



HAL
open science

Brain activity during reciprocal social interaction investigated using conversational robots as control condition

Birgit Rauchbauer, Bruno Nazarian, Morgane Bourhis, Magalie Ochs, Laurent Prevot, Thierry Chaminade

► **To cite this version:**

Birgit Rauchbauer, Bruno Nazarian, Morgane Bourhis, Magalie Ochs, Laurent Prevot, et al.. Brain activity during reciprocal social interaction investigated using conversational robots as control condition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2019, 374 (1771), pp.20180033. 10.1098/rstb.2018.0033 . hal-02067722

HAL Id: hal-02067722

<https://hal.science/hal-02067722>

Submitted on 25 Feb 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1
2 **Brain activity during reciprocal social interaction**
3 **investigated using conversational robots as control condition**
4

5 Birgit Rauchbauer^{1, 2, 4}, Bruno Nazarian¹, Morgane Bourhis¹,
6 Magalie Ochs³, Laurent Prévot^{4, 5} & Thierry Chaminade¹
7

- 8 1. Institut de Neurosciences de la Timone, Aix-Marseille Université, CNRS, France
9 2. Laboratoire de Neurosciences Cognitive, Aix-Marseille Université, CNRS, France
10 3. Laboratoire d'Informatique et des Systèmes, Aix-Marseille Université, CNRS, France
11 4. Laboratoire Parole et Langage, Aix-Marseille Université, CNRS, France
12 5. Institut Universitaire de France, Paris France
13

14 **Abstract**

15 We present a novel functional magnetic resonance imaging (fMRI) paradigm for second-person
16 neuroscience. The paradigm compares a human social interaction (human-human interaction,
17 HHI) to an interaction with a conversational robot (human-robot interaction, HRI). The social
18 interaction consists of 1-minute blocks of live bidirectional discussion between the scanned
19 participant and the human or robot agent. A final sample of 21 participants is included in the corpus
20 comprising physiological (BOLD, respiration and peripheral blood flow) and behavioural
21 (recorded speech from all interlocutors, eye tracking from scanned participant, face recording of
22 the human and robot agents) data. Here we present the first analysis of this corpus, contrasting
23 neural activity between HHI and HRI. We hypothesized that, independently of differences in
24 behaviour between interactions with the human and robot agent, neural markers of mentalizing
25 (temporo-parietal junction and medial-prefrontal cortex) and social motivation (hypothalamus and
26 amygdala) would only be active in HHI. Results confirmed significantly increased response
27 associated with HHI in the temporo-parietal junction, hypothalamus and amygdala, but not in the
28 medial prefrontal cortex. Future analysis of this corpus will include fine-grained characterization
29 of verbal and non-verbal behaviours recorded during the interaction to investigate their neural
30 correlates.

31 **Introduction**

32 Humans' social bonds are established and maintained through interactions with others.
33 These interactions "are characterized by intricate reciprocal relations with the perception of
34 socially relevant information prompting (re-) actions, which are themselves processed and reacted
35 to" (Schilbach et al., 2013). To date, the field of social neuroscience, investigating the
36 neurophysiological basis of social interactions, has mostly focused on the investigation of *the*
37 *observation* of social signals, rather than on truly interactive social settings. In an attempt to
38 capture the interactional dynamics in real-life, "second person neuroscience" (Schilbach, 2015;
39 Schilbach et al., 2013) encourages the investigation of naturalistic interactive paradigms for
40 enhanced ecological validity (Pan & Hamilton, 2018). This approach aims to shift social
41 neuroscience from a prevailing "passive spectator science" (Hari, Henriksson, Malinen, &
42 Parkkonen, 2015) to an approach investigating the dynamics of social exchange (Hari et al., 2015).
43 This requires that not only the experimental, but also control condition should preserve the
44 reciprocity of real-time interactions.

45 Computer-animated on-screen agents have been used to study the influence of animacy on
46 motor imitation (Klapper, Ramsey, Wigboldus, & Cross, 2014) and mechanisms of joint attention
47 (Schilbach et al., 2006, 2010; Wilms et al., 2010), taking advantage of the extensive control the
48 experimenter can have on their behaviour. This includes for example control over the direction of
49 the gaze, towards or away from a target, or its timing. Robots also have been used as control
50 conditions in social neuroscience experiments (Chaminade et al., 2010, 2012; Chaminade, Da
51 Fonseca, Rosset, Cheng, & Deruelle, 2015; Krach et al., 2008). This article presents a novel
52 second-person neuroscience paradigm for functional magnetic resonance imaging (fMRI) that uses
53 a conversational robot as control condition for a human social interaction (Hale et al., 2018)..
54 Social interaction is operationalized using language, the most ubiquitous form of human
55 interaction. The paradigm allows recording brain activity during 1-minute live bidirectional
56 discussions between the scanned participant and a fellow human (human-human interaction; HHI)
57 and similar discussions between the same participant and a conversational robot (human-robot
58 interaction; HRI).

59 The HHI represents the experimental condition, constituting the "social" condition. The
60 HRI represents the control condition, which preserves sensorimotor aspects of live, bidirectional
61 conversation. Indeed, the robot has an anthropomorphic outer appearance, including a human face
62 and voice, so that seeing, hearing and talking to the artificial agent is similar to the interaction with
63 the human agent. In addition, while participants believe the robot is autonomous, it is actually

64 controlled by the same individual participants discuss with in the HHI conditions. As a
65 consequence, participants are not aware that they interact with the same individual, the
66 confederate, in both HHI and HRI conditions (see Figure 1, top). On the other hand, the
67 conversational robot used in the experiment is clearly not human: the face is projected on a
68 moulded plastic screen, it has a limited number of pre-scripted sentences for conversation and it
69 doesn't exhibit meaningful facial expressions or speech intonations.

