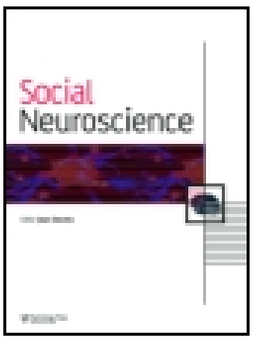

Social Neuroscience

ISSN: 1747-0919 (Print) 1747-0927 (Online) Journal homepage: http://www.tandfonline.com/loi/psns20

Believing androids – fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents

Ceylan Özdem, Eva Wiese, Agnieszka Wykowska, Hermann Müller, Marcel Brass & Frank Van Overwalle

To cite this article: Ceylan Özdem, Eva Wiese, Agnieszka Wykowska, Hermann Müller, Marcel Brass & Frank Van Overwalle (2016): Believing androids – fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents, Social Neuroscience, DOI: 10.1080/17470919.2016.1207702

To link to this article: http://dx.doi.org/10.1080/17470919.2016.1207702

Accepted author version posted online: 08 Jul 2016.
Published online: 02 Sep 2016.

SOCIAL NEUROSCIENCE, 2016
http://dx.doi.org/10.1080/17470919.2016.1207702

Routledge
Taylor & Francis Group

Believing androids – fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents

Ceylan Özdem[a*], Eva Wiese[b,c*], Agnieszka Wykowska[d,e], Hermann Müller[c], Marcel Brass[f] and Frank Van Overwalle [a]

[a]Department of Psychology, Vrije Universiteit Brussels, Brussels, Belgium; [b]Department of Psychology, George Mason University, Fairfax, VA, USA; [c]Department of Psychology, Ludwig Maximilians-Universiteit, Munchen, Germany; [d]Engineering Psychology, Division of Human Work Sciences, Luleå University of Technology, Luleå, Sweden; [e]Chair for Cognitive Systems, Technische Universität München, Munich, Germany; [f]Ghent Institute for Functional and Metabolic Imaging, University of Ghent, Ghent, Belgium

ABSTRACT

Attributing mind to interaction partners has been shown to increase the social relevance we ascribe to others' actions and to modulate the amount of attention dedicated to them. However, it remains unclear how the relationship between higher-order mind attribution and lower-level attention processes is established in the brain. In this neuroimaging study, participants saw images of an anthropomorphic robot that moved its eyes left- or rightwards to signal the appearance of an upcoming stimulus in the same (valid cue) or opposite location (invalid cue). Independently, participants' beliefs about the intentionality underlying the observed eye movements were manipulated by describing the eye movements as under human control or preprogrammed. As expected, we observed a validity effect behaviorally and neurologically (increased response times and activation in the invalid vs. valid condition). More importantly, we observed that this effect was more pronounced for the condition in which the robot's behavior was believed to be controlled by a human, as opposed to be preprogrammed. This interaction effect between cue validity and belief was, however, only found at the neural level and was manifested as a significant increase of activation in bilateral anterior temporoparietal junction.

ARTICLE HISTORY
Received 16 December 2015
Revised 20 April 2016
Published online
5 September 2016

KEYWORDS
Attentional reorientating; social mentalizing; intentional stance; TPJ

Introduction

Recent advancements in technology enlarge the human social sphere to encompass not only other human agents but also artificial entities. A rapidly growing branch of robotics – social robotics – aims at designing robots that allow for humans to communicate and interact with the robotic agents in an intuitive and socially engaging manner and so could ultimately be introduced into humans' daily lives. To achieve this, understanding how humans interact with each other, what signals are used for communication, and what sort of knowledge is needed to understand and predict the behavior of others is of crucial importance. Research in social cognitive neuroscience has demonstrated that when we interact with others we often make inferences about the others' internal states (i.e., intentions, beliefs) in order to explain, understand, and predict their behavior – a process commonly referred to as *mentalizing* (Baron-Cohen, 1995; Frith & Frith, 2006). Such inferences may be made automatically, without deliberate mental effort, or with conscious intent (Ma et al., 2012). Mentalizing is crucial for successful social interactions (Waytz et al., 2010), and designing robots that trigger mentalizing in human interaction partners will therefore, arguably, strongly modulate human–robot interaction.

Consequences of adopting the intentional stance toward others

Attribution of mind to an agent is a prerequisite for inferring their mental states. Mind and mental states are typically attributed to other humans by default, whereas machines are not believed – at least in the Western culture – to possess true mental states. Rather, they are perceived as physical entities with preprogrammed behaviors. Beliefs about the internal states of other agents are based on experience and preexisting knowledge humans have acquired about the world (Frith & Frith, 2006). The belief that an agent has a mind intuitively or consciously triggers the adoption of the *intentional stance* (Dennett, 2003), which involves

CONTACT Frank Van Overwalle Frank.VanOverwalle@vub.ac.be Department of Psychology, Vrije Universiteit Brussel, Pleinlaan 2, Brussel B - 1050, Belgium

*These authors contributed equally to this work.

treating the agent as a rational being with beliefs, desires, and action goals, thereby inducing mentalizing when predicting the agent's behavior. However, if there is doubt about the intentionality of the agent's behavior, the likelihood of adopting the intentional stance decreases. One might instead adopt a *design stance* (Dennett, 2003), as when explaining the behaviors of machines based on their designed functionality (e.g., select objects on assembly lines), or a *physical stance* (Dennett, 2003), to explain the behavior of simple entities based on the laws of physics or chemistry (e.g., a ball falls to the ground due to gravity).

The stance humans adopt toward agents has been shown to strongly influence the degree to which cognitive resources are dedicated to social interactions with these agents. This is presumably a result of how much social relevance is ascribed to the agents' actions. Wiese, Wykowska, Zwickel, and Müller (2012) have shown that observers exhibited gaze following, a fundamental mechanism of social cognition, to a higher degree when they observed the gaze behavior of a human agent relative to that of a preprogrammed "robot." Gaze following was stronger when observers adopted the intentional stance toward what they believed was a human agent, relative to a mechanistic agent. Adopting the intentional stance toward the (presumed) mind-controlled agent might have made its behavior more socially relevant, and hence participants responded more readily to the gaze behavior it displayed compared to when the agent was perceived as unintentional (i.e., socially not relevant). Importantly, Wiese et al. (2012) as well as Wykowska, Wiese, Prosser, and Müller (2014) showed this effect to be independent of the physical appearance of the agent, but dependent on what observers believed regarding the underlying cause of observed behavior (human-controlled vs. preprogrammed). In sum, it appears that the same robotic agent can be treated both as a socially relevant, intentional agent or a machine-like, preprogrammed "robot," depending on participants' beliefs about the cause of the agent's actions (Waytz et al., 2010; Wiese et al., 2012).

