

PHYSICAL APPEARANCE AS A TOP-DOWN INFLUENCER OF MIND
PERCEPTION IN HUMAN-ROBO SOCIAL ATTENTION

by

Abdulaziz Abubshait
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Psychology

Committee:

_____ Director

_____ Department Chairperson

_____ Program Director

_____ Dean, College of Humanities
and Social Sciences

Date: _____ Fall Semester 2019
George Mason University
Fairfax, VA

Physical Appearance as a Top-down Influencer of Mind Perception in Human-Robot
Social Attention

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy at George Mason University

by

Abdulaziz Abubshait
Master of Arts
George Mason University, 2017

Director: Eva Wiese, Professor
Department of Psychology

Fall Semester 2019
George Mason University
Fairfax, VA

Copyright 2019 Abdulaziz Abubshait
All Rights Reserved

DEDICATION

This dissertation is dedicated to my mother, father, and brother whose support and encouragement made this possible.

ACKNOWLEDGEMENTS

I would like to thank my family, friends, mentors, and colleagues for the continuous inspiration. I would also like to thank my committee members for their guidance throughout my graduate studies and making my educational experience unforgettable.

TABLE OF CONTENTS

	Page
List of Figures	viii
Abstract	xi
Introduction	1
Mind perception and the social brain	1
Joint attention	3
Mind perception, social attention, the social brain.....	5
Rationale.....	5
Study 1.	8
Abstract	8
Introduction	9
Methods and materials	14
Participants	14
Apparatus	15
Stimuli	15
Procedure	17
Analysis	21
Results	22
Discussion	25
Study 2.	34
Abstract	34
Introduction	35
Aim of study	41
Methods and materials	43
Participants	43
Stimuli	44
Tasks	45
Mind perception task.....	45
Social attention task	46
Procedure	49

Analysis	50
Behavioral data	50
fMRI data	52
Image acquisition and preprocessing	52
Neural activation associated with mind perception	53
Association of neural activation during mind perception and social attention ..	54
Results	55
Behavioral data	55
Mind perception task.....	55
Social attention task	55
Link between physical humanness, mind perception, and social attention	56
fMRI data.....	60
Neural activation associated with mind perception	60
Association of neural activation during mind perception and social attention	61
Discussion	62
Conclusion.....	71
Study 3.	73
Introduction	73
Methods	82
Participants	82
Apparatus.....	83
tDCS stimulation	83
Brain modelling	85
Stimuli	87
Procedure	88
Results	92
Questionnaires	92
Behavioural data	93
Discussion	97
Conclusion.....	103
General Discussion/Synopsis.....	105
References.....	108

LIST OF FIGURES

Figure 1. Manipulation of mind judgments. Human-like appearance (80% physical humanness) and reliable behavior (80% predictive cueing) should increase the likelihood that mind is attributed, while robot-like appearance (20% physical humanness) and random behavior (50% predictive cueing) should decrease the likelihood that mind is attributed to an agent.....	17
Figure 2. Sequence of events on a trial of gaze cueing. Participants first fixated on a fixation cross for 700–1000 ms and were then presented with an agent (human vs. robot) looking straight for 700–1000 ms, followed by a change in gaze direction (either to the left or right side of the screen). After a SOA of 500 ms, the target letter (F or T) appeared either where the face was looking (valid) or opposite of where the face was looking (invalid). The target remained on the screen until a response was given or a timeout of 1200 ms was reached. A blank screen marked the end of the trial and was presented for 680 ms.....	20
Figure 3. Gaze-cueing effects as a function of appearance and behavior. Reliable agents induced significantly larger gaze-cueing effects than agents showing random behavior, independent of appearance. There was neither a significant main effect of appearance on gaze-cueing effects, nor was there a significant interaction effect between appearance and behavior.....	23
Figure 4. Mind judgments before and after gaze following. Agent appearance affected mind ratings both before interacting with the agents during the gaze following task (A) and afterward (B), with higher agent ratings for the human than for the robot agent. The reliability with which agents cued the target location did not have an effect on agent ratings, neither at the pre-interaction stage (i.e., no baseline difference in agent ratings between participants in the reliable and the random condition), nor at the post-interaction stage (i.e., knowing about the reliability of the agents did not influence the degree to which mind was ascribed to them). The interaction effect was not significant for either of the two rating times.....	26
Figure 5. Change in mind judgments from pre- to post-interaction. Mind ratings for the human-like agent decreased significantly during gaze following, while mind ratings for the robot-like agent increased during gaze following. Interestingly, this effect is independent of the reliability with which the agents predicted the target location during gaze following.....	29
Figure 6. Stimuli used for the mind perception and social attention tasks. The images were created by morphing a robot face (Mekarobot; image on the very left) into a human face (adult male; image on the very right) in steps of 20%.	45
Figure 7. Mind perception task. While inside of an fMRI scanner, participants were presented with theory-of-mind questions and a series of morphed images to judge. Participants rated each image on a 1–8 scale.....	48
Figure 8. Social attention task. Outside of the fMRI scanner, participants performed the gaze cueing task using the same morphed images from the mind perception task.	

Participants were required to identify the identity of a target letter, presented in the periphery that was preceded by either a congruent or incongruent looking morph image	49
Figure 9. Equations tested in the nested model comparison. Data were modeled using a linear model (a), as well as a quadratic (b), cubic (c), and fourth-level (d) polynomial. The error terms and intercept have been omitted in all of the equations	51
Figure 10. Average mind ratings as a function of physical humanness. 1 = 0% physical humanness; 6 = 100% physical humanness, as modeled by a linear (a), quadratic (b), cubic (c), and fourth-level polynomial (d) model. The fourth-level polynomial model constituted the overall best model fit (see Table 1).	57
Figure 11. Average gaze-cueing effects as a function of the degree of physical humanness. For the x-axis, 1=0%human, 6=100%human, as modeled by a linear (a), quadratic (b), cubic (c), and fourth-level polynomial (d) model. The linear model constituted the overall best model fit.	60
Figure 12. Average gaze-cueing effects as a function of mind ratings. Data were modeled by a linear (a), quadratic (b), cubic (c), and a fourth-level polynomial (d) model. None of the nonlinear models fit significantly better than the linear model, which is evidence that the linear model was the best predictor of gaze-cueing behavior.	61
Figure 13. Path diagram illustrating the mediation model. The mediation analysis revealed both a significant and positive direct effect of physical humanness on gaze cueing, as well as a negative indirect effect, as mediated by mind ratings. Values over the directional arrows reflect standardized coefficients produced from each regression model in the mediation. * $p < .05$. ** $p < .01$	62
Figure 14. A priori parametric analysis of fMRI activations based on mind ratings. Z maps reflecting onset of the morph images, using mind ratings to weight the parametric regressor. From left to right: coronal ($y = 58$), sagittal ($x = -2$), and axial ($z = 2$) slices; no activations survived correction for multiple comparisons.	64
Figure 15. Post hoc parametric analysis of fMRI activations based on mind ratings. Z maps reflecting onset of the morph images, using mind ratings to weight the parametric regressor. From left to right: coronal ($y = 54$), sagittal ($x = -12$), and axial ($z = -14$) slices; cluster corrected ($Z = 1.96$, $p < .05$) at the whole-brain level.	66
Figure 16. Parametric analysis of fMRI activations based on gaze cueing effects, mind ratings, and their conjunction. Z maps reflecting onset of the morph images, using either mind ratings (orange) or gaze cueing (green) to weight the parametric regressor, along with their conjunction (yellow). From left to right: coronal ($y = 54$), sagittal ($x = -12$), and axial ($z = -14$) slices; cluster corrected ($Z = 1.96$, $p < .05$) at the whole-brain level.	67
Figure 17. Brain models illustrate the field intensities of stimulation in V/m^{21} per mA. Darker red areas illustrate higher stimulation intensities compared with blue areas. The circled regions in the ‘left lateral’ pane show the PFA region and the TPA region, respectively.	85
Figure 18. Human and robot stimuli. The human agent is represented by a female face taken from the Karolinska Directed Emotional Faces (KDEF) data- base (F07; written informed consent from the Karolinska Institute was received to use the photograph for experimental investigations and illustrations). The robot agent is the robot EDDIE (developed at the Technical University of Munich, Germany).	86

Figure 19. Timing of the experimental procedure. The experiment started with participants completing the baseline gaze-cueing task. Next the researcher set up the tDCS machine and participants completed a questionnaire about their sensations. The participants then completed a decoy survey, which asked about their video- game experience. Participants then completed a second sensation questionnaire followed by a gaze-cueing task under stimulation. After the stimulation gaze cueing task was completed, a final sensation questionnaire was administered and the tDCS stimulation was stopped.	88
Figure 20. Sequence of events on a trial of gaze cueing. Participants first fixated on a fixation cross for 700–1000 ms and were then presented with the gazing agent (human versus robot) looking straight for 700 – 1000 ms, followed by a change in gaze direction (to either the left or right side of the screen). After an SOA of 400–600 ms, the target letter (F or T) appeared either where the face was looking or opposite to where the face was looking. The target remained on the screen until a response was given or a timeout of 1200 ms was reached.	92
Figure 21. Gaze-cueing effects (in ms) as a function of Brain site (left PFA, left TPA), Session (baseline, stimulation) and Agent type (human, robot). There was a significant change in gaze cueing for active PFA stimulation, with no differences in gaze cueing between human and robot at baseline, but significantly larger gaze- cueing effects for the human versus the robot agent under stimulation. Active TPA stimulation did not have significant effects on gaze cueing (*p , 0.05).	96

ABSTRACT

PHYSICAL APPEARANCE AS A TOP-DOWN INFLUENCER OF MIND PERCEPTION IN HUMAN-ROBOT SOCIAL ATTENTION

Abdulaziz Abubshait, M.A.