70 A corpus of multimodal data is collected in addition to the fMRI data. Physiological
71 responses (respiration and peripheral blood flow pulse) are recorded synchronized with the MR
72 scanner and are used, currently, for modelling and removing physiological noise in the fMRI data.
73 Behavioural data is recorded to enable future exploration of brain-behaviour relations. This
74 includes speech production by the scanned participants and human and robot agent, the video
75 capture of the human and robot agent, and the gaze movement of the scanned participant. Given
76 the unconstrained nature of the conversation task, a fine-grained exploration of the behaviour, in
77 particular transcription and analyses of conversations, and exploration of dynamic gaze direction
78 to the human and robot agents' face, will be necessary to explore brain correlates in the corpus.
79 Here, we focus on the block analysis by contrasting conditions HHI and HRI. Given that the robot
80 control condition is designed to reproduce sensorimotor aspects of human conversation, both HHI
81 and HRI are expected to be associated with a neural network involved in visuomotor speech
82 perception and in speech production, including bilaterally the dorsal temporal lobes for speech
83 perception the ventral and lateral occipital cortex for face perception, as well as the bilateral ventral
84 primary motor cortex (speech motor control) and the left inferior frontal gyrus ("Broca's area")
85 for speech production (see Price, 2012 for review).

86 The contrast between conditions HHI and HRI is used to test specific hypotheses about the
87 neural correlates of social cognition, and hence confirming the quality and validity of the acquired
88 data. Social cognition (Adolphs, 1999) is broadly defined as "the sum of those processes that allow
89 individuals of the same species (conspecifics) to interact with one another." (Frith & Frith, 2007).
90 On the basis of previous work (Chaminade et al., 2010, 2012, 2015; Gallagher, Jack, Roepstorff,
91 & Frith, 2002; Krach et al., 2008), we specifically expected processes of mentalizing and enhanced
92 social motivation when interacting with the human compared to the robot. Mentalizing is the
93 ascription of mental states such as intentions and beliefs to explain the apparent behaviour of the
94 interaction partner (Frith & Frith, 1999). It requires the adoption of an intentional stance towards
95 the interaction partner – the assumption that the interacting agent actually has a mind supporting

96 mental states (Dennett, 1989). The adoption of an intentional stance towards a human versus a
97 computer interaction partner has been linked to activation in the paracingulate cortex (Gallagher
98 et al., 2002), a region of the medial prefrontal cortex (MPFC). It has been argued that humans do
99 not adopt an intentional stance towards robots, computers and more generally artificial agents
100 (Dennett, 1989). Indeed, increased activity in areas associated with mentalizing, not only in the
101 medial prefrontal cortex (MPFC), but also the temporoparietal junction (TPJ) has been repeatedly
102 found when interacting with a human compared to a robot or a computer (Chaminade et al., 2010,
103 2012, 2015; Gallagher et al., 2002; Krach et al., 2008). Such neural markers of mentalizing are
104 expected in the contrast HHI *versus* HRI.

105 Also on the basis of previous results in experiment contrasting human *versus* robot
106 interactions (e.g. Chaminade et al., 2012; Chaminade et al., 2015) we expected human interaction
107 to elicit activation of neural markers of social motivation, the human drive to interact, establish
108 and maintain bonds (Chevallier, Kohls, Troiani, Brodtkin, & Schultz, 2012). Chaminade and
109 colleagues (2015) report that a modulation of activity located in the paraventricular nucleus of the
110 hypothalamus by the social context (human *versus* robot) is present in neurotypical but not in
111 individuals diagnosed with autism spectrum disorder. This was associated with the proposal that
112 autism is associated with a deficit in social motivation, involving disrupted hypothalamic
113 regulation of oxytocin release (Chevallier et al., 2012). Consecutive works confirmed the
114 modulation of hypothalamus anatomy (Wolfe, Auzias, Deruelle, & Chaminade, 2015) and activity
115 (Wolfe, Deruelle, & Chaminade, 2018) by the social context. In general, social motivation and
116 reward have been associated with brain activation in the reward-circuit, comprising the ventral
117 striatum, orbitofrontal and ventromedial cortex (Chevallier et al., 2012), including amygdala
118 specifically for social reward (Rademacher et al., 2010). In line with these studies, we expected
119 that the interaction with a human would activate the previously reported subcortical areas (in
120 particular hypothalamus and amygdala) more than interaction with a robot. In contrast, we had no
121 specific hypothesis with regards to brain activity in the reverse contrast HRI *versus* HHI.

122 In the next sections we present the experimental paradigm, and the first results of the
123 reciprocal contrasts between conditions HHI and HRI, demonstrating not only the feasibility of
124 our approach, but also the scientific quality of the acquired data with regards to our hypotheses.

125

126 **Methods**

127 ***Participants***

128 Twenty-four native French-speaking participants (7 men) with an average age of 28.5
129 (SD=12.4) were fMRI scanned while having a conversation with a fellow human or a
130 retroprojected conversational robotic head (Furhat robotics, <https://www.furhatrobotics.com/>; Al
131 Moubayed, Beskow, Skantze, & Granström, 2012). Three participants were excluded due to
132 technical problems and insufficient task compliance. Twenty-one participants (mean age = 25.81,
133 SD = 7.49) were included in the analysis. Participants received information about the experiment,
134 confirmed their compatibility for MR scanning and gave their informed consent prior to scanning.
135 Eligibility entailed normal or corrected-to-normal vision and no history of psychiatric or
136 neurological conditions. Participants received a flat fee of 40 Euro for participation. The study was
137 approved by the ethics committee “Comité de Protection des Personnes Sud Méditerranée I”.

138

139 *Cover story for the experiment*

140 A recent behavioural study comparing human-human with human-robot conversations
141 (Chaminade, 2017) was adapted to the fMRI environment. The experimental factor was the nature
142 of the INTERACTING AGENT (HUMAN *versus* ROBOT), in a within-subject, block-design. A cover
143 story was a fundamental element of the study, as it provided a fake rationale for the experiment as
144 well as a frame for discussion and explanations for the experimental set-up. Volunteers were told
145 they participated in a neuromarketing experiment sponsored by an advertising company. The
146 company wanted to test whether the message of their forthcoming advertisement campaign can be
147 identified if a pair people are presented the images of the campaign and discuss about them. Two
148 series of three images presented anthropomorphized fruits and vegetables as superheroes or
149 appearing rotten respectively (see Supplementary Material, Figure S1). Participants were
150 instructed to talk freely about the presented image with the agent outside the scanner, either a
151 human or the conversational robotic head (controlled by the confederate, unbeknown to the
152 participant; see next section). The robot was a presented as an autonomous conversational agent
153 that had information about the advertisement campaign. As such, the discussion with the robot
154 could be used to gather information about the advertisement campaign.