Neural correlates of adopting the intentional stance and attributing intentional agency

Numerous neuroimaging studies exploring inferences about others' internal states demonstrated that the temporoparietal junction (TPJ), precuneus, and medial prefrontal cortex (mPFC) are core brain areas during mentalizing (see meta-analyses by Schurz, Radua, Aichhorn, Richlan, & Perner, 2014; Van Overwalle & Baetens, 2009). Specifically, activation in these areas was also observed while adopting the intentional stance and attributions of agency to others, including the TPJ (Chaminade et al., 2012; Krach et al., 2008), precuneus (Chaminade et al., 2012), and mPFC (Gallagher, Jack, Roepstorff, & Frith, 2002). For instance, when participants were led to believe that they played a competitive game against a human or a computer, activation in the TPJ and mPFC was increased in the human, but not the computer, condition (Gallagher et al., 2002; Krach et al., 2008).

Interestingly, however, during nonsocial tasks, such as orienting attention to locations in visual space, the TPJ has also been shown to be activated (Mitchell, 2008; Özdem, Brass, Van der Cruyssen, & Van Overwalle, 2016; Scholz, Triantafyllou, Whitfield-Gabrieli, Brown, & Saxe, 2009) as part of the ventral attention network that is driven by salient or unexpected "cues" in the environment (Cabeza, Ciaramelli, & Moscovitch, 2012; Corbetta, Kincade, Ollinger, McAvoy, & Shulman, 2000; Corbetta & Shulman, 2002). To study the neural correlates of attentional reorienting, neuroimaging studies have used Posner's spatial cueing task (Posner, 1980). Participants are presented with an advance cue, such as a central arrow or a person's gaze (Friesen & Kingstone, 1998), directing their attention to a given location on a screen. Their task is to respond to a target stimulus that appears either at the cued (i.e., valid) or at an uncued (i.e., invalid) location. If the target appears at an uncued location, a *reorienting* shift of attention from the cued to the actual target location is required. This shift results in significantly longer reaction times in the invalid compared to the valid condition. Neuroimaging studies using spatial cueing tasks demonstrated that such shifts of attention are accompanied by increased activation in the TPJ (Corbetta et al., 2000; Corbetta & Shulman, 2002; Doricchi, Macci, Silvetti, & Macaluso, 2010; see meta-analyses by Decety & Lamm, 2007; Van Overwalle & Baetens, 2009; and reviews by Cabeza et al., 2012; Geng & Vossel, 2013).

Interaction between mind attribution and attentional reorienting

The neuroimaging studies described above suggest a close link between the social process of mind attribution and the cognitive process of attentional orienting to spatial cues when interacting with other agents. This relationship has also been demonstrated in several behavioral studies. For example, Teufel and colleagues (2009) found that observing an agent's gaze direction resulted in smaller validity effects when participants believed that the observed person was wearing goggles that were opaque (in which case the person was believed to be effectively "blind") rather than transparent. Likewise, Wiese, Wykowska, and colleagues (2012)

showed that gaze validity effects were larger when eye movements were believed to be resulting from intentional, human-like agency, rather than from a preprogrammed algorithm. The modulatory effect of mind attribution on gaze cueing appeared to be due to enhanced sensory processing of stimuli presented at the attended location (Wykowska et al., 2014), as evidenced by event-related potentials of the EEG signal.

Earlier neuroimaging research demonstrated that the TPJ is preferentially recruited during the attribution of particular beliefs and intentions (Saxe, Moran, Scholz, & Gabrieli, 2006; Saxe & Powell, 2006), although the mPFC is sometimes also activated (Schurz et al., 2014; Van Overwalle & Baetens, 2009). Recent meta- and connectivity analyses of fMRI data suggests that the TPJ can be segregated into two subareas, and that these subareas subserve distinct processes of belief attribution and attention reorientation (Bzdok et al., 2013; Krall et al., 2015; Kubit & Jack, 2013). The posterior TPJ is believed to subserve mentalizing, while the anterior TPJ subserves primarily attention reorientation to unexpected stimuli (Bzdok et al., 2013; Kubit & Jack, 2013). However, Krall et al. (2015) argued, based on their meta-analytic data, that the anterior TPJ area is also responsible for mentalizing, including the intentional stance and belief inference.

Aim of study

Since previous studies showed that the anterior TPJ is involved in both attentional reorienting and mind attribution, it is plausible that the anterior TPJ is an area of the brain where interactive processes between mind attribution and attentional reorienting occur. To examine this hypothesis, we used a belief manipulation paradigm similar to those described above (e.g., Gallagher et al., 2002; Teufel et al., 2009; Wiese et al., 2012; Wykowska et al., 2014), in combination with a spatial reorienting protocol involving gaze cues. Specifically, we led participants to believe that the eyes of a robot were either preprogrammed or controlled by a human experimenter via a joystick. As described earlier, in a similar paradigm, Wykowska, Wiese, and colleagues (2014) and Wiese, Wykowska, and colleagues (2012) found larger gaze validity effects for the human compared to the preprogrammed condition, likely reflecting the increased social relevance ascribed to the observed eye movements in the human-controlled vs. the preprogrammed condition. Therefore, participants would be more likely to orient attention to gazed-at locations and expect to see the target at the gazed-at location when they believe the gaze results from operations of the mind, rather than being algorithmically determined. Extending on these findings, we expect that the human-controlled condition triggers stronger mind attributions (involving the posterior or anterior TPJ; see Krall et al., 2015) as well as stronger attention to the observed gaze (involving the anterior TPJ) than in the preprogrammed condition. More importantly, we expect to find interaction effects between cue validity and mind attribution, reflecting larger effects of attentional reorientation after invalid cues (involving the anterior TPJ) when the observed eye gaze is controlled by a human rather than being preprogrammed.

Method

Participants

Twenty-seven naive adults took part in this study (11 women; age range: 18–28 years; mean age: 21.14 years), and data of 21 participants were analyzed: 2 participants were excluded due to excessive head movements and another 4 participants because they expressed suspicion about the manipulation, as evidenced by their answers on a funneled questionnaire. All remaining participants believed the manipulation and were convinced that the experimenter steered the robot in the "human-controlled" condition. All participants were right-handed, as assessed by the Dutch version of the Edinburgh Inventory (Oldfield, 1971). They were paid 10 euros for their participation. Participants reported no abnormal neurological history and had normal or corrected-to-normal vision. Participants gave informed consent prior to the experiment in accordance with the guidelines of the Medical Ethics Committee at the Ghent University Hospital and Brussels University Hospital.

