMRI Analyses**

#### 775 **ROI**

776 *Pre-Registered.* Repeated measures ANOVAs were run for each ROI to assess the  
777 effect of game-partner and pairwise comparisons were run only if a main effect of  
778 game-partner was found. All pairwise comparisons were corrected for multiple  
779 comparisons (Bonferroni). Greenhouse-Geisser corrections were made if any  
780 rmANOVA was found to violate Mauchley's tests of sphericity. We assessed the  
781 linear effect of human-likeness using a linear repeated contrast in a within-subject  
782 ANOVA, which compares means across the different levels of the independent  
783 variable according to the following order: computer < mechanoid < humanoid <  
784 human.

785 *Exploratory.* Ratings results from the *Fun*, *Competitiveness*, and *Sympathy*  
786 questions in the Debrief, suggested swapping the robot orders in the linear model  
787 (see below). As an exploratory analysis, we ran a linear repeated contrast in a  
788 within-subject ANOVA to compare means across different levels of the independent  
789 variable according to the following order based on socialness ratings: computer <  
790 humanoid < mechanoid < human.

791 Additionally, we assessed whether the pSTS would show a linear pattern based on  
792 human-likeness or socialness during game play and whilst watching the video  
793 introduction which preceded each round.

#### 794 **Whole Brain**

795 *Pre-Registered.* A GLM comprising the four conditions (CP = Computer Partner, MR  
796 = Mechanoid Robot, HR= Humanoid Robot, HP = Human Partner) was specified for  
797 each participant. Simple contrasts were compared against: (1) HP > CP, (2) HR >  
798 CP, (3) MR > CP, (4) HR > MR, (5) HP > HR. Based on previous findings (Krach et  
799 al, 2010) and our hypothesis, we expected to see a linear increase in neural activity  
800 based on human-likeness of agent. To evaluate this, we calculated a parametric

801 modulation of gameplay partner (actual model weights used: CP = -3, MR = -1, HR =  
802 1, HP = 3). For the second level group analyses, we used a FWE-corrected  
803 threshold ( $p_{\text{uncorr}} < 0.001$ ) and a minimum cluster size ( $k = 100$ ).

804 *Exploratory.* While not pre-registered, we also included the following simple  
805 contrasts: (6) HP > MR, (7) MR > HR. We also calculated the parametric modulation  
806 of gameplay partners based on socialness (actual model weights used: CP = -3, HR  
807 = -1, MR = 1, HP = 3).

## 808 **Behavioral Analyses**

### 809 **Debrief Questions**

810 *Pre-Registered.* As pre-registered, rmANOVAs were run on each question to assess  
811 the effect of agent. Pairwise comparisons between agents were run only if an agent  
812 effect was identified. All pairwise comparisons were corrected for multiple  
813 comparisons (Bonferroni). Greenhouse-Geisser corrections were made if any  
814 rmANOVA was found to violate Mauchley's tests of sphericity. We assessed the  
815 linear effect of human-likeness using a linear repeated contrast in a within-subject  
816 ANOVA, which compares means across different levels of the independent variable.

817 *Exploratory.* Furthermore, based on participant-reported perceptions of socialness of  
818 the individual agents, we ran an exploratory (not pre-registered) linear repeated  
819 contrast in a within-subjects ANOVA that reversed the order of the robots in the 4-  
820 element hierarchy within the linear model.

### 821 **Inclusion of Others and Self (IOS)**

822 *Pre-Registered.* As pre-registered, rmANOVA was run to assess the effect of agent  
823 and pairwise comparisons were run only if an effect of agent was found. All pairwise  
824 comparisons were corrected for multiple comparisons (Bonferroni). Greenhouse-