George Mason University, 2019

dissertation Director: Dr. Eva Wiese

When we interact with others, we use their nonverbal gestures to predict what they are going to do next. This process, termed mentalizing, allows us to engage in successful social interaction where responses to gestures are appropriate. For example, you can predict that someone is hungry when you see them looking at an apple pie. Mentalizing processes are associated with activity in a network in the brain that is responsible for processing social interactions. Although humans generally mentalize with only other humans, humans can also mentalize with inanimate objects (i.e., robots) when the objects are perceived to have thoughts and intentions of their own (i.e., perceiving a mind to them). This mind perception process has been well understood and although neural substrates have implicated the mentalizing brain network for mind perception processes in the brain, little is known about how engaging the mentalizing network through mind perception is related to socio-cognitive processes. In this thesis we use a social attention task, the gaze-cueing task, to examine subjects' performance on social attention when

they encounter agents that have varying degrees of physical humanness. The first study in this proposal examined how physical appearance can influence gaze-cueing performance and subjective mind ratings separately. Results showed that physical appearance only influenced mind ratings but not gaze-cueing performance. In the second study, we illustrated how physical humanness (i.e., a mind trigger) influences gaze-cueing performance as a function of brain activity related to mind ratings (i.e., BOLD response as it relates to subjective mind ratings) and subjective mind ratings. In the last study, we used tDCS to establish a causal link between brain activity in regions related to mind perception and gaze-cueing performance in a social attention task and found that only brain stimulation to prefrontal brain regions modulated social attention performance.

INTRODUCTION

The use of robotics and automation has been seeing a large increase on a day-to-day basis (Tapus & Matarić, 2006). These uses include human-machine teams, therapy and clinical uses, and communication robots such as delivery robots (Wiese, Metta, & Wykowska, 2017). Studies on these types of robots have found that they can have a positive influence on their interaction partner's attitudes, however, to inform the design of social robotics, research must understand the neural underpinnings of social interactions between humans and robots and how the social brain reacts to interactions with inanimate objects. By understanding how the brain reacts to different characteristics of robots, we are able to better inform the design of social robotics and build robots that can elicit natural successful social interactions from their human interaction partners. To do so, research must combine behavioral, subjective, and neuroscientific methods to better understand how and why social robots are successful at eliciting these social interactions and when do these social interactions between humans and robots breakdown (Wiese et al., 2017).

Mind perception and the social brain

When we socially interact with others, we use their nonverbal behaviors such as gaze behavior, gestures, and facial expressions to make predictions about what they want to do next (Frith & Frith, 2006a). This process of mentalizing allows us to infer thoughts, mental states, and intentions from people's actions, which makes us engage in successful social interactions (Baron-Cohen, 1997; Frith & Frith, 2006a). Research has shown that a

specialized network in the brain is responsible for these mentalizing processes termed the social brain (Adolphs, 1999; Van Overwalle, 2009). This brain network includes prefrontal structures such as the Anterior Cingulate Cortex (ACC; Gallagher, Jack, Roepstorff, & Frith, 2002), the Medial Prefrontal Cortex (mPFC; Bzdok et al., 2013), and posterior structures such as the Temporoparietal Junction (TPJ; Santiesteban, Banissy, Catmur, & Bird, 2015), Superior Temporal Sulcus (STS; Schweinberger, Kloth, & Jenkins, 2007), and Fusiform Gyrus (FG; Gobbini et al., 2011).

As noted previously, when humans interact with others, these brain regions are engaged. However, when humans interact with social robots these structures are underactivated (Cross et al., 2019), which poses a problem for human-robot social interactions as their behaviors are not being processed as social. Since socio-cognitive that involve interacting with others are related to social brain activation (Gallagher et al., 2002; McCabe, Houser, Ryan, Smith, & Trouard, 2001), not engaging these regions will negatively influence these processes.

Although this may seem like humans cannot engage socially with nonhuman actors such as robots, studies have shown that the degree to which this brain network is engaged depends on how much mind is perceived to the actor through its perceived agency (i.e., is the actor perceived to be able to manipulate the external world?) and experience (i.e., is the actor perceived to be able to have thoughts, mental states, beliefs, and intentions; Gray, Gray, & Wegner, 2007). Mind perception has been shown to be influenced by characteristics such as having human-like physical appearance, having

human-like physical movement, and being likeable (Waytz, Gray, Epley, & Wegner, 2010).

Not only does perceiving a mind to an agent engage the mentalizing network of the brain, mind perception can also have positive outcomes on both attitudes towards robots where robots are perceived as more trustworthy (Kiesler, Powers, Fussell, & Torrey, 2008), and socio-cognitive processing where interactions with nonhuman agents are improved when nonhuman agents exhibit the capability of producing socio-cognitive behavior like joint attention (Breazeal, Kidd, Thomaz, Hoffman, & Berlin, 2005; Looije, Neerinx, & Cnossen, 2010).

Joint attention

Of many socio-cognitive processes involved in social interactions, joint attention is a crucial process that allows an observer to shift their attention in a social manner by following another's gaze behavior (Baron-Cohen, Leslie, & Frith, 1985). Gaze-following behavior develops in infants of 3 months and has also been found in primates (Nummenmaa & Calder, 2009). In experimental design, social attention has been investigated via the gaze-cueing task in which participants view a gazer that establishes joint attention by looking directly at them. The gazer then shifts their eyes to either the right or left. Next, a target appears. Participants then discriminate whether the target was an "F" or a "T". If the target appeared in the same direction as the cue, the trial is considered a valid trial, while if the target appeared in the opposite direction of the cue, then it is considered an invalid trial (Friesen & Kingstone, 1998). To measure the strength

of gaze-cueing, a gaze-cueing effect is calculated by deducting valid trials from invalid trials.

Traditional interpretations of the findings conclude that gaze-following is purely reflexive such that faster response times were found for valid trials compared to invalid trials an observer's attention is directed to where the target would appear (Friesen & Kingstone, 1998). However, at the appropriate Stimulus Onset Asynchrony (i.e., SOA: duration of time from the onset of the gaze to the onset of the target), recent evidence have shown that top-down factors can influence the strength of attentional orienting such as similarity to self (Hung & Hunt, 2012), political affiliation (Liuzza et al., 2011), social status (Cui, Zhang, & Geng, 2014; M. Dalmaso, Pavan, Castelli, & Galfano, 2012; Shepherd, Deaner, & Platt, 2006), trustworthiness (Süßenbach & Schönbrodt, 2014; Takao & Ariga, 2016), familiarity (Deaner, Shepherd, & Platt, 2007), group membership (Pavan, Dalmaso, Galfano, & Castelli, 2011), and when context information is provided (Wiese, Zwickel, & Müller, 2013). More importantly for this proposal, mind perception has also shown to be an effective method to exert top-down control on gaze-cueing behavior (Wiese, Wykowska, & Müller, 2014). In their study, Wiese and colleagues used a belief manipulation of mind perception in which participants saw a human face and a robot face as gazers in the gaze-cueing experiment. When participants believed that the robot's eye movements were controlled by a human in another room, they found no differences in attentional orienting between the human face and the robot face. This suggests that our perceptions of a robot's mind status - in this case believing that its

behavior is intentional as it is controlled by a human - can influence how we react to them (Wiese, Wykowska, et al., 2014).

Mind perception, social attention, the social brain

The previously mentioned studies have shown that manipulating the physical appearance of a robot can successfully modulate subjective mind ratings and activation in the mentalizing network of the brain. Studies also have shown that manipulations of mind status (i.e., does the gazer have a mind?) can exert top-down modulation of attentional orienting to gaze-cues as manipulated by physical appearance and not actual subjective ratings. This raises a question of how attentional orienting to gaze-cues is influenced by physical appearance depending on subjective mind ratings if it is indeed modulated by top-down processing. Another important question is where the locus of this top-down processing is in the brain (i.e., which brain structure is responsible for exerting top-down influence in gaze-behavior). While many studies have examined the link between gaze behavior and mind perception and the link between gaze behavior and the social brain, not many studies have investigated how gaze behavior is influenced by mind perception as a function of deferential responses in the social brain (i.e., how is gaze behavior is influenced by mind perception as a function of deferential responses in the social brain).