Stimuli

We used the same stimuli as described in Wykowska et al. (2014). As the gazing stimulus, we used a photo of an anthropomorphic robot ("EDDIE," developed by TU Munich). The robot's face was oriented frontally and did not move or change its orientation. The eyes were positioned on the central horizontal axis of the screen and moved to either the left or the right on a given trial. The (response-relevant) target stimuli were either a black capital letter "F" or a "T" (sized 0.2° × 0.2° of visual angle) presented peripherally and aligned with the eyes. The screen background was white (Figure 1).

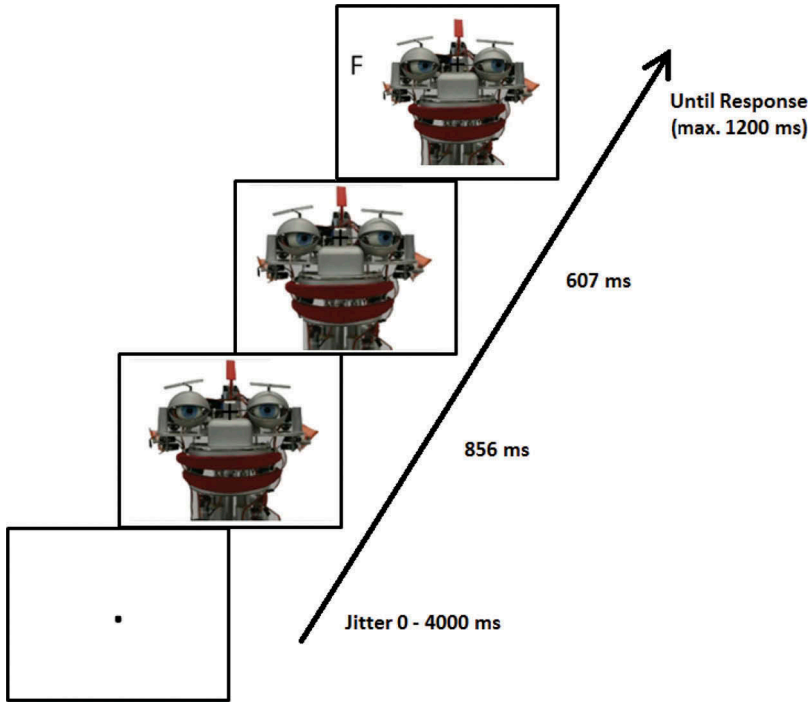

Figure 1. Stimuli and design of the experiment.

Procedure

Our procedure was adopted from Wykowska et al. (2014), with some notable changes regarding the belief manipulation and the order in which instructions were presented (see below for details). All trials started with a dot that was presented in the middle of the screen for a jittered duration between 0 and 4000 ms. Subsequently, the robot's face was presented on the screen for 856 ms, with gaze directed straight ahead (i.e., facing the observer). During that time, the fixation dot remained visible between the eyebrows of the face; see Figure 1. Afterwards, the robot changed its gaze direction to the left or the right side of the screen (i.e., gaze cue). After another 607 ms, the peripheral target letter was presented on either the left or the right side and remained visible until the participant produced a response (via button-press) to the target or after a timeout criterion (1200 ms) was reached. Participants were asked to respond as quickly and accurately as possible to the identity of the target (F vs. T) using the assigned buttons. Upon response, the screen went blank and the next trial started.

Crucially for this paradigm, the robot's gaze was directed either to the side on which the target appeared (valid trials, 50% of trial) or to the other side (invalid trials, 50% of trials). Nonpredictive gaze cueing was chosen because the goal of the study was to manipulate attributions of intentionality only via instructions (i.e., beliefs) and not via behavior (i.e., a highly reliable or unreliable cue would be expected to induce perceptions of humanness over and above the instruction manipulation of human vs. preprogrammed control). Participants were informed at the beginning of the experiment that gaze direction was not predictive of the target position. In addition to cue validity (valid vs. invalid), we also manipulated participants' belief of how the robot's eye movements were controlled. In the human-controlled condition, participants were told that the robot's eye movements were controlled by a human via joystick and transferred to the robot's eyes in real time. In the preprogrammed condition, they were told that the eye movements of the robot were preprogrammed prior to the experiment. These two manipulations were presented in two consecutive blocks, presented in a counterbalanced order. Altogether, there were 384 trials, 192 trials in each condition.

Participants received both instructions prior to scanning, while the order of the blocks during scanning was counterbalanced across participants in the experiment. After having read the instructions and before scanning, participants performed a practice run to ensure that they had understood the instructions: during the practice trials, the experimenter sat opposite to the participant with a laptop placed in front of the experimenter as well as the participant. In the human-controlled condition, a joystick was connected to the experimenter's laptop and participants could see and hear the

experimenter controlling the joystick: every time the joystick was moved, there was an audible click sound. In the preprogrammed condition, the joystick was not connected to the laptop and participants were told that the eye movements were preprogrammed. In reality, in both the preprogrammed and the human-controlled condition, the robot's eye movements were actually preprogrammed. Participants completed 16 practice trials (8 trials per instruction condition).

After the practice trials and in order to avoid any skepticism, the experimenter plugged the joystick into the computer in the control room of the scanner and participants were shown a microphone that would transmit the sound of the click to their headphones in the scanner. In reality, though, the click sound was recorded in advance and played during each trial in the human-controlled condition, in the same manner as in the practice session before scanning. After scanning, the participants filled in a funneled questionnaire about the manipulation of conditions, in which they were probed for suspicions with increasingly specific questions (see Appendix).

Note that the present procedure differs from the original protocol of Wykowska, Wiese, and colleagues (2014) in two major respects: first, Wykowska et al. (2014) did not use the joystick (and related sound) manipulation in the human-controlled condition, but instead told participants that the robot's eye movements represented real human eye movements that were performed by a human, detected by an eye tracker, and transferred to the robot in real time. Second, in Wykowska et al. (2014) instructions regarding the "human-controlled" and "preprogrammed" conditions were not presented together before the actual experiment, but instead individually at the beginning of each experimental block (i.e., participants did not know at the beginning of the experiment that there would be two conditions and what the nature of these conditions would be). Since it is difficult to provide novel instructions convincingly under scanner conditions, participants in the present study were presented with both instructions before entering the scanner. Given this, presentation of the click-sound in the human-controlled but not in the preprogrammed condition was designed to help participants to maintain a constant mental set; that is, the click sound or absence of it in the two conditions was meant to reinforce the instruction.