825 Geisser corrections were made if any rmANOVA was found to violate Mauchley's  
826 tests of sphericity.

827

Journal Pre-proof



## References

1. Turing, A.M. (2012). Computing machinery and intelligence. In *Machine Intelligence: Perspectives on the Computational Model* 10.7551/mitpress/6928.003.0012.
2. Shukla, A., and Karki, H. (2016). Application of robotics in onshore oil and gas industry—A review Part I. *Rob Auton Syst* 75, 490–507. 10.1016/j.robot.2015.09.012.
3. Kas, K.A., and Johnson, G.K. (2020). Using unmanned aerial vehicles and robotics in hazardous locations safely. *Process Safety Progress* 39. 10.1002/prs.12066.
4. Cifuentes, C.A., Pinto, M.J., Céspedes, N., and Múnera, M. (2020). Social Robots in Therapy and Care. *Current Robotics Reports* 1, 59–74. 10.1007/s43154-020-00009-2.
5. Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., and Tanaka, F. (2018). Social robots for education: A review. *Sci Robot* 3, eaat5954. 10.1126/scirobotics.aat5954.
6. Dawe, J., Sutherland, C., Barco, A., and Broadbent, E. (2019). Can social robots help children in healthcare contexts? A scoping review. *BMJ Paediatr Open* 3, e000371. 10.1136/bmjpo-2018-000371.
7. Drexler, N., and Lapré, V.B. (2019). For better or for worse: Shaping the hospitality industry through robotics and artificial intelligence. *Research in Hospitality Management* 9, 117–120. 10.1080/22243534.2019.1689701.
8. Mann, J.A., MacDonald, B.A., Kuo, I.-H., Li, X., and Broadbent, E. (2015). People respond better to robots than computer tablets delivering healthcare instructions. *Comput Human Behav* 43, 112–117. 10.1016/j.chb.2014.10.029.
9. Cross, E.S., and Ramsey, R. (2020). Mind Meets Machine: Towards a Cognitive Science of Human–Machine Interactions. *Trends Cogn Sci* 25, 200–212. 10.1016/j.tics.2020.11.009.
10. Caruana, N., and Cross, E.S. (2023). Autonomous social robots are real in the mind’s eye of many. *Behavioral and Brain Sciences* 46, e26. 10.1017/s0140525x22001625.
11. Frith, C.D., and Frith, U. (1999). Interacting Minds—A Biological Basis. *Science* (1979) 286, 1692–1695. 10.1126/science.286.5445.1692.
12. Saxe, R., Carey, S., and Kanwisher, N. (2004). Understanding Other Minds: Linking Developmental Psychology and Functional Neuroimaging. *Psychology* 55, 87–124. 10.1146/annurev.psych.55.090902.142044.
13. Schurz, M., Radua, J., Aichhorn, M., Richlan, F., and Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neurosci Biobehav Rev* 42, 9–34. 10.1016/j.neubiorev.2014.01.009.
14. Frith, C.D., and Frith, U. (2021). The Neural Basis of Mentalizing. 17–45. 10.1007/978-3-030-51890-5\_2.
15. Frith, U., and Frith, C.D. (2003). Development and neurophysiology of mentalizing. *Philos Trans R Soc Lond B Biol Sci* 358, 459–473. 10.1098/rstb.2002.1218.
16. Overwalle, F. Van, and Baetens, K. (2009). Understanding others’ actions and goals by mirror and mentalizing systems: A meta-analysis. *Neuroimage* 48, 564–584. 10.1016/j.neuroimage.2009.06.009.
17. Overwalle, F. Van (2009). Social cognition and the brain: a meta-analysis. *Hum Brain Mapp* 30, 829–858. 10.1002/hbm.20547.

18. Molenberghs, P., Johnson, H., Henry, J.D., and Mattingley, J.B. (2016). Understanding the minds of others: A neuroimaging meta-analysis. *Neurosci Biobehav Rev* 65, 276–291. 10.1016/j.neubiorev.2016.03.020.
19. Frith, C.D., and Frith, U. (2021). The Neural Basis of Mentalizing. 17–45. 10.1007/978-3-030-51890-5\_2.
20. Wiese, E., Wykowska, A., Zwickel, J., and Müller, H.J. (2012). I See What You Mean: How Attentional Selection Is Shaped by Ascribing Intentions to Others. *PLoS One* 7, e45391. 10.1371/journal.pone.0045391.
21. Wykowska, A., Wiese, E., Prosser, A., and Müller, H.J. (2014). Beliefs about the Minds of Others Influence How We Process Sensory Information. *PLoS One* 9, e94339. 10.1371/journal.pone.0094339.
22. Özdem, C., Wiese, E., Wykowska, A., Müller, H., Brass, M., and Overwalle, F. Van (2016). Believing androids – fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents. *Soc Neurosci* 12, 1–12. 10.1080/17470919.2016.1207702.
23. Henschel, A., Hortensius, R., and Cross, E.S. (2020). Social Cognition in the Age of Human–Robot Interaction. *Trends Neurosci* 43, 373–384. 10.1016/j.tins.2020.03.013.
24. Cross, E.S., Riddoch, K.A., Pratts, J., Titone, S., Chaudhury, B., and Hortensius, R. (2019). A neurocognitive investigation of the impact of socializing with a robot on empathy for pain. *Philosophical Transactions of the Royal Society B* 374, 20180034. 10.1098/rstb.2018.0034.
25. Pütten, A.M.R. der, Schulte, F.P., Eimler, S.C., Sobieraj, S., Hoffmann, L., Maderwald, S., Brand, M., and Krämer, N.C. (2014). Investigations on empathy towards humans and robots using fMRI. *Comput Human Behav* 33, 201–212. 10.1016/j.chb.2014.01.004.
26. Wiese, E., Buzzell, G.A., Abubshait, A., and Beatty, P.J. (2018). Seeing minds in others: Mind perception modulates low-level social-cognitive performance and relates to ventromedial prefrontal structures. *Cogn Affect Behav Neurosci* 18, 837–856. 10.3758/s13415-018-0608-2.
27. Chaminade, T., Rosset, D., Fonseca, D. Da, Nazarian, B., Lutcher, E., Cheng, G., and Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Front Hum Neurosci* 6, 103. 10.3389/fnhum.2012.00103.
28. Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., and Kircher, T. (2008). Can Machines Think? Interaction and Perspective Taking with Robots Investigated via fMRI. *PLoS One* 3, e2597. 10.1371/journal.pone.0002597.
29. Takahashi, H., Terada, K., Morita, T., Suzuki, S., Haji, T., Kozima, H., Yoshikawa, M., Matsumoto, Y., Omori, T., Asada, M., et al. (2014). Different impressions of other agents obtained through social interaction uniquely modulate dorsal and ventral pathway activities in the social human brain. *Cortex* 58, 289–300. 10.1016/j.cortex.2014.03.011.
30. Chaminade, T., Fonseca, D. Da, Rosset, D., Cheng, G., and Deruelle, C. (2015). Atypical modulation of hypothalamic activity by social context in ASD. *Res Autism Spectr Disord* 10, 41–50. 10.1016/j.rasd.2014.10.015.
31. Meltzoff, A.N. (2007). The ‘like me’ framework for recognizing and becoming an intentional agent. *Acta Psychol (Amst)* 124, 26–43. 10.1016/j.actpsy.2006.09.005.
32. Duffy, B.R., and Joue, G. (2004). I, robot being. *Intelligent Autonomous Systems* 8.