155 In practice, the cover story was presented to the participants by experimenter BR in the
156 lobby of the MR centre, later joined by the confederate. Confederates were gender-matched to
157 participants. Experimenter TC served as confederate for men and experimenter MB for women.
158 The participant was told that the confederate had already participated in the experiment inside the
159 scanner and had agreed to come back to play the role of the agent outside of the scanner. The

160 participant was then accompanied into the control room outside the scanner and shown the robot
161 (see next section). In the meantime, we asked the confederate to wait, telling him/her that we would
162 first get the participant ready into the scanner. At the end of the experiment, participants were
163 debriefed verbally to verify they still believed in the cover story and we revealed the true objective
164 of the experiment.

165

166 *Artificial agent*

167 The robotic head from Furhat robotics (<https://www.furhatrobotics.com/>; Al Moubayed,
168 Beskow, Skantze, & Granström, 2012) was used in this study. The robotic head is a semi-
169 transparent plastic mask moulded to mirror the shape of a human face on which the image of a
170 human face is retro projected. In order to match the robot appearance to the confederates, the face
171 and voice were gender-matched, a wig, a scarf and headphones were added as well as glasses for
172 confederate TC (see illustrations in Figure 1). Furhat OS allowed us to control its responses
173 through a Wizard of Oz (WOZ): unbeknown to the participant, the confederate was controlling the
174 robot remotely. The robot conversational feedbacks were largely based on actual human
175 interactions recorded during the previous the behavioural study (Chaminade, 2017). A WOZ user
176 interface was created with Furhat OS displaying buttons on a web browser running on a tablet
177 allowing the human controller to launch pre-programmed conversational feedback. For example,
178 clicking the button “yes” on the screen would make the robot say “yes”, clicking the button
179 “superhero” would launch the sentence “It looks like a superhero”. Conversational elements
180 included non-specific feedbacks, such as “yes”, “no”, or “maybe”, which could be used for all
181 images, as well as specific feedbacks for each of the images, such as “This lemon looks like a
182 superhero” or “Maybe this is a campaign to eat healthier food”. Note that the cover story allowed
183 to limit the number of targeted conversations for each image compared to unconstrained
184 discussion. Overall about 30 French conversational feedbacks were scripted for the robot for each
185 of the six images (see Supplementary Material, file S1 for robot statements).

186 The robot was controlled using this WOZ interface by the confederate acting as
187 conversational agent in the human condition, allowing for a realistic bidirectional conversation
188 similar to the interaction with the human. Thus, unbeknown to the participants, they discussed with
189 the same agent in both human and robot conditions. On the other hand, while the conversational
190 robot was able to reproduce superficial aspects of a human conversation, it lacked intonations in
191 speech, head movements, facial expressions and the ability to elaborate longer statements, thus

192 appearing clearly artificial and participants believed, according to debriefing, that it was
193 autonomous.

194

195 ***Experimental set-up***

196 The fMRI audio set-up allowed live conversation between the scanned participant lying
197 supine in the scanner and the agent outside of the scanner despite the noisy MRI environment. It
198 consisted of an active noise-cancelling MR compatible microphone (FORMI-III+ from
199 optoacoustics mounted on the head coil) and insert earphones from Sensimetrics. Live video of
200 the interacting agent (human or robot) was captured by webcams and projected to a mirror mounted
201 on the antenna in front of the scanned participant's eyes. Videos were recorded for future analysis.
202 Participants' direction of gaze on the projection mirror was recorded (Eyelink 1000 system, SR
203 Research). Stimulus presentation, audio and video routing and recording, synchronization with the
204 fMRI acquisition triggers and the eye tracker was implemented in a *Labview* (National Instrument)
205 virtual machine (see Figure 1, top). Finally, blood pulse and respiration were recorded with built-
206 in Prisma Siemens hardware and data format.

207 Altogether, we collected multimodal data including behaviour (speech from the participant
208 and human or robot agent, video capture of the human and robot agent, and the gaze movement of
209 the scanned participant) and physiology (BOLD signal, respiration and peripheral blood flow
210 pulse) to form a corpus. Transcribed speech data (more details on the transcription and an example
211 of the conversation is provided as Supplementary Material in section 2.1 and Files S3-S5) and
212 fMRI data, both raw and analysed, will be shared in online repositories.

213

214 ***Experimental paradigm***

215 The MRI recordings consisted of four sessions of each six 1-minute blocks of conversation
216 each, showing the “super-heroes” images in the first and third sessions and the “rotten fruits”
217 images in the second and fourth sessions (see Supplementary Material, Figure S1 and Table S1 for
218 details). The order was kept constant across participants each session alternating the three images
219 per session and two INTERACTING AGENTS (complete order of conditions is given in Supplementary
220 Material Table S1). Each image was thus shown twice in each session, once per INTERACTING
221 AGENT. Given the entertaining nature of the interaction, we did not expect habituation effects to
222 affect the brain imaging data and preferred to have the nature of the agent fully predictable. Hence,
223 we did not randomize the order of presentation of the human and robot agents.

256 realigned using a sinc interpolation algorithm that estimates rigid body transformations
257 (translations, rotations). Images were then spatially smoothed using an isotropic 5 mm full-width-
258 at-half-maximum Gaussian kernel. The first realigned and unwarped functional image was
259 coregistered with an unwarped single-band reference image recorded at the onset of each trial,
260 which was itself coregistered with the T1 and T2 anatomical images. These anatomical images
261 were segmented into grey matter (GM), white matter (WM), and cerebral spinal fluid (CSF) using
262 SPM12 “New segment”. GM, WM, and CSF tissue probability maps from our sample of 21
263 included participants were used to form a DARTEL template (Ashburner, 2007). The deformation
264 flow fields from individual spaces to this template were used to normalize the beta images resulting
265 from the individual subjects’ analyses (i.e. in subjects’ individual space) for use in a random-effect
266 second-level analysis.