MRI data acquisition

Images were obtained using a 3 T Magnetom Trio MRI scanner system (Siemens Medical Systems, Erlangen, Germany), using a 32-channel radiofrequency head coil. First, high-resolution anatomical images were collected using a T1-weighted 3D MPRAGE sequence [repetition time (TR) = 2530 ms, echo time (TE) = 2.58 ms, inversion time = 1100 ms, acquisition matrix = 256 × 256 × 176, sagittal field of view (FOV) = 220 mm, flip angle = 7, voxel size = 0.9 × 0.86 × 0.86 mm^3 (resized to 1 × 1 × 1 mm^3)]. Second, whole-brain functional images were acquired by using a T2*-weighted gradient echo sequence (TR = 2000 ms, TE = 35 ms, image matrix = 64 × 64, FOV = 224 mm, flip angle = 80°, slice thickness = 3.0 mm, distance factor = 17%, voxel size = 3.5 × 3.5 × 3.5 mm^3, 30 axial slices). In the scanner, stimuli were projected onto a screen at the end of the magnet bore and participants viewed the stimuli through an angled mirror located above their eyes on the head coil. Stimulus presentation was controlled by E-Prime 2.0 (www.pstnet.com/eprime; Psychology Software Tools) running under Windows XP. Participants were placed head first and supine in the scanner bore. They were instructed not to move their heads to avoid motion artifacts and foam cushions were placed to minimize head movements.

Image processing

The fMRI data were preprocessed and analyzed using SPM8 (Wellcome Department of Cognitive Neurology, London, UK). Prior to the statistical analysis, data were preprocessed to remove sources of noise and artifact. Slice-time correction was applied in order to amend differences in acquisition time between slices for each whole-brain volume, realigned within and across runs for the removal of the movement effects. The functional data were then transformed into a standard anatomical space (2 mm isotropic voxels) based on the ICBM152 brain template (Montreal Neurological Institute). Normalized data were then spatially smoothed (6 mm full-width at half-maximum) using a Gaussian Kernel. Finally, the preprocessed data were examined using the Artifact Detection Tool software package (ART; http://web.mit.edu/swg/art/art.pdf; http://www.nitrc.org/projects/artifact_detect/) for excessive motion artifacts and for correlations between motion and experimental design, and between global mean signal and experimental design. Outliers were identified in the temporal differences series by assessing between-scan differences (*Z*-threshold: 3.0 mm, scan-to-scan movement threshold: 0.5 mm; rotation threshold: 0.02 radians). These outliers were omitted in the analysis by including a single regressor for each outlier. No correlations between motion and experimental design or global signal and experimental design were identified. Six directions of motion parameters from the realignment step as well as outlier time points (defined by

ART) were included as nuisance regressors. We used a default high-pass filter of 128 s and serial correlations were accounted for by the default autoregressive AR (1) model.

Statistical analyses

Four regressors were defined reflecting the crossing of the factors validity (valid vs. invalid) and Instruction (human vs. preprogrammed), collapsing across the left and the right side of the screen where the target appeared. Onsets were specified at the moment when the target appeared on the screen after the robot had directed its gaze to the left or right and participants had to respond immediately using a canonical hemodynamic response function with event duration set to 0 s. The six head movement parameters were also included in the model. The regressors of interests were calculated at the single level for each subject and used at the second level.

At the second level, we calculated the main effect of instruction [($Human_{Valid}$ + $Human_{Invalid}$) > ($Preprogrammed_{Valid}$ + $Preprogrammed_{Invalid}$)] and validity [($Human_{Invalid}$ + $Preprogrammed_{Invalid}$)] > ($Human_{Valid}$ + $Preprogrammed_{Valid}$)] and their interaction [($Human_{Invalid}$ > $Human_{Valid}$) > ($Preprogrammed_{Invalid}$ > $Preprogrammed_{Valid}$)]; Specifically, if we use for the conditions in the following order – $Human_{Invalid}$, $Human_{Valid}$, $Preprogrammed_{Invalid}$, $Preprogrammed_{Valid}$ – the contrast values are for the main effect of instruction: 1 1 –1 –1; the main effect of Validity: 1 –1 1 –1; and the interaction: 3 –2 1 –2. The interaction contrast implements the pattern of response times from this study, showing that response times are highest for $Human_{Invalid}$ (511 ms) followed by $Preprogrammed_{Invalid}$ (507 ms) and approximately equivalent for $Human_{Valid}$ (503 ms) and $Preprogrammed_{Valid}$ (502 ms; Table 1). Moreover, this interaction pattern is theoretically preferable because it allows for a larger validity effect in the human condition as opposed to the preprogramed condition, without (as in a classic interaction) prescribing a reversed validity effect in the preprogrammed condition. After this whole-brain analysis, *a priori* regions of interest (ROI) analyses of the TPJ were performed with the small-volume correction and based on a sphere of 15 mm radius around the centers (in Montreal Neurological Institute coordinates) of areas that were identified in the meta-analysis by Bzdok et al. (2013) as involved in mentalizing: 54 −54 16.5 (posterior TPJ) and attention reorienting: 58.5 −39 16.5 (anterior TPJ). Note that taking such *a priori* coordinates from independent analyses as basis for defining ROIs is statistically and theoretically ideal because the ROIs are independent from the current data and based on a large data set of a meta-analysis. However, given individual variations, this sometimes requires to use a relatively large radius (see also Ma et al., 2012). Lastly, the mean percentage signal change in each ROI was extracted using the MarsBar toolbox (http://marsbar.sourceforge.net) and correlated with the behavioral response times.

Results

Behavioral results

The behavioral data were examined with a 2 × 2 analysis of variance (ANOVA) with instruction (human vs.

Table 1. Descriptive statistics and analysis of the behavioral data.