Rationale

The current proposal aims to expand on the literature linking socio-cognitive processes to mind attribution of human-robot interaction and the respective brain structures that are related to these socio-cognitive processes. Specifically, we are

interested in investigating how brain activation that is related to mind perception can be manipulated via equipping robotic agents with mind triggering characteristics such as human-physical appearance and whether this manipulation of brain activity can influence performance in a social attention task. Previous studies that have been conducted in this proposal have shown that human-physical appearance is related to subjective mind perception, and that social brain structures that are involved with mind perception are linked to performance in a social attention task. Specifically, Study 1 aimed to determine if human physical appearance can influence mind perception. Findings suggested that human-physical appearance only influenced subjective mind perception and not attentional orienting. Since attentional orienting could be dependent on subjective mind ratings, Study 2 sought out to determine whether physical-human appearance influences gaze-cueing performance as determined by subjective mind ratings and activation in social brain structures when subjectively judging the mind status of different robot faces. The second study found that mind ratings as well as activation in prefrontal structures that are implicated in the mentalizing network of the brain (i.e., ventromedial Prefrontal Cortex; vmPFC) are related to gaze-cueing performance. Although these two studies illustrate how human-physical appearance can trigger mind perception as indicated by activation in the vmPFC, and that the vmPFC is related to gaze-cueing performance, a causal link between brain activation in the mentalizing brain network and gaze-cueing performance has yet to be determined. The proposed study intends to determine causality between the activity in prefrontal brain structures, which have been implicated in the social brain and attentional orienting to gaze cues. To do so, we aim to causally activate

these brain regions to increase the likelihood of perceiving a mind when examining agents of varying degrees of human physical appearance. Specifically, we propose a third study in which we causally manipulate brain activation in prefrontal structures via tDCS stimulation to determine if prefrontal structures are indeed causally responsible for successful social attention performance when we manipulate mind perception through mimicking human-physical appearance.

STUDY 1.

Abstract

Gaze following occurs automatically in social interactions, but the degree to which gaze is followed depends on whether an agent is perceived to have a mind, making its behavior socially more relevant for the interaction. Mind perception also modulates the attitudes we have toward others, and determines the degree of empathy, prosociality, and morality invested in social interactions. Seeing mind in others is not exclusive to human agents, but mind can also be ascribed to non-human agents like robots, as long as their appearance and/or behavior allows them to be perceived as intentional beings. Previous studies have shown that human appearance and reliable behavior induce mind perception to robot agents, and positively affect attitudes and performance in human–robot interaction. What has not been investigated so far is whether different triggers of mind perception have an independent or interactive effect on attitudes and performance in human–robot interaction. We examine this question by manipulating agent appearance (human vs. robot) and behavior (reliable vs. random) within the same paradigm and examine how congruent (human/reliable vs. robot/random) versus incongruent (human/random vs. robot/reliable) combinations of these triggers affect performance (i.e., gaze following) and attitudes (i.e., agent ratings) in human–robot interaction. The results show that both appearance and behavior affect human–robot interaction but that the two triggers seem to operate in isolation, with appearance more strongly impacting attitudes,

and behavior more strongly affecting performance. The implications of these findings for human–robot interaction are discussed.

Introduction

In social interactions, we use information from gestures, facial expression or gaze direction to make inferences about what others think, feel or intend to do (i.e., *mentalizing*; Adolphs, 1999; Emery, 2000; Gallagher and Frith, 2003). How we react to these cues is determined by how much social relevance we ascribe to them and, specifically, to what degree they are believed to originate from an entity with a mind, capable of having internal states like emotions or intentions (i.e., *mind perception*; Gray et al., 2007). Changes in gaze direction, for instance, are followed more strongly when they are displayed by a face with a fearful rather than a neutral expression (Graham et al., 2010), or when they are believed to be intentional rather than pre-programmed or random (Teufel et al., 2009; Wiese et al., 2012; Wykowska et al., 2014; Özdem et al., 2016). Seeing minds in others is not exclusive to human agents, but intentionality can also be ascribed to agents who do not have minds (i.e., robots) or whose mind status is ambiguous (i.e., animals; Gray et al., 2007).

In order to trigger mind perception, non-human entities need to display signs of intentionality via appearance (Kiesler et al., 2008; Looser and Wheatley, 2010; Admoni et al., 2011) and/or behavior (Morewedge, 2009; Waytz et al., 2010; Wiese et al., 2014). Entities that physically resemble humans are more likely to be perceived as ‘having a mind’ than agents with a mechanistic appearance, in particular when they display human facial features (DiSalvo et al., 2002; Kiesler et al., 2008; Tung, 2011). Entities without

human appearance can still trigger mind perception when their behavior is predictable (Morewedge, 2009; Pfeiffer et al., 2011), leads to negative outcomes (Waytz et al., 2010), or resembles movement patterns reminiscent of human–human interactions (Heider and Simmel, 1944; Abell et al., 2000; Castelli et al., 2000). Behavior is also interpreted as intentional when it is believed to be reliable (Süßenbach and Schönbrodt, 2014; Wiese et al., 2014) or to be generated by a human (Wiese et al., 2012; Wykowska et al., 2014; Özdem et al., 2016).

A positive effect of mind perception on attitudes and performance has also been observed in human–robot interaction (Sidner et al., 2004; Bennewitz et al., 2005; Mutlu et al., 2006, 2012; Fussell et al., 2008; Yamazaki et al., 2010; Huang and Thomaz, 2011; Staudte and Crocker, 2011; Pfeiffer-Lessmann et al., 2012). Robots that exhibit human gestures like shrugging or nodding, for instance, have a positive impact on emotional reactions and perceived trustworthiness (Kiesler et al., 2008; Carter et al., 2014), and robots displaying human behavior lead to improved performance on joint tasks (Breazeal et al., 2005; Looije et al., 2010; Pak et al., 2012; Waytz et al., 2014). In contrast, robots that do not trigger mind perception have negative effects on performance in social interactions (Wiese et al., 2012; Wykowska et al., 2014; Caruana et al., 2016; Özdem et al., 2016), and fail to induce social facilitation effects (Bartneck, 2003; Woods et al., 2005; Park and Catrambone, 2007; Riether et al., 2012).

While these studies suggest that mind perception in non-human agents (a) has a beneficial effect on attitudes and performance in human–robot interaction, and (b) can be triggered experimentally via appearance and/or behavior, no study to date has examined

how congruent (i.e., cue A and B both trigger or inhibit mind perception) versus incongruent (i.e., cue A/B triggers/inhibits mind perception) combinations of these triggers affect attitudes and performance in human– robot interaction. The current study addresses this question by manipulating the likelihood that mind is ascribed to non-human agents via appearance (high: human-like vs. low: robot-like) and behavior (high: predictable vs. low: random), and examining how congruent (human-like/reliable, robot-like/random) versus incongruent (human-like/random, robot-like/reliable) combinations of these triggers affect gaze following (i.e., performance measure) and agent ratings (i.e., attitude measure). Gaze following was picked as performance measure in the present experiment, since gaze direction is one of the most important cues in social interactions indicating another’s focus of interest, and a pre-requisite for more complex social-cognitive functions like mentalizing (Baron-Cohen, 1995; Frith and Frith, 2006).

When examining gaze following experimentally, a face is presented centrally on the screen that first gazes straight ahead, and then changes gaze direction to trigger shifts of the observer’s attention to the left or right side of the screen (i.e., *gaze cueing*; Friesen and Kingstone, 1998). This gaze cue is followed by the presentation of a target either at the cued location (i.e., valid trial) or an uncued location (i.e., invalid trial), with reactions to targets appearing at the cued location being faster than reactions to targets appearing at an uncued location (*gaze-cueing effect*; Friesen and Kingstone, 1998; Frischen et al., 2007). Positive effects of gaze cues have also been observed in human–robot interaction, where robots that shift their gaze during social interactions are perceived as more enjoyable than robots that do not shift their gaze (Kanda et al., 2001), and robots that

conjointly attend to where the human partner is looking are perceived as more competent than robots that do not engage in joint attention (Huang and Thomaz, 2011). Robot gaze also helps performance on joint human–robot tasks, for instance, by improving the accuracy of predictions in an object selection game (Mutlu et al., 2009), or by improving recollection in a memory task by gazing at relevant objects (Mutlu et al., 2006).