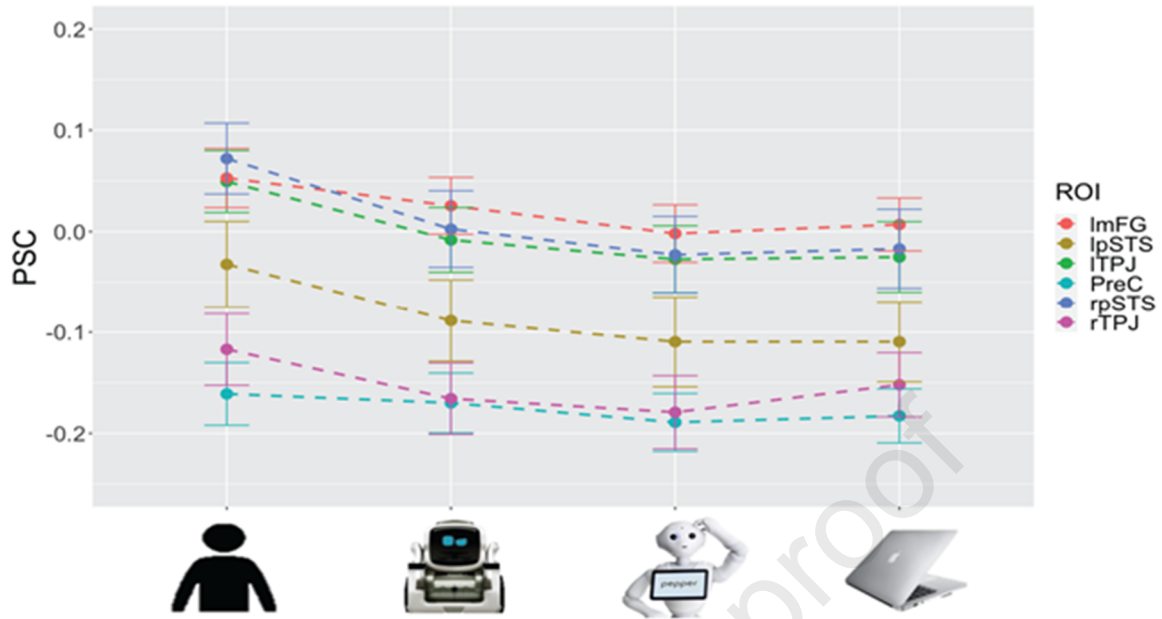
33. Goetz, J., Kiesler, S., and Powers, A. (2003). Matching Robot Appearance and Behavior to Tasks to Improve Human-Robot Cooperation. The 12th IEEE International Workshop on Robot and Human Interactive Communication, 2003. Proceedings. ROMAN 2003, 55–60. 10.1109/roman.2003.1251796.
34. Abubshait, A., and Wiese, E. (2017). You Look Human, But Act Like a Machine: Agent Appearance and Behavior Modulate Different Aspects of Human-Robot Interaction. *Front Psychol* 8, 1393. 10.3389/fpsyg.2017.01393.
35. Cross, E.S., Liepelt, R., Hamilton, A.F. de C., Parkinson, J., Ramsey, R., Stadler, W., and Prinz, W. (2012). Robotic movement preferentially engages the action observation network. *Hum Brain Mapp* 33, 2238–2254. 10.1002/hbm.21361.
36. Teufel, C., Fletcher, P.C., and Davis, G. (2010). Seeing other minds: attributed mental states influence perception. *Trends Cogn Sci* 14, 376–382. 10.1016/j.tics.2010.05.005.
37. Klapper, A., Ramsey, R., Wigboldus, D., and Cross, E.S. (2014). The Control of Automatic Imitation Based on Bottom–Up and Top–Down Cues to Animacy: Insights from Brain and Behavior. *J Cogn Neurosci* 26, 2503–2513. 10.1162/jocn\_a\_00651.
38. Epley, N., Waytz, A., and Cacioppo, J.T. (2007). On Seeing Human: A Three-Factor Theory of Anthropomorphism. *Psychol Rev* 114, 864–886. 10.1037/0033-295x.114.4.864.
39. Kiesler, S., Powers, A., Fussell, S.R., and Torrey, C. (2008). Anthropomorphic Interactions with a Robot and Robot–like Agent. *Soc Cogn* 26, 169–181. 10.1521/soco.2008.26.2.169.
40. Tung, F.-W. (2011). Human-Computer Interaction. Users and Applications, 14th International Conference, HCI International 2011, Orlando, FL, USA, July 9-14, 2011, Proceedings, Part IV. *Lecture Notes in Computer Science*, 637–646. 10.1007/978-3-642-21619-0\_76.
41. DiSalvo, C.F., Gemperle, F., Forlizzi, J., and Kiesler, S. (2002). All robots are not created equal: the design and perception of humanoid robot heads. Proceedings of the 4th conference on Designing interactive systems: processes, practices, methods, and techniques, 321–326. 10.1145/778712.778756.
42. Henschel, A., Laban, G., and Cross, E.S. (2021). What Makes a Robot Social? A Review of Social Robots from Science Fiction to a Home or Hospital Near You. *Current Robotics Reports* 2, 9–19. 10.1007/s43154-020-00035-0.
43. Cross, E.S., Ramsey, R., Liepelt, R., Prinz, W., and Hamilton, A.F. de C. (2016). The shaping of social perception by stimulus and knowledge cues to human animacy. *Philosophical Transactions of the Royal Society B: Biological Sciences* 371, 20150075. 10.1098/rstb.2015.0075.
44. Hsieh, T.-Y., Chaudhury, B., and Cross, E.S. (2023). Human–Robot Cooperation in Economic Games: People Show Strong Reciprocity but Conditional Prosociality Toward Robots. *Int J Soc Robot* 15, 791–805. 10.1007/s12369-023-00981-7.
45. Hsieh, T.-Y., and Cross, E.S. (2022). People’s dispositional cooperative tendencies towards robots are unaffected by robots’ negative emotional displays in prisoner’s dilemma games. *Cogn Emot* 36, 995–1019. 10.1080/02699931.2022.2054781.
46. Andriella, A., Siqueira, H., Fu, D., Magg, S., Barros, P., Wermter, S., Torras, C., and Alenyà, G. (2021). Do I Have a Personality? Endowing Care Robots with Context-Dependent Personality Traits. *Int J Soc Robot* 13, 2081–2102. 10.1007/s12369-020-00690-5.