	All participants (n = 27)				Nonsuspicious participants (n = 21)			
	Descriptive statistics							
	Human		Preprogrammed		Human		Preprogrammed	
	Valid	Invalid	Valid	Invalid	Valid	Invalid	Valid	Invalid
Response time (ms)								
Mean	504	512	505	510	503	511	502	507
SD	24	28	19	24	22	27	18	23
Accuracy (%)								
Mean	93	93	93	92	93	93	93	93
SD	3	3	3	4	3	3	3	3

	ANOVAs					
	F(1,26)	p	np^2	F(1,20)	p	np^2
Response time						
Instruction	0.10	0.760	0.004	0.34	0.568	0.017
Cue validity	19.28	0.000	0.426	13.04	0.002	0.395
Interaction	1.08	0.308	0.040	1.19	0.289	0.056
Accuracy						
Instruction	1.22	0.279	0.045	0.42	0.527	0.02
Cue validity	0.52	0.477	0.020	0.00	1	0
Interaction	0.21	0.645	0.008	0.00	1	0

preprogrammed) and cue validity (valid vs. invalid) as within-participant factors. The analysis involved individual participants' means per condition, and for the response times this was calculated only for correct-response trials and after excluding outliers beyond two standard deviations from the mean of each condition across all participants. As can be seen in Table 1, the main effect of instruction was not significant, while the main effect of the cue validity was significant. The interaction between cue validity and instruction was not significant; although numerically the pattern of cueing effects was in line with expectations (effects of 8 vs. 5 ms in the human-controlled vs. preprogrammed conditions), simple *t*-tests failed to yield a significant difference. The accuracy data did not show any main or interaction effects. These results did not change when all participants were included in the analysis (Table 1).

Imaging results

Whole-brain analysis

The whole-brain analysis (Table 2, Figure 2) revealed that the validity contrast (invalid > valid) showed activation in the right precuneus and right TPJ. The instruction contrast (human > preprogrammed) showed activation in the bilateral TPJ, the cuneus, the right superior parietal lobule and the right postcentral gyrus. Critically, the interaction contrast revealed activation in the bilateral TPJ as predicted, and additional activation in the bilateral superior temporal gyrus, right lentiform nucleus, left Insula, and left claustrum.

ROI analysis

To explore our hypothesis, we identified two ROIs based on the meta-analysis by Bzdok et al. (2013), which reflect two distinct functions of the TPJ: the posterior TPJ involved in mentalizing and the anterior TPJ involved primarily in attentional reorienting. We analyzed these ROIs using small-volume correction, applying the same contrasts (Table 2, Figure 2). The validity contrast was only active in the ROI of the right posterior TPJ (attributed to mentalizing). Consistent with our hypothesis, the instruction contrast as well as the interaction contrast activated the bilateral anterior TPJ (attributed to attention reorienting).

Correlations with behavioral response data

For exploratory purposes, we computed Pearson correlations between the activity (% signal change) for the ROIs of the bilateral anterior TPJ and the behavioral response times. Note that these anterior TPJ ROIs were obtained based on the activations in the

Table 2. Whole-brain and regions of interest (ROI) analysis of the effect of validity and instruction.

Comparison and anatomical area		*x*	*y*	*z*	Voxels	Max *t*
Whole-brain analysis						
Invalid > valid						
	Right TPJ	42	−50	30	222	4.25*
	Right precuneus	40	−76	34	274	4.39**
Human > preprogrammed						
	Right TPJ	56	−22	10	2671	8.26***
	Left TPJ	−44	−32	10	1561	7.12***
	Cuneus	8	−80	26	1894	4.67***
	Right superior parietal lobule	26	−50	64	195	4.41*
	Right postcentral gyrus	58	−12	50	619	4.65***
Interaction: human (invalid > valid) > preprogrammed (invalid > valid) with contrast [3 −2 1 −2]						
	Right TPJ	56	−22	10	783	5.73***
	Right lentiform nucleus	34	−18	0	783	4.35***
	Right superior temporal gyrus	52	−12	4	783	4.90***
	Left TPJ	−44	−32	10	337	4.73**
	Left insula	−42	−36	24	337	4.25**
	Left superior temporal gyrus	−50	−10	2	242	4.61*
	Left claustrum	−34	−6	−2	242	4.25*
Regions of interest						
Invalid > valid						
	Right posterior TPJ mentalizing	46	−50	28	17	3.63*
Human > preprogrammed						
	Right anterior TPJ reorienting	54	−26	8	357	7.59***
	Left anterior TPJ reorienting	−46	−32	10	334	6.80***
Interaction: human (invalid > valid) > preprogrammed (invalid > valid) with contrast [3 −2 1 −2]						
	Right anterior TPJ reorienting	54	−26	8	82	5.00**
	Left anterior TPJ reorienting	−46	−32	10	66	4.35**

Note: *x*, *y*, and *z* = Montreal Neurological Institute coordinates of the peak values; *t* = *t*-score of the peak values.
ROIs are spheres with 15 mm radius around coordinates ± 58.5–39 16.5 (aTPJ reorienting), ±54–54 16.5 (pTPJ mentalizing) according to Bzdok et al. (2013). Whole-brain analysis with cluster extent > 10 voxels and small-volume analysis (only for ROIs) with cluster extent > 1 voxel, both with $p < .001$. Listed are clusters that are significant at $p < .05$, family-wise error (FWE) cluster-corrected. The weights in the interaction contrast have the same order as the conditions in that row.
*$p \leq .05$; **$p < .01$; ***$p < .001$ (FWE-corrected).