Attentional orienting to gaze cues has traditionally been thought of as a bottom–up process that is observable in infants as young as 3 months of age (Hood et al., 1998), and can be triggered by any kind of stimulus with eye-like configurations (Friesen and Kingstone, 1998; Langton and Bruce, 1999; Quadflieg et al., 2004). Confirming its reflexive nature, gaze following cannot be suppressed even when gaze direction is unlikely to predict the location of a target (Friesen et al., 2004; Vecera and Rizzo, 2006), and is not modulated by the gazer’s *animacy* (Quadflieg et al., 2004), *familiarity* (Frischen and Tipper, 2004), *facial expression* (Hietanen and Leppänen, 2003; Bayliss et al., 2007), or *trustworthiness* (Bayliss and Tipper, 2006). The few modulatory effects that were originally reported in the context of gaze following strongly depended on age (i.e., stronger gaze following in children; Hori et al., 2005), and individual traits (i.e., stronger gaze cueing in highly anxious individuals; Tipples, 2006; Fox et al., 2007).

More recently, however, studies have shown that gaze following *can* be top–down modulated when gaze behavior is embedded in a context that enhances its social relevance for the observer (Tipples, 2006; Fox et al., 2007; Bonifacci et al., 2008; Graham et al., 2010; Kawai, 2011; Hungr and Hunt, 2012; Süßenbach and Schönbrodt, 2014; Wiese et al., 2014; Wykowska et al., 2014; Cazzato et al., 2015; Dalmaso et al.,

2016). Using this updated version of the original gaze-cueing paradigm, researchers were able to show that variables like *similarity-to-self* (Hung and Hunt, 2012; Porciello et al., 2014), *physical humanness* (Admoni et al., 2011; Martini et al., 2015), *facial expression* (Bonifacci et al., 2008; Graham et al., 2010), *social status* (Jones et al., 2010; Dalmaso et al., 2012, 2014, 2015, 2016; Ohlsen et al., 2013), *membership in ingroup* (Dodd et al., 2011, 2016; Liuzza et al., 2011; Pavan et al., 2011; Ciardo et al., 2014; Cazzato et al., 2015; Dalmaso et al., 2015), or *familiarity* (Frischen and Tipper, 2006; Deaner et al., 2007) are able to modulate the degree to which gaze is followed by increasing or decreasing its social relevance.

Believing that an agent is intentional rather than pre-programmed is another factor that can increase the social relevance of observed behavior, with the effect that malevolent actions believed to be intentional are experienced more intensely (Gilbert et al., 2004; Gray and Wegner, 2008), and judged more harshly (Ohtsubo, 2007; Cushman, 2008) than unintentional ones. Similarly, believing that changes in gaze direction are intentional versus unintentional increases the degree to which they are followed (Teufel et al., 2009; Wiese et al., 2012, 2014; Wykowska et al., 2014), and positively affects how the gazer is evaluated (Bayliss and Tipper, 2006). Altogether, these studies indicate that perceiving robots as agents with a mind and the ability to execute intentional actions has the potential to positively impact performance and attitudes in human–robot interaction. What is still unclear is, which agent features most effectively trigger mind perception and how attitudes and performance in human–robot interaction are affected when two triggers, like appearance and behavior, are in conflict. The effect of conflicting agent

features on attitudes and performance, however, is an important issue in human–robot interaction since a subset of contemporary robots either display human appearance or intentional behavior, but usually not both (Fong et al., 2003).

In the current experiment, we examine how behavior and appearance interact in triggering mind perception, and measure how social-cognitive performance (i.e., gaze following) and agent ratings (i.e., judgments of mind status) are affected in congruent versus incongruent conditions. Based on previous studies, we expected that reliable gaze behavior (i.e., cue predicts target location in 80% of trials) and human-like appearance (i.e., 80% physical humanness) would increase the likelihood that mind is perceived in artificial agents, while random gaze behavior (i.e., cue predicts target location in 50% of trials) and robot-like appearance (i.e., 20% physical humanness) were expected to decrease the likelihood for mind perception; see **Figure 1**.

Methods and materials

Participants

Eighty-six undergraduate students at George Mason University were originally recruited for the experiment. The data of 23 participants had to be excluded from analysis since they did not meet the *a priori* accuracy cut off of 90%; the data of the remaining 63 participants was analyzed (47 females, *M* age: 21, *SD* = 3.3, 10 left handed). Participants were recruited using the participant management website SONA Systems at George Mason University. Participants were randomly assigned to one of the two reliability conditions (i.e., human behavior: 80% reliable vs. robot behavior: 50% reliable), with 32 participants (24 females, *M* age: 20.6, *SD* = 3.9, three left-handed) in the 80% reliability

condition and 31 participants (23 females, mean age: 19.7, $SD = 2.5$, six left-handed) in the 50% reliability condition. Approval by the Internal Review Board (IRB) was obtained prior to data collection. Participant data was collected according to George Mason University's ethics committee. All participants gave informed consent, and reported normal to corrected-to-normal vision. Participant data was stored anonymously according to IRB guidelines. Testing time was about 30 min.

Apparatus

Stimuli were presented on a 19-inch ASUS VB Series VB198T-P monitor with the refresh rate set at 85 Hz. RT measures were based on standard keyboard responses. Participants were seated approximately 57 cm from the monitor, and the experimenter ensured that participants were centered with respect to the monitor. The experiment was programmed using the software *Experiment Builder* (SR Research, Ltd., Ottawa, ON, Canada).

Stimuli

Images of two agents were used for the gaze-cueing task: a robot-like agent and a human-like agent. The agent images were created by morphing a human face (i.e., male face from the Karolinska Institute database; Lundqvist et al., 1998) into a robot face (i.e., Meka S2 robot head) in steps of 10% using the software Fantamorph. Out of the morphing spectrum, the morph with 80% physical humanness was used as human-like gazer and the morph with 20% physical humanness as robot-like gazer. The left-and rightward gazing faces were created using Photoshop by shifting the irises and pupils in the eyes of the original faces until they deviated 0.4° from direct gaze, which was then

followed by another round of morphing for the left- and the rightward gazing faces separately. As a last step, Gimp was used for all images to touch up any minor imperfections in images and to make the sequencing of the images smooth. The face stimuli were 6.4° wide and 10.0° high on the screen, depicted on a white background and presented in full frontal orientation with eyes positioned on the central horizontal axis of the screen; see **Figure 1**.

The target stimuli for the gaze-cueing procedure were black capital letters (F or T), measuring 0.8° in width and 1.3° in height. Targets appeared on the horizontal axis, and were located 6.0° from the center of the screen. Targets appeared at the gazed-at location in 80% of the trials in the reliable condition (i.e., gaze direction predictive of target location), and in 50% of the trials in the random condition (i.e., gaze direction non-predictive of target location); see **Figure 1**.





		Behavior	
		Reliable	Random
Appearance	Robot	 80% T	 50% T
	Human	 80% T	 50% T

Figure 1. Manipulation of mind judgments. Human-like appearance (80% physical humanness) and reliable behavior (80% predictive cueing) should increase the likelihood that mind is attributed, while robot-like appearance (20% physical humanness) and random behavior (50% predictive cueing) should decrease the likelihood that mind is attributed to an agent.

Procedure

At the beginning of the session, participants gave informed consent and were randomly assigned to one of two reliability conditions (80% vs. 50%). They were then told that they would perform a gaze following task together with two different agents (introduced via images), which required discriminating target letters (F or T) by pressing one of two response keys: for half of the participants, F was assigned to the “D” key and T to the “K” key of the keyboard; for the other half of the participants, stimulus-response

mapping was reversed. The original key labels on the keyboard were covered with stickers to prevent letter interference effects. Participants were informed that agent gaze either validly or invalidly cued the location of the target, and were told that the experiment started with a practice block consisting of 20 trials, followed by two experimental blocks of gaze following (one per agent). They were instructed to fix their gaze on a centrally presented fixation cross at the beginning of each trial and to remain fixated until the trial was over. After the fixation cross, the image of one of the agents would appear in the center of the screen, which would then shift its gaze left- or rightward to either validly or invalidly cue the location of the target. Participants were asked to respond as quickly and accurately as possible to the identity of the target letter as soon as it appeared on the screen. In addition to gaze following, participants were also instructed to rate the agents regarding their capability of having a mind (i.e., “Do you think this agent has a mind?”) on an eight-point Likert-scale, once at the beginning and the end of each block (all instructions were given in written form). This question was used in the current experiment to be consistent with previous literature that operationalized mind perception as the degree to which agents were judged as having a mind as a function of their physical humanness (Looser and Wheatley, 2010; Hackel et al., 2014; Martini et al., 2016). Although this question has been commonly used in the literature to assess mind perception, we would like to point out that it most likely does not measure *perceptions* of mind (i.e., actually seeing mind in the agent), but more likely measures *judgments* of mind (i.e., how similar does this agent look to agents that have a mind). In consequence, ratings probably do not reflect the degree to which participants

thought the depicted agents actually *have* minds, but more likely reflect how similar they thought the agents *looked* to human agents (leaving aside that they do not actually have a mind). To account for this, the results of the agent ratings will be referred to as *mind judgments*.