267 Potential artefacts from blood pulse and respiration were controlled using the Translational
268 Algorithms for Psychiatry-Advancing Science (TAPAS) toolbox standard procedure
269 (<https://www.tnu.ethz.ch/de/software/tapas/documentations/physio-toolbox.html>; Kasper et al.,
270 2017). Realignment parameters (translation and rotation) as well as their derivatives and the square
271 product of both parameters and their derivatives were used as covariates to control for movement-
272 related artefacts. We also used the Artefact Detection Tools (ART) to control for any movement-
273 related artefacts (www.nitrc.org/projects/artifact_detect/) using the standard threshold of 2 mm.

274 The fMRI time series were analysed using the General Linear Model (GLM) approach
275 implemented in SPM. Single-subject models consisted of one regressor representing the one-
276 minute discussion for each of the two INTERACTING AGENTS, and another one representing the
277 presentation of the images.

278 After normalization, beta estimates images were entered in a mixed-model analysis of
279 variance (using SPM “full ANOVA”) with participants and sessions as random factors and the
280 nature of the INTERACTING AGENT as factor of interest for inferences at the population level. A
281 mask was created on the basis of the mean of DARTEL normalized anatomical GM and WM tissue
282 classes of each participant, also used for rendering results in Figures 2 & 3.

283 We first assessed the main effect of the conversation with both agents against the implicit
284 baseline. We then looked specifically at the effects of each of the INTERACTING AGENT contrasted
285 to the other one, with a clear focus on brain areas involved in mentalizing and social motivation in
286 the contrast HHI *versus* HRI.

287 All statistical inference was performed applying a threshold of $p = 0.05$ False-Discovery
288 Rate (FDR) corrected for the whole brain at the cluster-level (Friston, Holmes, Poline, Price, &
289 Frith, 1996). Anatomical localization of the resulting clusters relied on the projection of the results
290 onto the mean anatomical image of our pool participants resulting from DARTEL coregistration.

291

292 **Results**

293 *Cover story debriefing*

294 A verbal debriefing was performed in an undirected and open format, to allow the
295 participants to report their experience in an unbiased manner. None of the participants reported
296 feelings of distress during the experiment with either interaction partner, or doubts about the
297 autonomous nature of the conversational robot. In conclusion, all participants still believed in the
298 cover-story at the end of the recordings.

299

300 *Assessment of participants' movements during scanning*

301 No participant was excluded on the basis of the assessment of movement using the toolbox
302 ART (https://www.nitrc.org/projects/artifact_detect/). At the movement threshold used, between
303 0 and a maximum of 3 volumes per session and participant were considered as outliers. In the
304 absence of large artefacts, all scans and sessions from the twenty-one participants were included
305 in the analysis. Moreover, using the same metric to calculate a global movement per block of
306 discussion, session and subject, an analysis of variance showed no effect of the INTERACTING
307 AGENTS ($F(1,495)=2.22$, $p = 0.14$; see Supplementary Figure S2).

308

309 *Participants' behaviour*

310 The full transcription of the 504 minutes of discussion collected for the corpus is ongoing
311 (examples, as well as the link to the data repository are presented in supplementary material; see
312 Files S3-S5). Yet it has been observed by the confederates that discussions between the two agents
313 differed in terms of the speed and emotion conveyed by participant's voice. Participants spoke in
314 general faster and with increased prosodic variations with the human than the robot agent. Humour
315 was also observed in the conversation with the human, but not with the robot. These observations
316 are expected given the differences in conversational competence between the two agents.

317

318 *fMRI results*

319 The main effect of conversation for the human and for the robot largely overlapped (Figure
320 2, top). As predicted given the nature of the task, common activation clusters are found bilaterally
321 along the superior temporal sulcus and gyrus, the central operculum, the lateral and ventral
322 occipital cortex, the lateral premotor cortex, the supplementary motor area and the ventral and
323 dorsal cerebellum, as well as the left inferior frontal gyrus. Differences between the resulting
324 activation maps for the human and robot agents were quantitative rather than qualitative, with
325 larger clusters mostly related to motor control (in region of the precentral and postcentral gyri) for
326 the robot and to speech processing (in the temporal cortex) for the human.

327 The contrast HHI *versus* HRI (see Figure 2, bottom left) revealed bilateral activation in the
328 superior temporal gyrus and sulcus that overlapped partly with the temporal areas associated with
329 the main effects of the conversations. It extended anteriorly to the temporal poles and to the
330 posterior lateral orbitofrontal cortex. Posteriorly, it covered the temporo-parietal junction and
331 lateral occipital cortex. Another significant cluster covered a number of subcortical structures: the
332 bilateral thalamus, hypothalamus, hippocampus, amygdala, caudate nucleus and the subthalamic
333 area. We also found bilateral activation in the cerebellum centred on the horizontal fissure. No
334 medial prefrontal cluster was found at the threshold used.

335 The reverse contrast HRI *versus* HHI identified a number of bilateral activation clusters.
336 In the occipital region, a cluster centred on the striate cortex extended to the lingual and fusiform
337 gyri. Furthermore, a strong activation was found bilaterally within the intraparietal sulcus
338 extending to the supramarginal gyrus. Clusters were also found in the middle frontal gyrus and the
339 centred on the lateral central sulcus.

340

341 - Figure 2 and 3 around here -

342 **Discussion**

343 We introduce a novel paradigm to investigate the neural bases of natural interactions
344 between humans, in line with a second-person neuroscience approach. We choose live
345 bidirectional conversation as operationalization of natural interactions given it is the most common
346 form of communication between humans. The scientific challenge is twofold.

347 Methodologically, investigating natural interactions implies that the classical experimental
348 approach, in which only one parameter is changed between experimental conditions, isn't
349 applicable. Here, we use a robot for a high-level control condition: the conversational robot
350 reproduces a number of sensorimotor aspects of the conversation, yet is far from mimicking a real

351 human, and it does not elicit the adoption of an intentional stance according to Dennett (1989).
352 Hence, interacting with the robot in the current paradigm can be considered to be non social,
353 yielding a unique control condition for the social interaction with a fellow human.