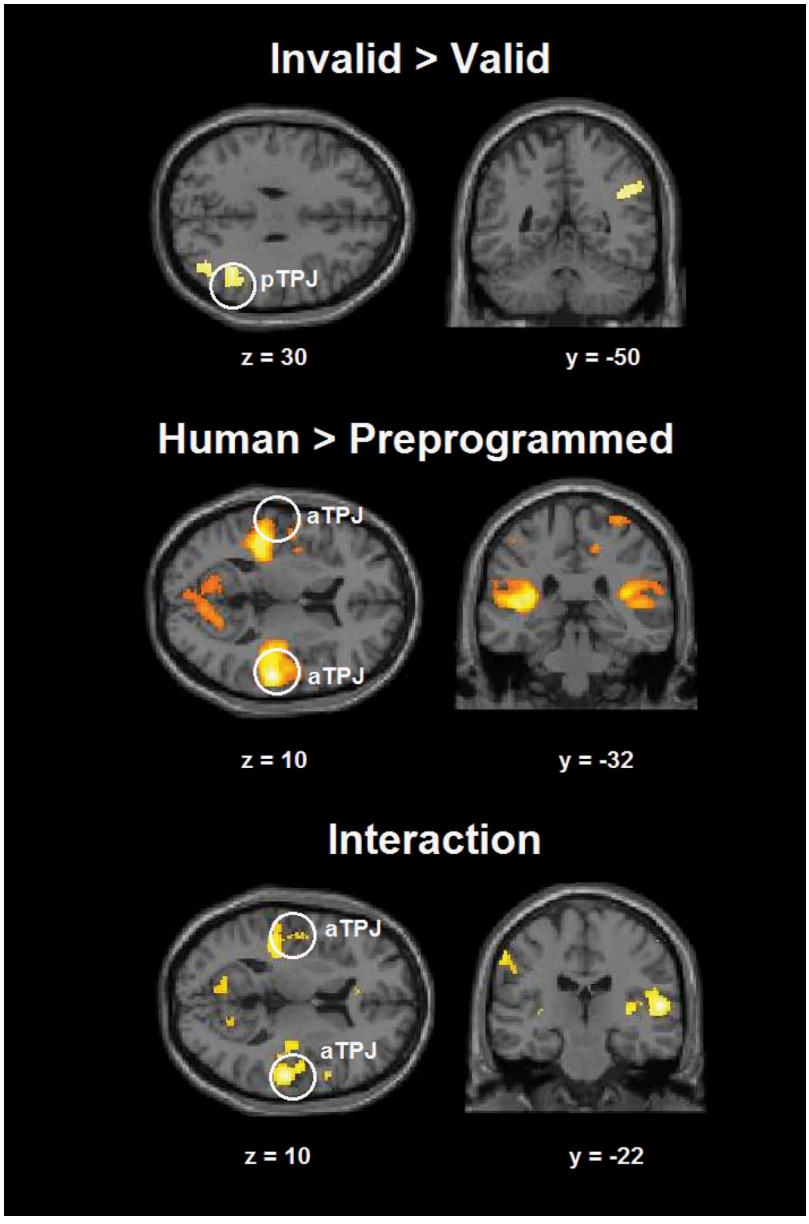

Figure 2. Whole-brain activation ($p < .001$ uncorrected) for the main effects (invalid > valid and human > preprogrammed contrasts) and their interaction. Circles indicate ROIs with significant activation in the anterior TPJ (aTPJ) and posterior TPJ (pTPJ) (ROIs have a 15 mm radius centered around the mean coordinates reported in Bzdok et al., 2013).

instruction main contrast (human > preprogrammed) as well as in the interaction between cue validity and Instruction. We found negative correlations across participants for the bilateral anterior TPJ in the human conditions, which reached significance when pooled together for the right TPJ ($r = -.28$, $p < .05$ one-sided) and the left TPJ ($r = -.41$, $p < .01$ one-sided). Thus, the stronger the activation in the anterior TPJ (main effect of instruction or interaction between cue validity and instruction), the faster participants responded in the human conditions. There were no significant correlations for the preprogrammed conditions.

Discussion

The aim of the current study was to investigate – using a gaze-cueing paradigm – how the relationship between higher-order social cognition (i.e., mind attribution) and low-level cognitive processes (i.e., attentional reorientation) is established in the brain. Beliefs about mind involvement were manipulated via instructions: in the human condition, participants were told that the eye movements of a robotic agent were controlled by the experimenter (via joystick), while in the preprogrammed condition participants believed that the eye movements were predetermined (prior to the

experiment). In fact, eye movements were always preprogrammed and replayed to the participants during the experiment. We hypothesized that gaze behavior believed to result from intentional human agency might be treated as being socially more relevant than gaze behavior generated by a preprogrammed algorithm. Consequently, we expected that participants would follow the gaze cues more in the human-controlled condition compared to the preprogrammed condition, resulting in larger cueing effects. Critically, given that the neural substrates of attentional reorienting are located in the anterior TPJ, we predicted that a modulatory effect of mind attribution on attentional reorienting would raise activation in this brain area. We were less specific about the neural substrates of the human-controlled instruction itself because this might depend on mind attribution (involving the anterior or posterior TPJ) or greater orientation to the cue itself (involving the anterior TPJ).

On the neuronal level, the fMRI data showed the expected interaction between validity and beliefs, with greater activation in the anterior TPJ after an invalid cue. This area is known to subserve primarily attention regulation and reorientation to unexpected stimuli (Bzdok et al., 2013; Krall et al., 2015; Kubit & Jack, 2013). This supports the notion that the anterior TPJ is a hub area involved in both attention and mentalizing (Krall et al., 2015). In line with this, it is interesting to note that Bzdok et al., (2013, p. 381) argued that the anterior TPJ "links two antagonistic brain networks processing external versus internal information" involved in attention reorientation and mentalizing, respectively. Similarly, Carter and Huettel (2013) speculated that the TPJ is a nexus area where lower-level functions of different processing streams are combined to produce higher-order social-cognitive functions. Holding the belief that the robot was controlled by a human increased activation also in a nearby brain area: the bilateral superior temporal gyrus. This area is responsible for the detection of important auditory and visual stimuli (Kubit & Jack, 2013; Schönwiesner et al., 2007), and especially facial stimuli (Baron-Cohen et al., 1999). Accordingly, our finding suggests that participants paid increased attention to the face and gaze of the robot under the belief that the robot's eyes were human-controlled because changes in gaze direction coming from a human partner would constitute a more important visual stimulus than those generated by a machine. This interpretation is in line with what participants reported in the postscanning suspicion questionnaire.

We also found that instructing the participants that the cue was human-controlled activated the anterior TPJ. This is consistent with our hypothesis that this part of the TPJ supports belief attribution about human mind and control (Krall et al., 2015; Saxe et al., 2006; Saxe & Powell, 2006). This belief may further trigger a greater attention to the cue, which may additionally recruit the same anterior TPJ (Cabeza et al., 2012; Corbetta et al., 2000; Corbetta & Shulman, 2002). In line with our expectation that the instruction activates attributions of beliefs of human control and so activates primarily the TPJ (Saxe et al., 2006; Saxe & Powell, 2006), we found no activation of other areas that are sometimes correlated with the intentional stance and mentalizing (e.g., Gallagher et al., 2002; Chaminade et al., 2012; Schurz et al., 2014; Van Overwalle & Baetens, 2009). It is unclear what exact beliefs participants held about the human who controlled the gaze, especially given that they were told that gaze validity was essentially random. This manipulation was chosen to make sure that gaze behavior does not induce attributions of humanness that might interfere with the belief manipulation via instruction. Although this might constitute a potential confound, the responses from the postscanning suspicion questionnaire indicate that participants indeed did attribute more intentionality to the eye movements in the human-controlled versus the preprogrammed condition (i.e., they expected to be able to use the eye movements from the experimenter in the human-controlled condition as an indicator of whether the cue was reliable or not) despite the fact that they were told that gaze behavior is random. Thus, telling participants about the randomness of the eye movements did not hold them back from ascribing more intentionality to the eye movements in the human-controlled versus the preprogrammed condition. Some participants in the postscanning questionnaire referred to the joystick click sound as a cue of human control. Note that because the neural effects reported here are in line with previous literature on mentalizing and attentional reorienting, they are most likely attributable to the instruction manipulation, and not the sound of the joystick click itself.