Figure 2 illustrates the sequence of events on a given trial of gaze cueing: the trial started with the presentation of a fixation cross in the center of the screen for a random time interval of 700–1000 ms. Afterward, one of the agents appeared in the center of the screen with straight gaze for a random time interval of 700–1000 ms. The agent then changed gaze direction either looking to the left or the right, followed by the appearance of one of the two target letters either at the valid or invalid location after a stimulus onset asynchrony (SOA) of 500 ms. Agent and target remained on the screen until a response was given or a time-out of 1200 ms was reached, whichever came first. At the end of each trial, a blank screen was presented for an inter-trial interval (ITI) of 680 ms before the next trial started. Each session of the experiment was composed of 340 trials total, with a block of 20 practice trials preceding two experimental blocks of 160 trials each (one block per agent). The order in which the blocks were presented was counterbalanced across participants. Gaze direction (left, right), target side (left, right), target identity (F, T) and agent were selected pseudo-randomly and every combination appeared with equal frequency. Gaze validity was calculated based on the combination of gaze direction and target direction: on valid trials, the target appeared where the face was looking, while on invalid trials the target appeared opposite of where the face was looking. In the random condition, valid and invalid trials appeared with equal frequencies (i.e., 80 valid trials and

80 invalid trials per agent), whereas in the reliable condition, 80% of the trials were valid and 20% invalid (i.e., 128 valid trials and 32 invalid trials per agent); agent reliability was manipulated between participants. At the beginning and the end of each agent block, participants were asked to rate the agent's capability of having a mind. For this purpose, the image of the respective agent was presented with a eight-point Likert scale presented underneath and participants were instructed to type in the number rating they wanted to assign to the agent into a free response box on the screen. No information about the actual reliability of the agents was disclosed at any time during the experiment.

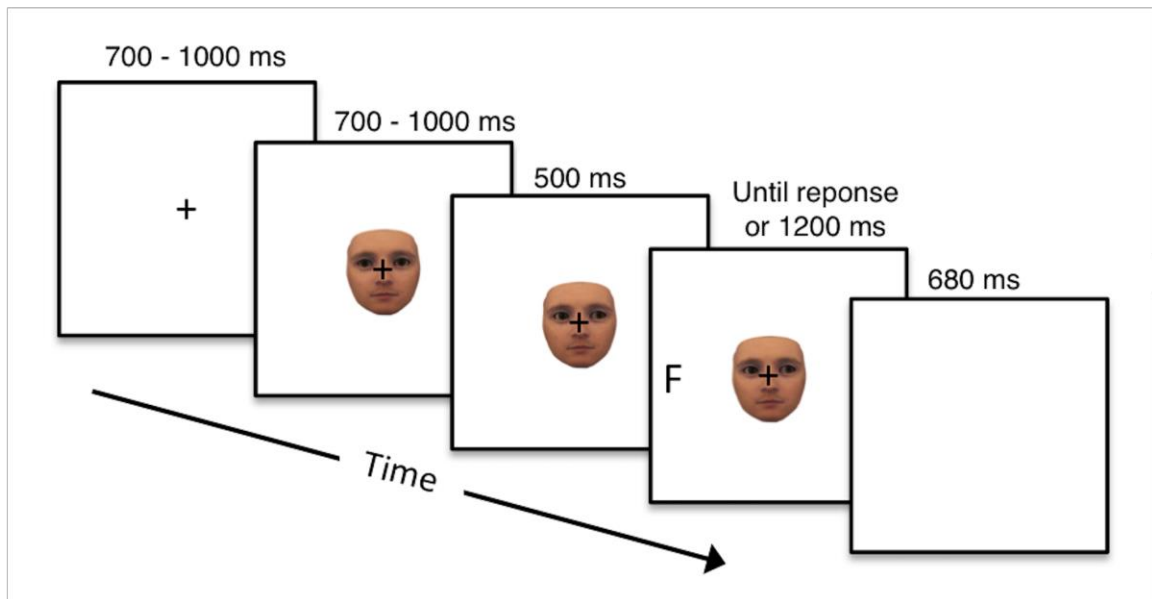


Figure 2. Sequence of events on a trial of gaze cueing. Participants first fixated on a fixation cross for 700–1000 ms and were then presented with an agent (human vs. robot) looking straight for 700–1000 ms, followed by a change in gaze direction (either to the left or right side of the screen). After a SOA of 500 ms, the target letter (F or T) appeared either where the face was looking (valid) or opposite of where the face was looking (invalid). The

target remained on the screen until a response was given or a timeout of 1200 ms was reached. A blank screen marked the end of the trial and was presented for 680 ms.

Analysis

Data was analyzed using R 3.2.4. Misses and incorrect responses, as well as data from participants with an accuracy rate below of 90% were removed prior to analyses (27% of trials). The data was analyzed with regard to the combined effect of appearance and behavior on (a) social-cognitive performance as measured in gaze-cueing effects and (b) agent ratings as measured in the degree to which mind was attributed to the agents. Gaze-cueing effects were calculated by subtracting the average reaction time for valid trials from the average reaction time for invalid trials (per participant, for agent and reliability conditions separately), and subjected to a 2×2 ANOVA with the within-factor Appearance (robot-like vs. human-like) and the between-factor Behavior (random vs. reliable). The more positive the difference score, the more strongly participants followed the gaze of the agent.

To examine how exposure to different appearances and behaviors changed the participants' attitudes toward the agents, we calculated three mixed 2×2 ANOVAs with the within-factor Appearance (robot- vs. human-like) and the between-factor Behavior (random vs. reliable), and pre-interaction ratings, post-interaction ratings and difference scores between pre- and post-interaction ratings as dependent variables. A positive difference score between pre- and post-ratings reflects an increase in mind ratings after completing the gaze-cueing task (i.e., agent is perceived as more mindful after the interaction), while a negative difference score reflects a decrease in mind ratings after

completing the task (i.e., agent is perceived as less mindful after the interaction). The higher the agent ratings at the pre- and post-interaction stage, the more willing participants were to ascribe mind to the gazer.

Results

The results of the analysis of the *gaze-cueing* data are shown in **Figure 3**. The 2×2 ANOVA revealed a main effect of Behavior [$F(1,61) = 5.33, p = 0.024, \eta^2_p = 0.04$], with larger cueing effects for reliable versus random gaze behavior (26.4 ms vs. 15.4 ms). The main effect of Appearance was not significant [$F(1,61) = 0.18, p = 0.67, \eta^2_p = 0.001$], suggesting that the gaze of the human-like agent was not followed more strongly than the gaze of the robot-like agent. The interaction effect between Appearance and Behavior was also not significant [$F(1,61) = 0.02, p = 0.89, \eta^2_p < 0.001$], suggesting that the reliability with which the agent indicated target location influenced gaze following to the same degree for the human- and the robot-like gazer.

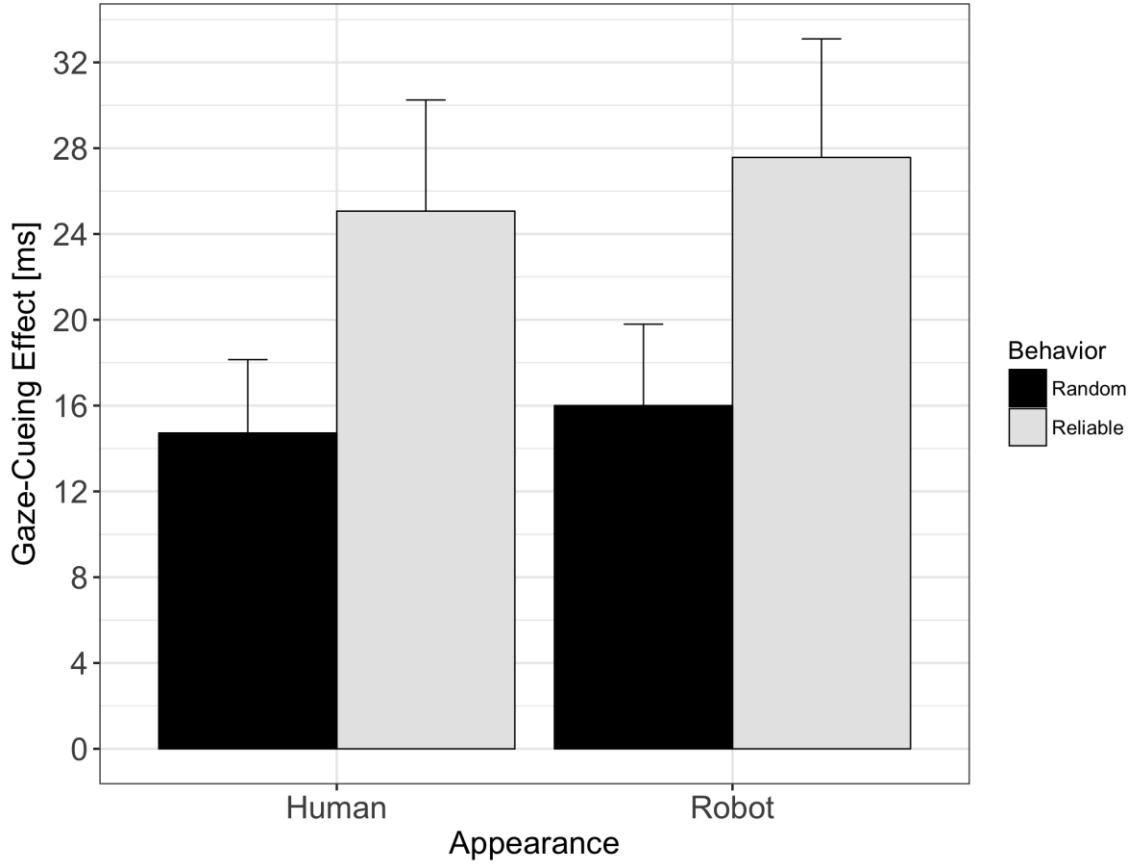


Figure 3. Gaze-cueing effects as a function of appearance and behavior. Reliable agents induced significantly larger gaze-cueing effects than agents showing random behavior, independent of appearance. There was neither a significant main effect of appearance on gaze-cueing effects, nor was there a significant interaction effect between appearance and behavior.