354 Technically, the constraints of MRI recordings are numerous for a live bidirectional
355 conversation during fMRI scanning: participants lie supine in a very noisy environment and are
356 required to avoid any movement to ensure the quality of the data. We decided to hold the head
357 firmly using foam pads while keeping the jaw free. Importantly, post-hoc assessment of individual
358 participants' movements showed very limited motion and no quantitative difference between the
359 human and robot condition, confirming the feasibility of the task.

360 The main objective of the analysis presented in the present article is to evaluate the quality
361 of the recorded fMRI data, the main part of a unique corpus of neural, physiological and
362 behavioural data. We have strong hypotheses about brain responses expected to be common during
363 conversation with the two agents, as well as for the difference between interaction with the human
364 *versus* the robot.

365
366 *Commonly activated areas.* We report a large number of common activated areas in the main
367 effects of HHI and HRI that can be directly related to sensorimotor aspects of the conversation. As
368 expected, they cover the dorsal half of the posterior temporal cortex bilaterally, known as the main
369 brain region for auditory speech perception, comprising functional areas such as the primary
370 auditory cortex or temporal voice areas (Belin, Fecteau, & Bedard, 2004; Belin, Zatorre, Lafaille,
371 Ahad, & Pike, 2000). Common activations are also found in motor-related areas which are
372 involved in the motor aspects of speech production. In particular, the ventral and opercular region
373 below the central sulcus and adjacent precentral and postcentral gyri is likely to include primary
374 motor and sensory regions involved in verbalization (e.g., Price, 2012), while the lateral cluster in
375 the central sulcus area maps into the sensorimotor representation of the larynx (Brown et al., 2009).
376 The lateralized inferior frontal gyrus corresponds to Broca's area, crucial for the production of
377 speech. The medial premotor areas and the cerebellum are generally associated with the timing of
378 action, which is crucial for articulation (see for review Price, 2012). Note that these motor areas
379 could also be involved in speech perception according to the motor theory of speech
380 perception (Galantucci, Fowler, & Turvey, 2006). Indeed, a recent study revealed correlated
381 activation in the temporal auditory areas and the inferior frontal gyrus during successful coupling
382 between a speaker and a listener during a delayed interaction (Stephens, Silbert, & Hasson, 2010).

383 Current results show that live bidirectional conversation, irrespective of the agent, activates a
384 network of brain regions previously associated with speech perception and production.
385 Unfortunately, speech production and perception can't be distinguished in the current analysis, but
386 will be the object of future exploration of this corpus. Finally, the large cluster spanning the lateral
387 and ventral occipital cortex most likely responds to the processing of visual information, namely
388 the face of the human or robot agent talking.

389
390 *Increased activity areas in HHI condition.* Interaction with a fellow human as compared to the
391 robot, revealed activation in the temporal cortex, including the bilateral temporoparietal junction
392 (TPJ), and subcortical activation in the hypothalamus, the thalamus, the hippocampus, the
393 amygdala and the subthalamic area. The results are in line with our predictions, except for the
394 absence of activation in the anterior medial frontal cortex. Activation in the TPJ and hypothalamus
395 has been reported in previous studies comparing human to robot interaction (Chaminade et al.,
396 2012, 2015; Krach et al., 2008). TPJ activation has recently been reported when explicitly
397 ascribing human intention to robot behavior (Özdem et al., 2017). Hypothalamus activation during
398 HHI *versus* HRI was linked to enhanced social motivation (Chaminade et al., 2015; Chevallier et
399 al., 2012), given the release of oxytocin by hypothalamus subnuclei (Bartz et al., 2011; Heinrichs,
400 von Dawans, & Domes, 2009). The amygdala has specifically been related to social as compared
401 to monetary reward (Rademacher et al., 2010). It is a key neural node in the processing of
402 emotionally and socially relevant information, coding saliency, reward and value of social stimuli
403 (Adolphs, 2010).

404
405 *Increased activity areas in HRI condition.* The contrast HRI *versus* HHI showed significant
406 activation in visual areas, including the fusiform cortex, hosting the fusiform face area, in the
407 intraparietal sulcus and in anterior parts of the middle frontal gyrus. We did not have specific
408 hypotheses for the effect of HRI compared to HHI, so all interpretation remains speculative. The
409 finding of enhanced activity in an area prominently involved in human face perception (FFA) has
410 been previously reported for action observation comparing robotic to human movements (Cross,
411 Ramsey, Liepelt, Prinz, & Hamilton, 2016), and, alongside enhanced activation in visual areas, for
412 the perception of robot compared to human faces (Chaminade et al., 2010). This has been
413 interpreted as additional visual processing effort to identify an unfamiliar, robotic face (Chaminade
414 et al., 2010). Interestingly, the intraparietal sulcus was associated with the "uncanny valley" effect

415 (Saygin, Chaminade, & Ishiguro, 2010), and interpreted as reflecting an increase of attention
416 towards unfamiliar stimuli. Enhanced response in visual areas, the IPS and the MFG seems in line
417 with studies investigating mechanistic versus social reasoning (Jack, Dawson, Begany, et al., 2013)
418 and ratings of images depicting machines (including robots) versus humans (Jack, Dawson, &
419 Norr, 2013).

420 Overall, we largely confirmed our hypotheses for brain activation in response to human
421 compared to robot conversation. They support that processes of mentalizing and social motivation
422 are enhanced in our paradigm when interacting with a human rather than with a robot. These results
423 further confirm the quality and validity of the brain imaging data recorded, the main part of corpus
424 also including behavioural and physiological data collected with the approach presented in this
425 paper.

426

427 **Limitations**

428 We present an approach towards truly reciprocal, interactive social neuroscience and first
429 supporting neurophysiological results. One major concern in fMRI studies involving language is
430 the risk of extensive movement artefacts induced by motor-related aspects of speech-production.
431 Yet, in the present study, we observed hardly any speech-induced movement during recording (see
432 Supplementary Figure S2).