Interestingly, the interaction effect also showed up in neural patches located in areas surrounding the TPJ at some distance. One of these clusters is the lentiform nucleus in the basal ganglia, which has been shown to be involved in attentional reorientation in children (8–12 years old; Konrad et al., 2005). Another area sensitive to the combined effect of belief and validity manipulation is the claustrum (lateral to the putamen), which has been found to be responsible for coding the salience of incoming information and the reallocation of attentional resources (Mathur, 2014). Both structures showed

stronger activation for invalid cues in the human-controlled compared to the preprogrammed condition, which indicates that attentional reorientation to gaze direction might be partially driven by subcortical structures in conditions in which gaze behavior was thought to be human-controlled.

Unexpectedly, the main effect of validity was revealed in the posterior TPJ and precuneus, two mentalizing areas (see Krall et al., 2015; Schurz et al., 2014; Van Overwalle & Baetens, 2009). One possibility is that the posterior TPJ area reflects the overlap between attention reorienting and mentalizing, although we hypothesized, based on earlier meta-analyses (Bzdok et al., 2013; Decety & Lamm, 2007; Kubit & Jack, 2013; Krall et al., 2015), that this overlap is rather consistently found in the anterior TPJ – and this was confirmed by the interaction in the present study, which activated the anterior TPJ. Alternatively, the validity effect may show up in mentalizing areas of the present study because the validity manipulation was most effective when our participants believed that the gaze was human-controlled. Stated differently, that the validity effect activates the posterior TPJ may be indirectly due to the fact that participants attended to the validity of the gaze cue only when it was under human control. Still another explanation is that the validity manipulation was completely random (50% valid and 50% invalid cues), which may have decreased the impact of attention reorientation altogether. Instead, a valid as opposed to an invalid cue may have increased reference to human-like purposeful and predictable behavior, leading to more activity in mentalizing areas.

The behavioral results replicated the typical behavioral finding that target discrimination takes longer after an invalid compared to a valid cue. However, contrary to our expectation, the interaction between the validity effect and mind attribution was not significant. Note, however, that there were sizeable correlations between this interaction contrast and response times, indicating that participants who showed stronger contrast activations in the anterior TPJ responded faster in the human-controlled conditions. The lack of a statistically significant interaction at the behavioral level may be attributable to several factors, which apparently affected the behavioral results more than the neural effects: (i) given that the behavioral validity effect, although significant, is relatively small in the order of 5–10 ms, the lower number of trials under the scanner (only 40% of the behavioral studies) due to the longer trial duration caused by the addition of random jitters and longer intertrial intervals (to let the activation return back to baseline) might have increased the overall variability of the response times. (ii) In order to accommodate the constraints of the scanner environment, in the present paradigm, participants received both instructions (human-controlled vs. preprogrammed) together, prior to the experiment – as opposed to receiving the respective instruction only before performing the corresponding condition (as in Wykowska et al., 2014). (iii) The switch from a more offline paradigm (as in Wiese et al., 2012) to a more "online" social interaction (with more direct sensory feedback through the joystick click sounds in the present paradigm) might also have affected the pattern of results (Schilbach et al., 2013). Even if these reasons affected only some of the participants, they would have diluted or altered the effects of the belief manipulation, making it harder to observe the behavioral interaction pattern repeatedly found by Wiese et al. (2012) and Wykowska et al. (2014).

Conclusion

To summarize, the goal of the present study was to investigate the neural correlates underlying the effect of mind attribution on attentional reorienting to gaze direction. We found that higher-order social attribution processes and lower-level attention mechanisms intersect at the anterior TPJ responsible for attentional reorientation and belief inference (Krall et al., 2015), which exhibited particularly strong activation for invalid cues under the belief that eye movements were human-controlled. Based on our findings, gaze cues believed to be produced by a human rather than a machine seem to attract more attention and most likely reflect the fact that they are socially more relevant and informative than preprogrammed ones. This adds to the growing evidence that mind attribution and attentional orienting interact in a common, underlying attentional control process (Mitchell, 2008; Özdem et al., 2016; Scholz et al., 2009), which ensures that more attentional resources are devoted to interactions with agents who are believed to have a mind, as compared to machine-like agents.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was supported by Research Foundation Flanders (FWO) Grant to Frank Van Overwalle, and performed at GIfMI (Ghent Institute for Functional and Metabolic Imaging); Fonds Wetenschappelijk Onderzoek [FWOAL556].

ORCID

Frank Van Overwalle http://orcid.org/0000-0002-2538-9847

References

Baron-Cohen, S. (1995). *Mindblindness: An essay on autism and theory of mind*. Boston, MA: MIT Press/Bradford Books.

Baron-Cohen, S., Ring, H. A., Wheelwright, S., Bullmore, E. T., Brammer, M. J., Simmons, A., & Williams, S. C. (1999). Social intelligence in the normal and autistic brain: An fMRI study. *The European Journal of Neuroscience*, *11*(6), 1891–1898. doi:10.1046/j.1460-9568.1999.00621.x

Bzdok, D., Langner, R., Schilbach, L., Jakobs, O., Roski, C., Caspers, S., … Eickhoff, S. B. (2013). Characterization of the temporo-parietal junction by combining data-driven parcellation, complementary connectivity analyses, and functional decoding. *NeuroImage*, *81*, 381–392. doi:10.1016/j.neuroimage.2013.05.046

Cabeza, R., Ciaramelli, E., & Moscovitch, M. (2012). Cognitive contributions of the ventral parietal cortex: An integrative theoretical account. *Trends in Cognitive Sciences*, *16*(6), 338–352. doi:10.1016/j.tics.2012.04.008

Carter, R. M., & Huettel, S. A. (2013). A nexus model of the temporal-parietal junction. *Trends in Cognitive Sciences*, *17*(7), 328–336. doi:10.1016/j.tics.2013.05.007

Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutcher, 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*, *6*(May), 103. doi:10.3389/fnhum.2012.00103

Corbetta, M., Kincade, J. M., Ollinger, J. M., McAvoy, M. P., & Shulman, G. L. (2000). Voluntary orienting is dissociated from target detection in human posterior parietal cortex. *Nature Neuroscience*, *3*(3), 292–297. doi:10.1038/73009

Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews. Neuroscience*, *3*(3), 215–229. doi:10.1038/nrn755

Decety, J., & Lamm, C. (2007). The role of the right temporoparietal junction in social interaction: How low-level computational processes contribute to meta-cognition. *The Neuroscientist*, *13*(6), 580–593. doi:10.1177/1073858407304654

Dennett, D. C. (2003). True believers: The intentional strategy and why it works. In T. O 'Connor & D. Robb (Eds.), *Philosophy of mind: Contemporary readings* (pp. 370–390). London: Routledge.