47. Whittaker, S., Rogers, Y., Petrovskaya, E., and Zhuang, H. (2021). Designing Personas for Expressive Robots. *ACM Transactions on Human-Robot Interaction (THRI)* 10, 1–25. 10.1145/3424153.
48. Kirby, R., Forlizzi, J., and Simmons, R. (2010). Affective social robots. *Rob Auton Syst* 58, 322–332. 10.1016/j.robot.2009.09.015.
49. Graaf, M.M.A. de, and Allouch, S. Ben (2013). Exploring influencing variables for the acceptance of social robots. *Rob Auton Syst* 61, 1476–1486. 10.1016/j.robot.2013.07.007.
50. Jacoby, N., Bruneau, E., Koster-Hale, J., and Saxe, R. (2016). Localizing Pain Matrix and Theory of Mind networks with both verbal and non-verbal stimuli. *Neuroimage* 126, 39–48. 10.1016/j.neuroimage.2015.11.025.
51. Schilbach, L. (2014). On the relationship of online and offline social cognition. *Front Hum Neurosci* 8, 278. 10.3389/fnhum.2014.00278.
52. Walbrin, J., and Koldewyn, K. (2019). Dyadic interaction processing in the posterior temporal cortex. *Neuroimage* 198, 296–302. 10.1016/j.neuroimage.2019.05.027.
53. Deen, B., Koldewyn, K., Kanwisher, N., and Saxe, R. (2015). Functional Organization of Social Perception and Cognition in the Superior Temporal Sulcus. *Cerebral Cortex* 25, 4596–4609. 10.1093/cercor/bhv111.
54. Landsiedel, J., Daughters, K., Downing, P.E., and Koldewyn, K. (2022). The role of motion in the neural representation of social interactions in the posterior temporal cortex. *Neuroimage* 262, 119533. 10.1016/j.neuroimage.2022.119533.
55. Castelli, F., Happé, F., Frith, U., and Frith, C. (2000). Movement and Mind: A Functional Imaging Study of Perception and Interpretation of Complex Intentional Movement Patterns. *Neuroimage* 12, 314–325. 10.1006/nimg.2000.0612.
56. Frith, U., and Frith, C. (2010). The social brain: allowing humans to boldly go where no other species has been. *Philosophical Transactions of the Royal Society B: Biological Sciences* 365, 165–176. 10.1098/rstb.2009.0160.
57. Isik, L., Koldewyn, K., Beeler, D., and Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences* 114, E9145–E9152. 10.1073/pnas.1714471114.
58. Walbrin, J., Downing, P., and Koldewyn, K. (2018). Neural responses to visually observed social interactions. *Neuropsychologia* 112, 31–39. 10.1016/j.neuropsychologia.2018.02.023.
59. Wheatley, T., Milleville, S.C., and Martin, A. (2006). Understanding Animate Agents. *Psychol Sci* 18, 469–474. 10.1111/j.1467-9280.2007.01923.x.
60. Gallagher, H.L., Jack, A.I., Roepstorff, A., and Frith, C.D. (2002). Imaging the Intentional Stance in a Competitive Game. *Neuroimage* 16, 814–821. 10.1006/nimg.2002.1117.
61. McKiernan, K.A., Kaufman, J.N., Kucera-Thompson, J., and Binder, J.R. (2003). A Parametric Manipulation of Factors Affecting Task-induced Deactivation in Functional Neuroimaging. *J Cogn Neurosci* 15, 394–408. 10.1162/089892903321593117.
62. Visser, M., Jefferies, E., and Ralph, M.A.L. (2010). Semantic Processing in the Anterior Temporal Lobes: A Meta-analysis of the Functional Neuroimaging Literature. *J Cogn Neurosci* 22, 1083–1094. 10.1162/jocn.2009.21309.