433 The pre-scripted sentences of the robot were shorter and more limited than the human's.
434 The robot intonation, and more generally of head and face movements, were not controlled in the
435 current experiment. Thus, it is expected that the human conversation differed from the robot
436 conversation. This is likely to explain some of the differences in brain activity reported here.

437 The univariate fMRI analysis presented here is not sufficient to investigate the complex
438 dynamics of the interactions. The corpus collected contained not only fMRI but also behavioural
439 (linguistic, eye-tracking of the participant, video of the other agent during the interaction) and
440 physiological (respiration and blood pulse) data. Future work on the corpus will entail fine-grained
441 description of the behaviour, that will fuel the analysis of fMRI data. Transcription of speech
442 recordings is under way (see supplementary information for an example of transcription) and will
443 be made publicly available together with the fMRI data.

444 Also, future studies should include explicit measures of the perception of robots in general,
445 and of the conversational robot used in the experiment more specifically, in the form of

446 questionnaires that would provide insights about individuals' variations in their expectations about
447 the robot's capacity.

448

449 **Conclusion**

450 We investigated natural interaction comparing Human-Human Interaction (HHI) and
451 Human-Robot Interaction (HRI) using fMRI. Using a conversational robot as control condition
452 allowed to preserve reciprocal dynamics during interaction. Results for HHI showed activity in
453 brain areas associated with mentalizing and social motivation. The article introduces an innovative
454 paradigm in a second-person neuroscience approach. As such, it could be used as a starting point
455 for social neuroscience to investigate specificities of human social cognition as well as to quantify,
456 and thus participate in the improvement of, the social competence of robots interacting with
457 humans.

458

459 **Acknowledgement**

460 Research supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036 (BLRI) and
461 AAP-ID-17-46-170301-11.1 by the Excellence Initiative of Aix-Marseille University
462 (A*MIDEX), a French "Investissement d'Avenir" programme. BR is supported by the Fondation
463 pour la Recherche Médicale (FRM, SPF20171039127).

References

- Adolphs, R. (1999). Social cognition and the human brain. *Trends in Cognitive Sciences*, 3(12), 469–479.
- Adolphs, R. (2010). What does the amygdala contribute to social cognition? *Annals of the New York Academy of Sciences*, 1191(1), 42–61.
- Al Moubayed, S., Beskow, J., Skantze, G., & Granström, B. (2012). Furhat: a back-projected human-like robot head for multiparty human-machine interaction. In *Cognitive behavioural systems* (pp. 114–130). Springer.
- Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage*, 38(1), 95–113.
- Bartz, J. A., Zaki, J., Bolger, N., & Ochsner, K. N. (2011). Social effects of oxytocin in humans: context and person matter. *Trends in Cognitive Sciences*, 15(7), 301–309.
- Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: neural correlates of voice perception. *Trends in Cognitive Sciences*, 8(3), 129–135.
- Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, 403(6767), 309.
- Brown, S., Laird, A. R., Pfordresher, P. Q., Thelen, S. M., Turkeltaub, P., & Liotti, M. (2009). The somatotopy of speech: Phonation and articulation in the human motor cortex. *Brain and Cognition*, 70(1), 31–41.
- Chaminade, T. (2017). An experimental approach to study the physiology of natural social interactions. *Interaction Studies*, 18(2).
- Chaminade, T., Da Fonseca, D., Rosset, D., Cheng, G., & Deruelle, C. (2015). Atypical modulation of hypothalamic activity by social context in ASD. *Research in Autism Spectrum Disorders*, 10, 41–50.
- Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutscher, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience*. Retrieved from <http://journal.frontiersin.org/article/10.3389/fnhum.2012.00103>
- Chaminade, T., Zecca, M., Blakemore, S.-J., Takanishi, A., Frith, C. D., Micera, S., ... Umiltà, M. A. (2010). Brain response to a humanoid robot in areas implicated in the perception of human emotional gestures. *PLoS One*, 5(7), e11577.
- Chevallier, C., Kohls, G., Troiani, V., Brodtkin, E. S., & Schultz, R. T. (2012). The social

- motivation theory of autism. *Trends in Cognitive Sciences*, 16(4), 231–239.
- Cross, E. S., Ramsey, R., Liepelt, R., Prinz, W., & Hamilton, A. F. de C. (2016). The shaping of social perception by stimulus and knowledge cues to human animacy. *Phil. Trans. R. Soc. B*, 371(1686), 20150075.
- Dennett, D. C. (1989). *The intentional stance*. MIT press.
- Depue, R. A., & Morrone-Strupinsky, J. V. (2005). A neurobehavioral model of affiliative bonding: Implications for conceptualizing a human trait of affiliation. *Behavioral and Brain Sciences*, 28(3), 313–+.
- Friston, K. J., Holmes, A., Poline, J.-B., Price, C. J., & Frith, C. D. (1996). Detecting activations in PET and fMRI: levels of inference and power. *Neuroimage*, 4(3), 223–235.
- Frith, C. D., & Frith, U. (1999). Interacting minds--a biological basis. *Science*, 286(5445), 1692–1695.
- Frith, C. D., & Frith, U. (2007). Social cognition in humans. *Current Biology*, 17(16), R724–R732.
- Galantucci, B., Fowler, C. A., & Turvey, M. T. (2006). The motor theory of speech perception reviewed. *Psychonomic Bulletin & Review*, 13(3), 361–377.
- Gallagher, H. L., Jack, A. I., Roepstorff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *Neuroimage*, 16(3), 814–821.
- Hari, R., Henriksson, L., Malinen, S., & Parkkonen, L. (2015). Centrality of social interaction in human brain function. *Neuron*, 88(1), 181–193.
- Heinrichs, M., von Dawans, B., & Domes, G. (2009). Oxytocin, vasopressin, and human social behavior. *Frontiers in Neuroendocrinology*, 30(4), 548–557.
- Jack, A. I., Dawson, A. J., Begany, K. L., Leckie, R. L., Barry, K. P., Ciccio, A. H., & Snyder, A. Z. (2013). fMRI reveals reciprocal inhibition between social and physical cognitive domains. *NeuroImage*, 66, 385–401.
- Jack, A. I., Dawson, A. J., & Norr, M. E. (2013). Seeing human: Distinct and overlapping neural signatures associated with two forms of dehumanization. *Neuroimage*, 79, 313–328.
- Klapper, A., Ramsey, R., Wigboldus, D., & Cross, E. S. (2014). The control of automatic imitation based on Bottom–Up and Top–Down cues to animacy: Insights from brain and behavior. *Journal of Cognitive Neuroscience*.
- Kasper, L., Bollmann, S., Diaconescu, A. O., Hutton, C., Heinzle, J., Iglesias, S., ... Stephan, K. E. (2017). The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data. *Journal of*

Neuroscience Methods, 276, 56–72.