Doricchi, F., Macci, E., Silvetti, M., & Macaluso, E. (2010). Neural correlates of the spatial and expectancy components of endogenous and stimulus-driven orienting of attention in the Posner task. *Cerebral Cortex (New York, N.Y. : 1991)*, *20*(7), 1574–1585. doi:10.1093/cercor/bhp215

Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, *5*, 490–495.

Frith, C. D., & Frith, U. (2006). How we predict what other people are going to do. *Brain Research*, *1079*, 36–46.

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. doi:10.1006/nimg.2002.1117

Geng, J. J., & Vossel, S. (2013). Re-evaluating the role of TPJ in attentional control: Contextual updating? *Neuroscience and Biobehavioral Reviews*, 1–13. doi:10.1016/j.neubiorev.2013.08.010

Konrad, K., Neufang, S., Thiel, C. M., Specht, K., Hanisch, C., Fan, J., … Fink, G. R. (2005). Development of attentional networks: An fMRI study with children and adults. *NeuroImage*, *28*(2), 429–439. doi:10.1016/j.neuroimage.2005.06.065

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. doi:10.1371/journal.pone.0002597

Krall, S. C., Rottschy, C., Oberwelland, E., Bzdok, D., Fox, P. T., Eickhoff, S. B., … Konrad, K. (2015). The role of the right temporoparietal junction in attention and social interaction as revealed by ALE meta-analysis. *Brain Structure and Function*, *220*(2), 587–604.

Kubit, B., & Jack, A. I. (2013). Rethinking the role of the rTPJ in attention and social cognition in light of the opposing domains hypothesis : Findings from an ALE-based meta-analysis and resting-state functional connectivity. *Frontiers in Human Neuroscience*, *7*(May). Retrieved from http://www.frontiersin.org/Journal/Abstract.aspx?s=537&name=human_neuroscience&ART_DOI=10.3389/fnhum.2013.00323

Ma, N., Vandekerckhove, M., Baetens, K., Overwalle, F., Van Seurinck, R., & Fias, W. (2012). Inconsistencies in spontaneous and intentional trait inferences. *Social Cognitive and Affective Neuroscience*, *7*, 937–950. doi:10.1093/scan/nsr064

Mathur, B. N. (2014). The claustrum in review. *Frontiers in Systems Neuroscience*, *8*(April), 1–11. doi:10.3389/fnsys.2014.00048

Mitchell, J. P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex*, *18*(2), 262–271. doi:10.1093/cercor/bhm051

Oldfield, R. C. (1971). The assessment and analysis of handedness: The Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. doi:10.1016/0028-3932(71)90067-4

Özdem, C., Brass, M., Van der Cruyssen, L., & Van Overwalle, F. (2016). The overlap between false belief and spatial reorientation in the temporo-parietal junction: The role of input modality and task. *Social Neuroscience*, 1–11. doi:10.1080/17470919.2016.1143027

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3–25. doi:10.1080/00335558008248231

Saxe, R. R., Moran, J. M., Scholz, J., & Gabrieli, J. (2006). Overlapping and non-overlapping brain regions for theory of mind and self reflection in individual subjects. *Social Cognitive and Affective Neuroscience*, *1*(3), 229–234. doi:10.1093/scan/nsl034

Saxe, R. R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, *17*(8), 692–699. doi:10.1111/j.1467-9280.2006.01768.x

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience. *The Behavioral and Brain Sciences*, *36*, 393–414. doi:10.1017/S0140525X12000660

Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., & Saxe, R. R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PloS One*, *4*(3), 1–7. doi:10.1371/journal.pone.0004869

Schönwiesner, M., Novitski, N., Pakarinen, S., Carlson, S., Tervaniemi, M., & Naatanen, R. (2007). Heschl's gyrus,

posterior superior temporal gyrus, and mid-ventrolateral prefrontal cortex have different roles in the detection of acoustic changes. *Journal of Neurophysiology*, *97*(3), 2075–2082. doi:10.1152/jn.01083.2006

Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience & Biobehavioral Reviews*, *42*, 9–34. doi:10.1016/j.neubiorev.2014.01.009

Teufel, C., Alexis, D. M., Todd, H., Lawrance-Owen, A. J., Clayton, N. S., & Davis, G. (2009). Social cognition modulates the sensory coding of observed gaze direction. *Current Biology*, *19*(15), 1274–1277.

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage*, *48*(3), 564–584. doi:10.1016/j.neuroimage.2009.06.009

Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., & Cacioppo, J. T. (2010). Making sense by making sentient: Effectance motivation increases anthropomorphism. *Journal of Personality and Social Psychology*, *99*(3), 410–435. doi:10.1037/a0020240

Wiese, E., Wykowska, A., Zwickel, J., & Müller, H. J. (2012). I see what you mean: How attentional selection is shaped by ascribing intentions to others. *PloS One*, *7*(9), e45391. doi:10.1371/journal.pone.0045391

Wykowska, A., Wiese, E., Prosser, A., & Müller, H. J. (2014). Beliefs about the minds of others influence how we process sensory information. *PLoS ONE*, *9*(4), e94339. doi:10.1371/journal.pone.0094339

Appendix

Postscanning funneled questionnaire probing for suspicion after the preprogrammed (in bold) and human-controlled conditions. In a funneled questionnaire or interview, general questions are asked at the beginning and are followed by increasingly specific questions probing directly for suspicion.

(1) Was the control of the **[preprogrammed]** eye movements [by the experimenter] successful? If yes, how did you establish that? If no, why did it not work?
(2) How good was the **[programming of]** [control of the experimenter on] the eye movements? (using a seven-point scale with anchors 1 = *not vivid at all* to 7 = *very vivid*)
(3) What were the differences between the conditions? Please note them all.
(4) Did you believe that the **[eye movements were preprogrammed]** [experimenter had influence on the eye movements of the robot]? If yes, how did you establish that? If no, why did it not work?
(5) How large was the influence of the [**computer program**] [experimenter] on the eye movements? (using a seven-point scale with anchors 1 = *none at all* to 7 = *complete influence*)
(6) Did you have any suspicion that the [**computer program**] [human experimenter] did not really control the eye movements? If yes, how did you establish that? If no, why not?
(7) How strong was your suspicion? (using a seven-point scale with anchors 1 = *none at all* to 7 = *very much*)