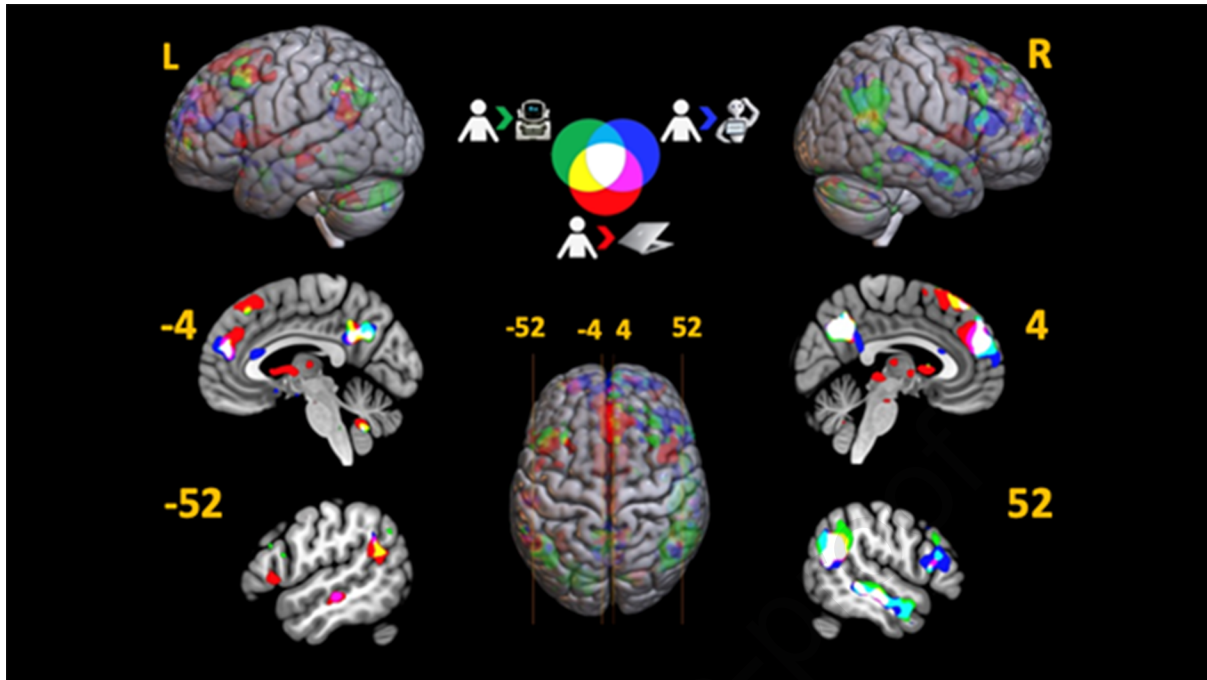
63. Binder, J.R., Desai, R.H., Graves, W.W., and Conant, L.L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex* *19*. 10.1093/cercor/bhp055.
64. Shine, J.M., and Breakspear, M. (2018). Understanding the Brain, By Default. *Trends Neurosci* *41*, 244–247. 10.1016/j.tins.2018.03.004.
65. Lu, L., Zhang, P., and Zhang, T. (Christina) (2021). Leveraging “human-likeness” of robotic service at restaurants. *Int J Hosp Manag* *94*, 102823. 10.1016/j.ijhm.2020.102823.
66. Fink, J. (2012). *Lecture Notes in Computer Science*. 199–208. 10.1007/978-3-642-34103-8\_20.
67. Rothstein, N., Kounios, J., Ayaz, H., and Visser, E.J. de (2020). Advances in Neuroergonomics and Cognitive Engineering, Proceedings of the AHFE 2020 Virtual Conferences on Neuroergonomics and Cognitive Engineering, and Industrial Cognitive Ergonomics and Engineering Psychology, July 16-20, 2020, USA. *Advances in Intelligent Systems and Computing*, 190–196. 10.1007/978-3-030-51041-1\_26.
68. Roselli, C., Ciardo, F., Tommaso, D. De, and Wykowska, A. (2022). Human-likeness and attribution of intentionality predict vicarious sense of agency over humanoid robot actions. *Sci Rep* *12*, 13845. 10.1038/s41598-022-18151-6.
69. Mejia, C., and Kajikawa, Y. (2017). Assessing the Sentiment of Social Expectations of Robotic Technologies. 2017 Portland International Conference on Management of Engineering and Technology (PICMET), 1–7. 10.23919/picmet.2017.8125441.
70. Ishiguro, H., and Nishio, S. (2007). Building artificial humans to understand humans. *Journal of Artificial Organs* *10*, 133–142. 10.1007/s10047-007-0381-4.
71. Reeves, B., Hancock, J., and Liu, X. (2020). Social robots are like real people: First impressions, attributes, and stereotyping of social robots. *Technology, Mind, and Behavior* *1*. 10.1037/tmb0000018.
72. Coradeschi, S., Ishiguro, H., Asada, M., Shapiro, S.C., Thielscher, M., Breazeal, C., Mataric, M.J., and Ishida, H. (2006). Human-Inspired Robots. *IEEE Intell Syst* *21*, 74–85. 10.1109/mis.2006.72.
73. Polakow, T., Laban, G., Teodorescu, A., Busemeyer, J.R., and Gordon, G. (2022). Social robot advisors: effects of robot judgmental fallacies and context. *Intell Serv Robot* *15*, 593–609. 10.1007/s11370-022-00438-2.
74. Breazeal, C. (2003). Emotion and sociable humanoid robots. *Int J Hum Comput Stud* *59*, 119–155. 10.1016/s1071-5819(03)00018-1.
75. Hortensius, R., and Cross, E.S. (2018). From automata to animate beings: the scope and limits of attributing socialness to artificial agents. *Ann N Y Acad Sci* *1426*, 93–110. 10.1111/nyas.13727.
76. Wykowska, A., Chaminade, T., and Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences* *371*, 20150375. 10.1098/rstb.2015.0375.
77. Press, C., Gillmeister, H., and Heyes, C. (2006). Bottom-up, not top-down, modulation of imitation by human and robotic models. *European Journal of Neuroscience* *24*, 2415–2419. 10.1111/j.1460-9568.2006.05115.x.