The results of the analysis of the *mind judgments* are shown in **Figures 4, 5**. The 2×2 ANOVA of the pre-interaction ratings revealed a significant main effect of Appearance [$F(1,61) = 116.07, p < 0.001, \eta^2_p = 0.34$], with higher agent ratings for the human- than the robot-like agent (6.18 vs. 3.14). Neither the main effect of Behavior [$F(1,61) = 2.39, p = 0.12, \eta^2_p = 0.02$], nor the interaction effect of Appearance and Behavior were significant [$F(1,61) = 3.43, p = 0.06, \eta^2_p = 0.01$], indicating that there

was no difference in the degree to which mind was attributed to the agents between reliability conditions prior to gaze following; see **Figure 4A**. The 2×2 ANOVA at the post-interaction stage showed a significant main effect of Appearance [$F(1,61) = 38.95, p < 0.001, \eta^2_p = 0.1$], with higher ratings for the human- than the robot-like agent (5.29 vs. 3.63). Neither the main effect of Behavior [$F(1,61) = 1.14, p = 0.28, \eta^2_p = 0.01$], nor the interaction effect of Appearance and Behavior [$F(1,61) = 3.44, p = 0.06, \eta^2_p = 0.01$] were significant, showing that the agents' reliability during gaze following did not influence how much mind was attributed toward them; see **Figure 4B**. The effect of Appearance on post-ratings was further modulated by participant gender with significantly lower ratings for the human-like agent by male participants than female participants [$F(1,57) = 4.02, p = 0.04, \eta^2_p = 0.05$].

The 2×2 ANOVA of the difference scores between pre-and post-interaction ratings revealed a significant main effect of Appearance [$F(1,61) = 25.13, p < 0.001, \eta^2_p = 0.14$], with a decrease in mind ratings for the human-like agent ($\square 80\% = -1$), and a slight increase in mind ratings for the robot-like agent ($\square 20\% = +0.5$). Neither the main effect of Behavior [$F(1,61) = 0.16, p = 0.69, \eta^2_p < 0.01$], nor the interaction effect of Appearance and Behavior were significant [$F(1,61) = 0.02, p = 0.89, \eta^2_p < 0.01$], showing that the agents' behavior during gaze following did not affect how their mind status was rated; see **Figure 5**. The effect of appearance on changes in ratings from the pre- to post-interaction stage was further modulated by participant gender, with a

significantly more negative change in ratings for the human-like agent for male than female participants [$F(1,57) = 6.6, p = 0.01, \eta^2_p = 6.26$].

Discussion

The goal of the current experiment was to examine whether appearance and behavior interact in their ability to trigger mind perception to non-human agents, and if so, how congruent (human-like/reliable vs. robot-like/random) versus incongruent (human-like/random vs. robot-like/reliable) combinations of these triggers affect social-cognitive performance (i.e., gaze following) and agent ratings (i.e., do you think the agent has a mind?). Based on previous studies, reliable gaze behavior and human-like appearance were expected to *increase* the likelihood that mind is perceived in artificial agents, while random gaze behavior and robot-like appearance was expected to *decrease* the likelihood for mind perception. To investigate whether and how these two triggers for mind perception interact, appearance and behavior were both manipulated within a gaze following paradigm, where either a human- or robot-like agent reliably or randomly cued the location of an upcoming target. If mind perception played a role for social-cognitive performance, gaze following should be stronger in conditions where mind was likely to be attributed to the gazer (i.e., human-like appearance, reliable behavior) compared to conditions where mind attribution was not likely (i.e., robot-like appearance, random behavior). Likewise, if appearance and behavior affected how agents were rated, more mind status should be attributed to them in conditions where mind perception was likely compared to conditions where it was unlikely.

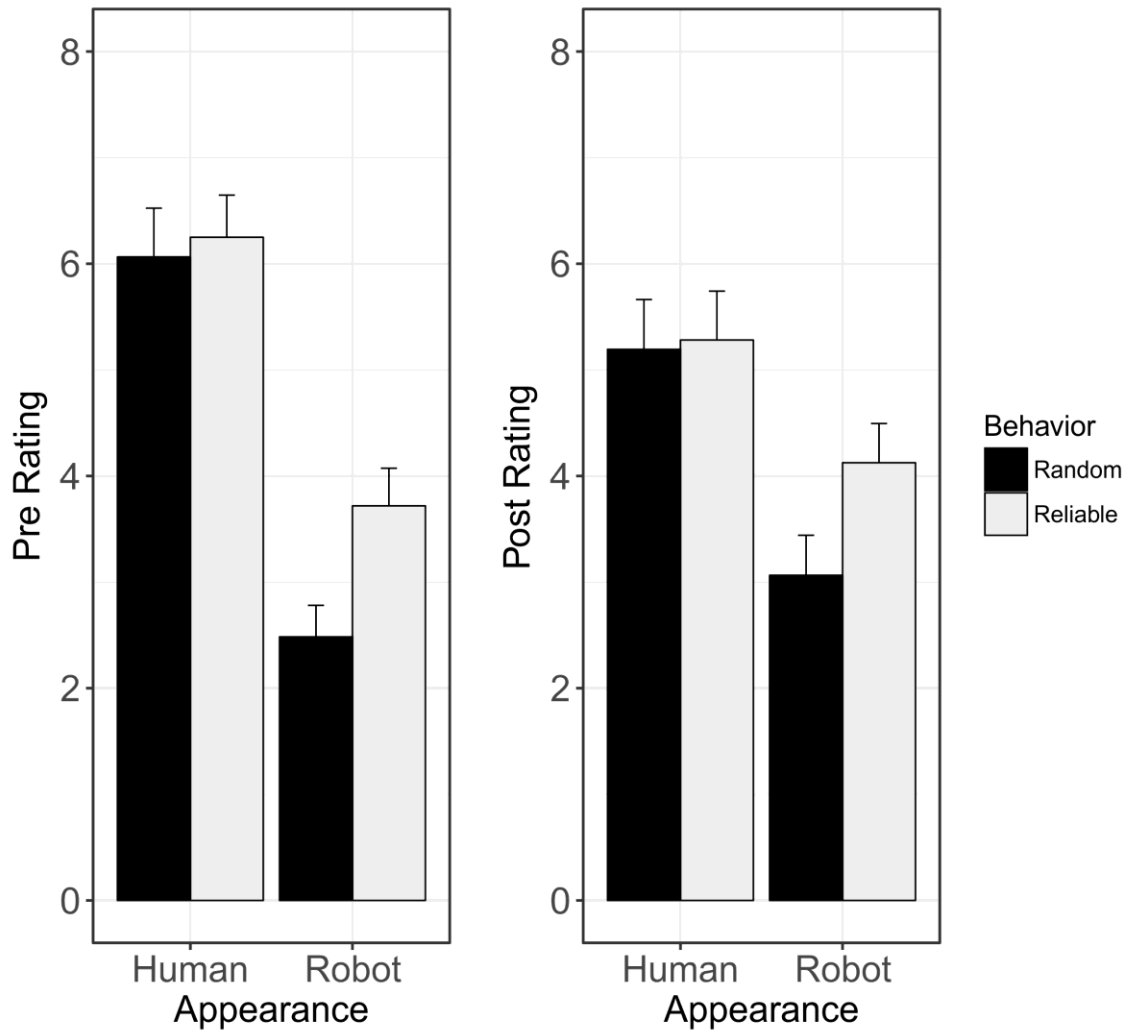


Figure 4. Mind judgments before and after gaze following. Agent appearance affected mind ratings both before interacting with the agents during the gaze following task (A) and afterward (B), with higher agent ratings for the human than for the robot agent. The reliability with which agents cued the target location did not have an effect on agent ratings, neither at the pre-interaction stage (i.e., no baseline difference in agent ratings between participants in the reliable and the random condition), nor at the post-interaction stage (i.e., knowing about the reliability of the agents did not influence the degree to which mind was ascribed to them). The interaction effect was not significant for either of the two rating times.