- Krach, S., Hegel, F., Wrede, B., Sagerer, G., Binkofski, F., & Kircher, T. (2008). Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PloS One*, 3(7), e2597.
- Özdem, C., Wiese, E., Wykowska, A., Müller, H., Brass, M., & Van Overwalle, F. (2017). Believing androids—fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents. *Social Neuroscience*, 12(5), 582–593.
- Pan, X. and Hamilton, A. F. (2018), Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*, 109: 395-417.
- Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *Neuroimage*, 62(2), 816–847.
- Rademacher, L., Krach, S., Kohls, G., Irmak, A., Gründer, G., & Spreckelmeyer, K. N. (2010). Dissociation of neural networks for anticipation and consumption of monetary and social rewards. *Neuroimage*, 49(4), 3276–3285.
- Saygin, A. P., Chaminade, T., & Ishiguro, H. (2010). The perception of humans and robots: Uncanny hills in parietal cortex. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32).
- Schilbach, L. (2015). Eye to eye, face to face and brain to brain: Novel approaches to study the behavioral dynamics and neural mechanisms of social interactions. *Current Opinion in Behavioral Sciences*. <https://doi.org/10.1016/j.cobeha.2015.03.006>
- Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *Behavioral and Brain Sciences*, 36(04), 393–414.
- Schilbach, L., Wilms, M., Eickhoff, S. B., Romanzetti, S., Tepest, R., Bente, G., ... Vogeley, K. (2010). Minds made for sharing: initiating joint attention recruits reward-related neurocircuitry. *Journal of Cognitive Neuroscience*, 22(12), 2702–2715.
- Schilbach, L., Wohlschlaeger, A. M., Kraemer, N. C., Newen, A., Shah, N. J., Fink, G. R., & Vogeley, K. (2006). Being with virtual others: Neural correlates of social interaction. *Neuropsychologia*, 44(5), 718–730.
- Stephens, G. J., Silbert, L. J., & Hasson, U. (2010). Speaker–listener neural coupling underlies successful communication. *Proceedings of the National Academy of Sciences*, 107(32),

14425–14430.

Wilms, M., Schilbach, L., Pfeiffer, U., Bente, G., Fink, G. R., & Vogeley, K. (2010). It's in your eyes—using gaze-contingent stimuli to create truly interactive paradigms for social cognitive and affective neuroscience. *Social Cognitive and Affective Neuroscience*, *5*(1), 98–107.

Wolfe, F. H., Auzias, G., Deruelle, C., & Chaminade, T. (2015). Focal atrophy of the hypothalamus associated with third ventricle enlargement in autism spectrum disorder. *NeuroReport*, *26*(17), 1017–1022.

Wolfe, F. H., Deruelle, C., & Chaminade, T. (2018). Are friends really the family we choose? Local variations of hypothalamus activity when viewing personally known faces. *Social Neuroscience*, *13*(3), 289–300.

Figure 1:

Experimental design showing (a) the communication between the scanned participant and the other conversation agent, either the confederate or the robot, as well as the recording modalities; (b) the timeline of the experiment showing the alternation between the stimuli and conversation periods, as well as the relative timing. The fruit pictures correspond to the images used in the cover story, while the robot and confederate pictures illustrates episodes of live bidirectional conversations.

Figure 2:

Render of the brain surface of the mean of the coregistered and normalized brains from our participants sample. Overlaid are the results of the contrasts of interest ($p < 0.05$ FDR-corrected at the cluster level). Upper row shows the contrast of the human-human interaction (HHI) *versus* baseline in blue, and of the human-robot interaction (HRI) *versus* baseline in red. Lower row shows the contrast HHI *versus* HRI in blue, and HRI *versus* HHI in red.

Figure 3:

Coronal (top images), sagittal (middle images) and axial (bottom images) sections focusing on the cluster identifying subcortical structures significantly activated in HHI *versus* HRI.

Figure 1.