78. Stanley, J., Gowen, E., and Miall, R.C. (2010). How instructions modify perception: An fMRI study investigating brain areas involved in attributing human agency. *Neuroimage* 52, 389–400. 10.1016/j.neuroimage.2010.04.025.
79. MacDorman, K.F., and Ishiguro, H. (2006). The uncanny advantage of using androids in cognitive and social science research. *Interaction Studies. Social Behaviour and Communication in Biological and Artificial Systems* 7, 297–337. 10.1075/is.7.3.03mac.
80. Wang, S., Lilienfeld, S.O., and Rochat, P. (2015). The Uncanny Valley: Existence and Explanations. *Review of General Psychology* 19, 393–407. 10.1037/gpr0000056.
81. Redcay, E., and Schilbach, L. (2019). Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nat Rev Neurosci* 20, 495–505. 10.1038/s41583-019-0179-4.
82. Rauchbauer, B., Nazarian, B., Bourhis, M., Ochs, M., Prévot, L., and Chaminade, T. (2019). Brain activity during reciprocal social interaction investigated using conversational robots as control condition. *Philosophical Transactions of the Royal Society B* 374, 20180033. 10.1098/rstb.2018.0033.
83. Wykowska, A. (2021). Robots as Mirrors of the Human Mind. *Curr Dir Psychol Sci* 30, 34–40. 10.1177/0963721420978609.
84. Marchesi, S., Ghiglini, D., Ciardo, F., Perez-Osorio, J., Baykara, E., and Wykowska, A. (2019). Do We Adopt the Intentional Stance Toward Humanoid Robots? *Front Psychol* 10, 450. 10.3389/fpsyg.2019.00450.
85. Bossi, F., Willemse, C., Cavazza, J., Marchesi, S., Murino, V., and Wykowska, A. (2020). The human brain reveals resting state activity patterns that are predictive of biases in attitudes toward robots. *Sci Robot* 5. 10.1126/scirobotics.abb6652.
86. Rutherford, H.J.V., Wareham, J.D., Vrouva, I., Mayes, L.C., Fonagy, P., and Potenza, M.N. (2012). Sex differences moderate the relationship between adolescent language and mentalization. *Personality Disorders: Theory, Research, and Treatment* 3. 10.1037/a0028938.
87. McDonald, B., and Kanske, P. (2023). Gender differences in empathy, compassion, and prosocial donations, but not theory of mind in a naturalistic social task. *Sci Rep* 13, 20748. 10.1038/s41598-023-47747-9.
88. Kirkland, R.A., Peterson, E., Baker, C.A., Miller, S., and Pulos, S. (2013). Meta-analysis reveals adult female superiority in “Reading the mind in the eyes test.” *N Am J Psychol* 15.
89. Baron-Cohen, S., Radecki, M.A., Greenberg, D.M., Warrier, V., Holt, R.J., and Allison, C. (2022). Sex differences in theory of mind: The on-average female advantage on the Reading the Mind in the Eyes Test. Preprint, 10.1111/dmnc.15364 10.1111/dmnc.15364.
90. Greenberg, D.M., Warrier, V., Abu-Akel, A., Allison, C., Gajos, K.Z., Reinecke, K., Rentfrow, P.J., Radecki, M.A., and Baron-Cohen, S. (2023). Sex and age differences in “theory of mind” across 57 countries using the English version of the “Reading the Mind in the Eyes” Test. *Proceedings of the National Academy of Sciences* 120, e2022385119. 10.1073/pnas.2022385119.
91. Poznyak, E., Morosan, L., Perroud, N., Speranza, M., Badoud, D., and Debbané, M. (2019). Roles of age, gender and psychological difficulties in adolescent mentalizing. *J Adolesc* 74, 120–129. 10.1016/j.adolescence.2019.06.007.