The results show that agent behavior but not appearance affected gaze following, while agent appearance but not behavior affected mind judgments: gaze was followed more strongly in conditions where the agents showed reliable versus random gaze

behavior, but this perception of reliability did not affect how much mind was ascribed to the agents after gaze following. In contrast, agent appearance did not have an impact on how strongly agent gaze was followed, but exclusively influenced mind attribution to the agents. Importantly, the positive effect of human appearance on mind ratings was observable both before and after participants interacted with the agent images during gaze following, and was not modulated by the reliability with which the agents cued an upcoming target location. Interestingly, however, the observed positivity bias caused by human-like appearance at first encounter seemed to fade over time (i.e., mind ratings for the human-like agent decreased between pre- and post-testing), while mind judgments for the robot-like agent increased from pre- to post-interaction ratings.

The observation that appearance and behavior influence how we interact with non-human agents is in line with previous reports showing that the two variables affect agent ratings (Looser and Wheatley, 2010; Waytz et al., 2010; Hackel et al., 2014; Martini et al., 2016), and performance (Kiesler et al., 2008; Morewedge, 2009; Süßenbach and Schönbrodt, 2014; Wiese et al., 2014; Mandell et al., 2015). Surprisingly, however, the current study shows that appearance and behavior differ significantly in their capacity to modulate performance versus mind judgments, with appearance having a stronger impact on agent ratings and behavior having a stronger impact on performance. This finding can be interpreted in two ways: first, it might indicate that judging one's

mind status is a qualitative rather than quantitative process, where agents either get mind status or no mind status ascribed, but nothing in between. If that were to be the case, it is possible that participants base their decision of whether an agent has a mind on just one mind trigger and ignore dissonant information from additional triggers to reduce potential cognitive conflicts. This interpretation is in line with previous studies showing that mind perception follows a qualitative pattern (i.e., significant increase in mind perception only after a certain threshold is passed; Cheetham et al., 2014; Hackel et al., 2014; Martini et al., 2016), and that conflicting information as to whether an agent has a mind or not has the potential to induce a cognitive conflict (Mandell et al., 2017; Weis and Wiese, 2017). Alternatively, the results could also indicate that mind perception is not a unified process that affects performance and attitudes in human–robot interaction in the same way, but instead that behavioral cues matter more in situations when participants actively interact with a robot agent, while physical cues have a stronger weight when making judgments about specific agent traits outside an interactive scenario. If that were to be the case, social roboticists would have to accentuate a robot’s perceived intentionality via behavioral cues when the robot’s main purpose is to engage in joint actions with human partners, as opposed to via physical cues when the focus of the interaction is on the robot’s personality.

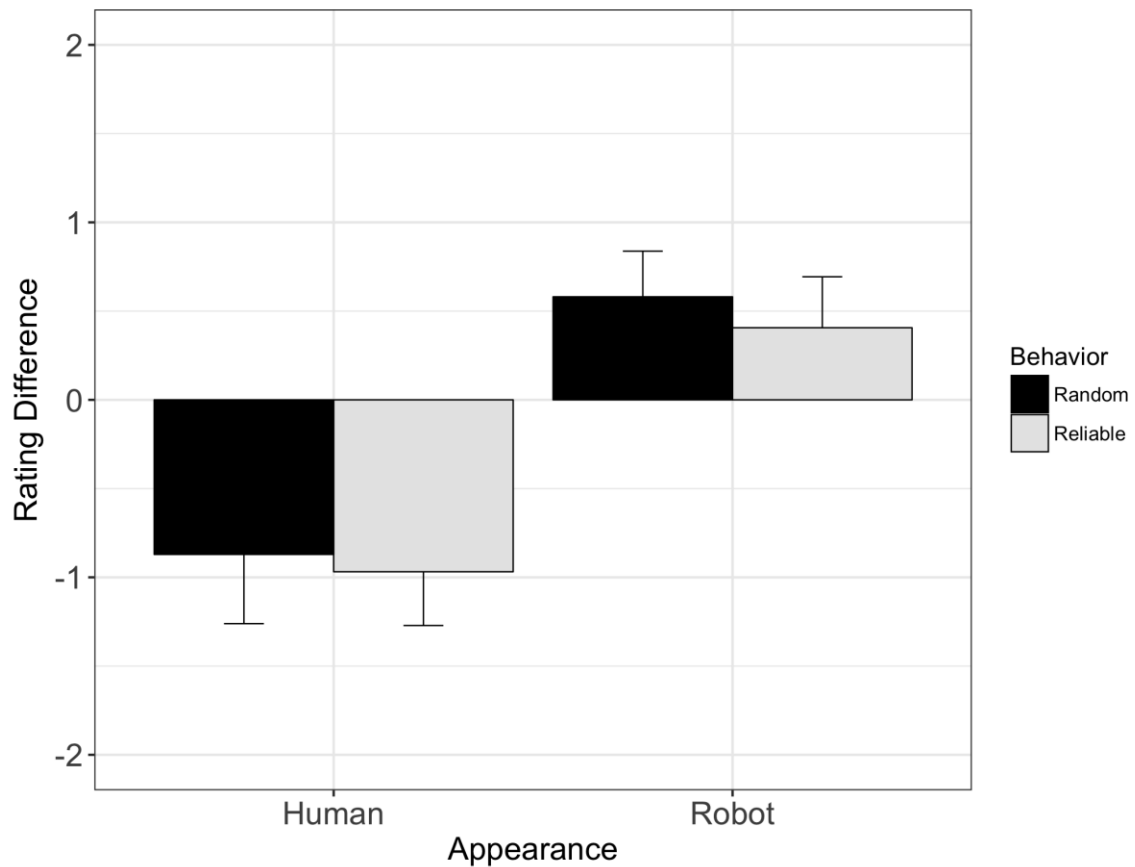


Figure 5. Change in mind judgments from pre- to post-interaction. Mind ratings for the human-like agent decreased significantly during gaze following, while mind ratings for the robot-like agent increased during gaze following. Interestingly, this effect is independent of the reliability with which the agents predicted the target location during gaze following.

Another unexpected observation in the current experiment was that mind ratings for the human-like agent decreased over time, while ratings for the robot-like agent slightly increased over time (both independent of reliability). With regard to the increase in ratings for the robot-like agent, it is possible that its mechanistic appearance might have primed participants to expect it to behave like a machine, incapable of engaging in social interactions. When they then experienced it sending social signals during the gaze

following task, participants might have been positively surprised by the agent's socialness, which in turn might have led to an increase in mind ratings. With regard to the decrease in ratings for the humanoid agent, it is possible that participants perceived a mismatch between its human-like appearance on the one hand and its mechanistic eye-movements on the other hand, with potentially negative effects on mind ratings. In gaze following paradigms, the impression of eye movements is caused by first presenting an agent looking straight and then, after a predefined time interval, the same agent looking to the side. Although this manipulation is effective in inducing shifts of attention to the gazed-at location, the eye movements usually do not match the biological motion patterns prototypically seen in human gazers. Since humans are quite sensitive to distinguishing biological motion from non-biological motion patterns (MacDorman and Ishiguro, 2006; Kätsyri et al., 2015; Wykowska et al., 2017), it is possible that perceiving a mismatch between human appearance and non-human motion might have triggered feelings of discomfort (MacDorman and Ishiguro, 2006; Saygin et al., 2012; Kätsyri et al., 2015), and therefore led to a decrease in mind ratings for the humanoid agent. The change in ratings from pre- to post-interaction was also more pronounced in male than in female participants, pointing at potential gender differences in perceiving mind in non-human agents.

The findings have several implications for the role perceptions of intentionality play in human–robot interaction. First, and foremost, the current study shows that expecting a robot agent to behave like an intentional being modulated attitudes and performance in human–robot interaction, and designing robots that trigger mind

perception should therefore be an important goal to social roboticists. Second, although previous research has identified physical and behavioral factors that trigger mind perception in isolation, it seems like these triggers do not modulate attitudes and performance in human–robot interaction to the same extent. Rather, it seems that human-like behavior has a stronger impact on performance in human–robot interaction, while human-like appearance matters more when rating an agent regarding stable traits, such as ‘having a mind.’ Third, the current experiment shows that although human appearance has a positive effect on attitudes at first encounter, its effect seems to be short-lived and have detrimental consequences on human–robot interaction if the positive expectation caused by a robot’s appearance (i.e., agent behaves like a human) is not met by its actual behavior (i.e., agent does not behave human-like).