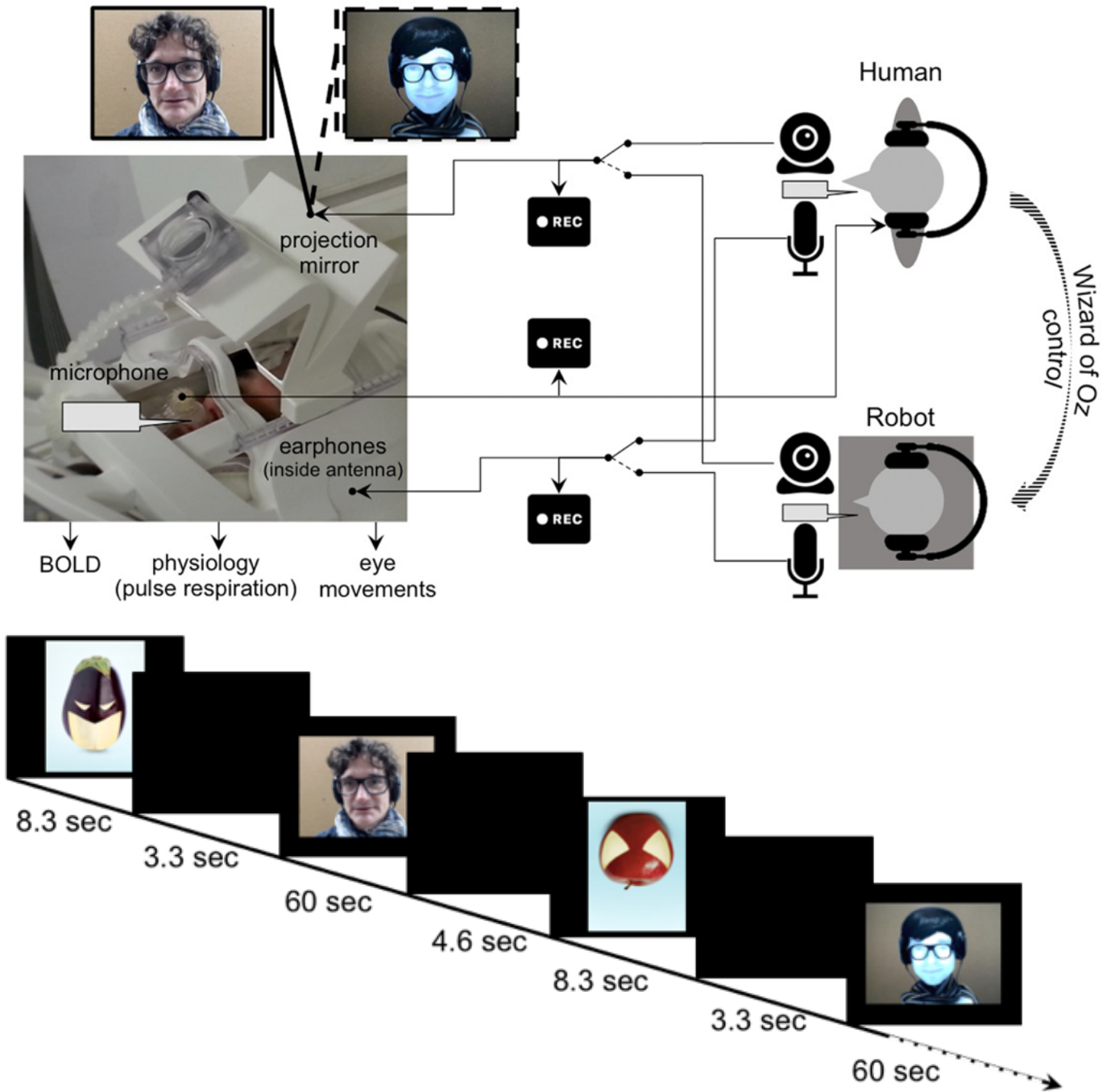


Figure 2.

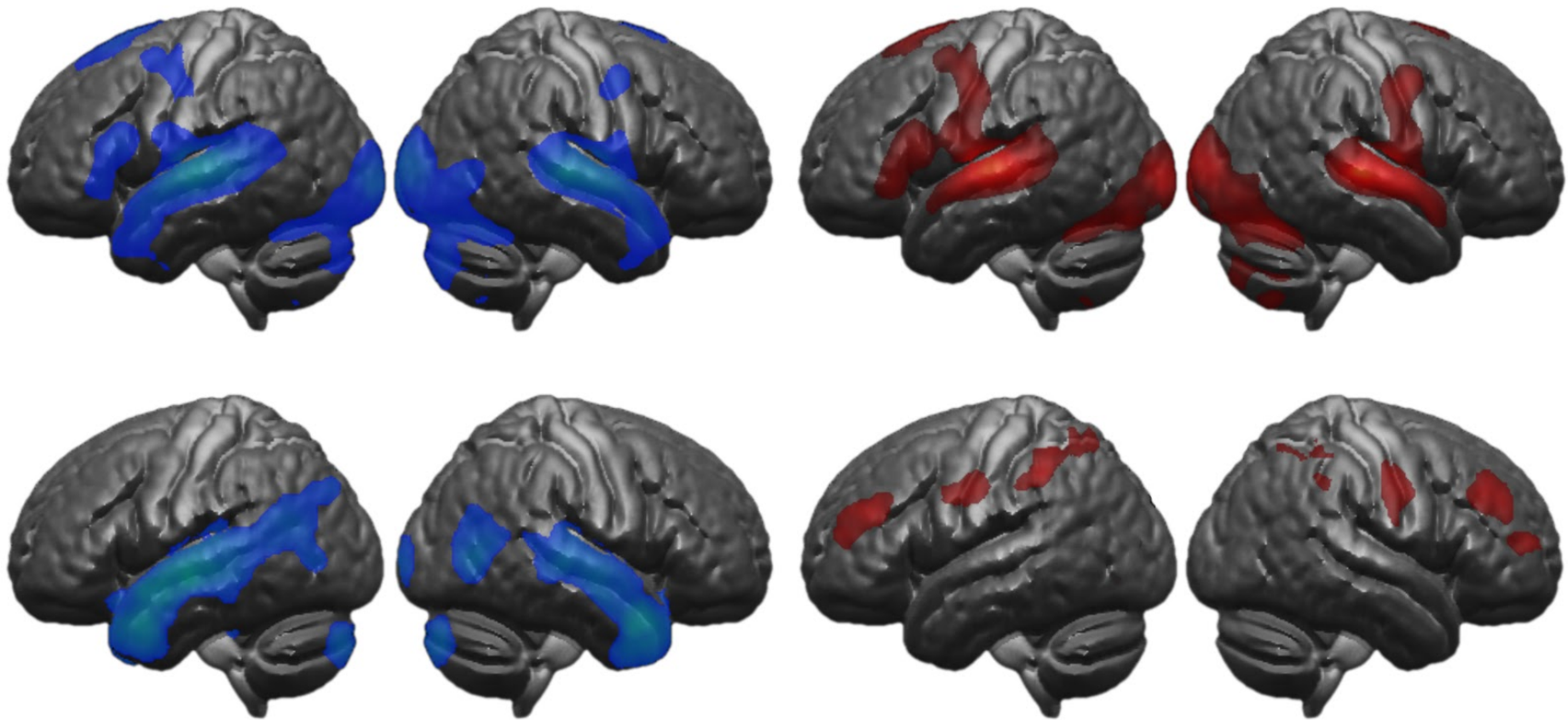


Figure 3.