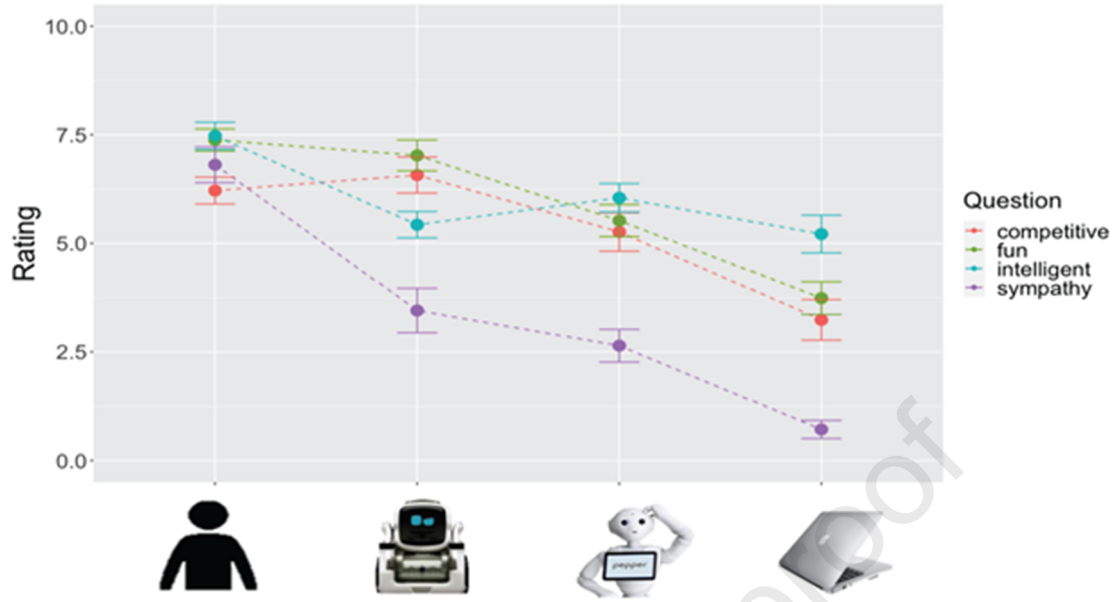
92. Krach, S., Blümel, I., Marjoram, D., Lataster, T., Krabbendam, L., Weber, J., Os, J. van, and Kircher, T. (2009). Are women better mindreaders? Sex differences in neural correlates of mentalizing detected with functional MRI. *BMC Neurosci* *10*, 9. 10.1186/1471-2202-10-9.
93. Skotko, V., Langmeyer, D., and Lundgren, D. (1974). Sex Differences as Artifact in the Prisoner's Dilemma Game. *Journal of Conflict Resolution* *18*, 707–713. 10.1177/002200277401800411.
94. Balliet, D., Li, N.P., Macfarlan, S.J., and Vugt, M. Van (2011). Sex Differences in Cooperation: A Meta-Analytic Review of Social Dilemmas. *Psychol Bull* *137*, 881–909. 10.1037/a0025354.
95. Oldfield, R.C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia* *9*, 97–113. 10.1016/0028-3932(71)90067-4.
96. Riek, L.D., Rabinowitch, T., Chakrabarti, B., and Robinson, P. (2009). How anthropomorphism affects empathy toward robots. 2009 4th ACM/IEEE International Conference on Human-Robot Interaction (HRI), 245–246. 10.1145/1514095.1514158.
97. Gallagher, H.L., Jack, A.I., Roepstorff, A., and Frith, C.D. (2002). Imaging the Intentional Stance in a Competitive Game. *Neuroimage* *16*, 814–821. 10.1006/nimg.2002.1117.
98. Halai, A.D., Welbourne, S.R., Embleton, K., and Parkes, L.M. (2014). A comparison of dual gradient-echo and spin-echo fMRI of the inferior temporal lobe. *Hum Brain Mapp* *35*, 4118–4128. 10.1002/hbm.22463.
99. Maldjian, J.A., Laurienti, P.J., Kraft, R.A., and Burdette, J.H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage* *19*, 1233–1239. 10.1016/s1053-8119(03)00169-1.
100. Arslan, R.C., Walther, M.P., and Tata, C.S. (2020). formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behav Res Methods* *52*, 376–387. 10.3758/s13428-019-01236-y.
101. Aron, A., Aron, E.N., and Smollan, D. (1992). Inclusion of Other in the Self Scale and the Structure of Interpersonal Closeness. *J Pers Soc Psychol* *63*, 596–612. 10.1037/0022-3514.63.4.596.

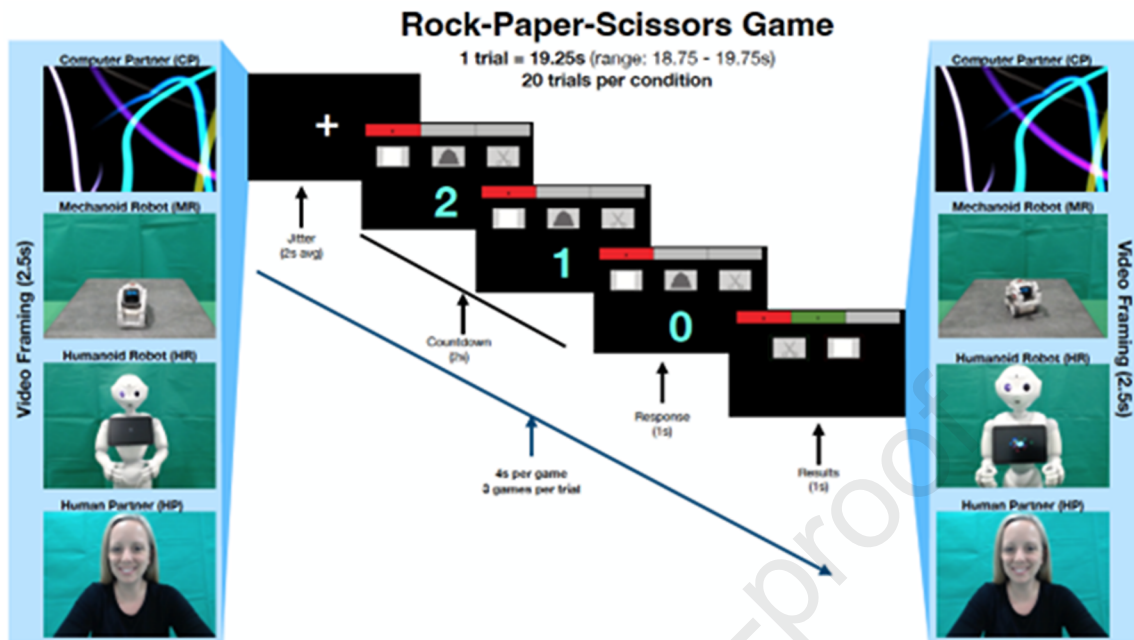






Journal Pre-proof





## Highlights

- The more human-like an agent, the more we engage the mentalizing network in the brain.
- Perceived socialness was even more influential in engaging the mentalizing network.
- Humans still hold a unique advantage over robots during social interactions.
- Implications for robotic design and the flexibility of human social cognition.

**Key resources table**

REAGENT or RESOURCE	SOURCE	IDENTIFIER
<b>Antibodies</b>		
<b>Bacterial and virus strains</b>		
<b>Biological samples</b>		
<b>Chemicals, peptides, and recombinant proteins</b>		
<b>Critical commercial assays</b>		
<b>Deposited data</b>		
fMRI data	Mendeley Data	DOI: 10.17632/2x9ykks2x x.1 DOI: 10.17632/693ty6chc d.1 DOI: 10.17632/c48324drr w.1
Group level whole brain results	This paper; Neurovault	<a href="https://identifiers.org/neurovault.collection:17268">https://identifiers.org/neurovault.collection:17268</a>
Behavioral data, stimuli, and additional analyses	This paper; OSF	<a href="https://osf.io/t4apv/">https://osf.io/t4apv/</a>
Preregistration	AsPredicted.org	<a href="https://aspredicted.org/CBG_ZPG">https://aspredicted.org/CBG_ZPG</a>

Experimental models: Cell lines		
Experimental models: Organisms/strains		
Oligonucleotides		
Recombinant DNA		
Software and algorithms		
MATLAB 2018a	MathWorks Inc	RRID:SCR_001622
Statistical Parametric Mapping 12 (SPM12)	<a href="https://www.fil.ion.ucl.ac.uk/spm/">https://www.fil.ion.ucl.ac.uk/spm/</a>	RRID:SCR_007037
Python 2.7	Python Software Foundation	RRID:SCR_008394
Python 3.5	Python Software Foundation	RRID:SCR_008394
Psychopy	<a href="https://www.psychopy.org/">https://www.psychopy.org/</a>	RRID:SCR_006571
Psychtoolbox-3 (PTB-3)	Psychophysics Toolbox	RRID:SCR_002881
R Studio	The R Foundation	RRID:SCR_001905
Code for robot introduction & main experiment	GitHub	<a href="https://github.com/chaudhuryB">https://github.com/chaudhuryB</a>
Other		