There are some limitations related to the current experiment. First, it is not clear to what extent the gender of the participant and the gender of the gazer (i.e., only a white male face was used as a basis for the morphed stimuli) had an influence on the reported results. While gaze-cueing effects and ratings at the pre-interaction stage were not influenced by participant gender, the post-interaction ratings and, in consequence, the changes in ratings over time, were modulated by participant gender, with a stronger decline in agent ratings for the human-like agent for male than female participants. Whether this effect is due to gender differences in mind perception or due to systematic biases of the current experimental setup cannot be determined based on the current data. Similarly, it is unclear whether using a wider range of gazing stimuli (i.e., different gender, age, ethnic background) would change the pattern of results reported in this

paper. While it is common sense to control for perceptual features of the gazer by just using one gazing stimulus (i.e., of a particular gender, age and ethnic background) in gaze-cueing experiments (e.g., Hori et al., 2005; Bonifacci et al., 2008; Wiese et al., 2012; Wykowska et al., 2014; Graham et al., 2010), it cannot be ruled out completely that diversifying the features of the gazer might change the effects on mind ratings and gaze following reported in the current experiment. Second, we cannot fully rule out that changes in mind ratings from pre- to post-interaction are not simply due to a pragmatic effect, related to the fact that participants had to answer the same question (i.e., “Do you think the agent has a mind?”) twice, potentially suggesting to participants that the in-between manipulation was supposed to change their initial response. While this explanation is certainly possible, we do not believe that it is very likely, since asking the same question twice influenced participant answers differently for the human-like agent and the robot-like agent, with a decrease in ratings for the former and an increase in ratings for the latter. If asking the mind-rating question twice systematically impacted mind ratings in the current experiment, we would expect to see a similar effect for both the human-like agent and the robot-like agent. Since ratings do not change in the same way in both conditions, we believe that the observed changes in ratings are unlikely the result of a pragmatic effect. Third, asking participants whether they think an agent has a mind, might not actually measure *perceptions* of mind, but rather *judgments* of mind, that is: the reported ratings might not reflect the degree to which participants thought the depicted agents actually had minds, but rather how similar they thought they were to agents with mind (i.e., humans). While this limitation does not affect the general

observation that mind ratings are affected by agent appearance, it might overestimate the degree to which participants actually see non-human agents as having a mind. Future studies need to address this issue by being more specific about whether they investigate mind perception or mind judgments.

STUDY 2.

Abstract

In social interactions, we rely on nonverbal cues like gaze direction to understand the behavior of others. How we react to these cues is affected by whether they are believed to originate from an entity with a mind, capable of having internal states (i.e., mind perception). While prior work has established a set of neural regions linked to social-cognitive processes like mind perception, the degree to which activation within this network relates to performance in subsequent social-cognitive tasks remains unclear. In the current study, participants performed a mind perception task (i.e., judging the likelihood that faces, varying in physical human- likeness, have internal states) while event-related fMRI was collected. Afterwards, participants performed a social attention task outside the scanner, during which they were cued by the gaze of the same faces that they previously judged within the mind perception task. Parametric analyses of the fMRI data revealed that activity within ventromedial prefrontal cortex (vmPFC) was related to both mind ratings inside the scanner and gaze-cueing performance outside the scanner. In addition, other social brain regions were related to gaze-cueing performance, including frontal areas like the left insula, dorsolateral prefrontal cortex, and inferior frontal gyrus, as well as temporal areas like the left temporo-parietal junction and bilateral temporal gyri. The findings suggest that functions subserved by the vmPFC are relevant to both mind perception and social attention, implicating a role of vmPFC in the top-down modulation of low-level social-cognitive processes.

Introduction

Engaging in social interactions requires the ability to infer internal states of others, such as beliefs, intentions, and emotions (mentalizing; Baron-Cohen, 1997), and to use this information to predict their behavior (C. D. Frith & Frith, 2006). The primate brain is equipped with neural networks specialized in processing social information (social brain; Adolphs, 2009), responsible for making inferences about internal states and understanding the goals that underlie observed actions (Brothers, 2002; Bzdok et al., 2013; C. D. Frith & Frith, 2006; Van Overwalle, 2009; Van Overwalle & Baetens, 2009). Activation within social brain areas is modulated by the degree to which others are perceived as "having a mind" (Spunt, Meyer, & Lieberman, 2015) and the ability to experience internal states and execute goal-directed actions (mind perception; H. M. Gray, Gray, & Wegner, 2007). Mind perception is not exclusive to interactions with human agents; it can also be triggered in social interactions with nonhuman entities like animals or robots, as long as their behavior and/or appearance evoke associations with humanness (anthropomorphism; Abell, Happé, & Frith, 2000; Castelli, Happé, Frith, & Frith, 2000; DiSalvo, Gemperle, Forlizzi, & Kiesler, 2002; Kiesler, Powers, Fussell, & Torrey, 2008; Looser & Wheatley, 2010; Pfeiffer, Timmermans, Bente, Vogeley, & Schilbach, 2011; Waytz, Gray, Epley, & Wegner, 2010). Agents that do not trigger mind perception recruit social brain areas less than agents believed to have a mind (Gallagher, Jack, Roepstorff, & Frith, 2002; Harris & Fiske, 2006; Krach et al., 2008; Özdem et al., 2016; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; Waytz, Gray, et al., 2010), and have a negative impact on performance during social interactions (Caruana,

McArthur, Woolgar, & Brock, 2016; Wiese, Wykowska, Zwickel, & Müller, 2012; Wykowska, Wiese, Prosser, & Müller, 2014). What has not been investigated so far is whether the degree to which mind perception activates social brain areas is directly related to human performance during social-cognitive tasks. To address this question, the current experiment employed parametric analyses of fMRI data to relate brain activation during a mind perception task (i.e., judging the likelihood that agents, varying in physical human-likeness, have internal states) with performance on a separate social attention task (i.e., attentional orienting to agents' gaze cues).

We expect networks that are activated during mind perception and social attention to be located in the social brain network, consisting of the action perception system (APS) involved in understanding the goals underlying observed actions, and the mentalizing system (MS) involved in inferring others' internal states (Adolphs, 2009). The APS consists of a distributed network of temporal areas like the extrastriate body area (EBA) and posterior superior temporal sulcus (pSTS), as well as parietal and frontal areas like the inferior parietal cortex (IPC) and ventral premotor cortex (vPMC); the temporal areas are thought to detect the presence of intentional agents and label their actions as goal-directed based on observed motion patterns, while the parietal and frontal areas are believed to identify particular goals underlying these actions (e.g., "What is the outcome of an action?"; Becchio, Adenzato, & Bara, 2006; Grafton & Hamilton, 2007; Pobric & Hamilton, 2006; Saxe, 2006; Saygin, 2007; Saygin, Wilson, Hagler, Bates, & Sereno, 2004). Action understanding in the primate brain is based on the principles of resonance, where shared representations are activated both when an action is

executed and when a similar action is observed in others (Gallese, Fadiga, Fogassi, & Rizzolatti, 1996). In nonhuman primates, resonance is associated with mirror neurons located in the IPC and vPMC, which fire during both action observation and execution, and may support inferences about the goals underlying observed actions of others (Gallese et al., 1996; Gallese, Keysers, & Rizzolatti, 2004; Iacoboni, 2005; Keysers & Perrett, 2004; Rizzolatti & Craighero, 2004). Although there is agreement that action understanding in humans is also based on the principles of resonance (Kilner, Paulignan, & Blakemore, 2003; Oztop, Franklin, Chaminade, & Cheng, 2005; Press, Bird, Flach, & Heyes, 2005; Rizzolatti & Craighero, 2004; Umiltà et al., 2001), the particular role of mirror neurons in this process is still a matter of debate (Chong, Cunnington, Williams, Kanwisher, & Mattingley, 2008; Dinstein, Hasson, Rubin, & Heeger, 2007; Kilner, Neal, Weiskopf, Friston, & Frith, 2009; Mukamel, Ekstrom, Kaplan, Iacoboni, & Fried, 2010; Saygin, Chaminade, Ishiguro, Driver, & Frith, 2012).

The mentalizing system is a distributed network involving posterior areas like the temporo-parietal junction (TPJ), superior temporal sulcus (STS), and fusiform gyrus (FG), as well as anterior areas like the medial and ventromedial prefrontal cortex (mPFC, vmPFC), and anterior cingulate cortex (ACC; Saygin et al., 2012; Van Overwalle, 2009). Within the posterior part of the network, the STS is involved in processing biological motion and inferring intentions underlying biological cues, like changes in gaze or head direction, while the FG is responsible for encoding invariable facial information, such as identity (Nummenmaa & Calder, 2009). The TPJ is involved in inferring particular intentions, beliefs and higher-order action goals in a situation-specific manner (“Why is

an observed action executed?”; Chaminade & Decety, 2002; Farrer et al., 2003; Gallagher et al., 2000; Grèzes, Berthoz, & Passingham, 2006; Grèzes, Frith, & Passingham, 2004; Ohnishi et al., 2004; Perner, Aichhorn, Kronbichler, Staffen, & Ladurner, 2006; Ruby & Decety, 2001; Saxe & Kanwisher, 2003; Saxe & Powell, 2006), and allows differentiating self from other intentions via perspective taking (Chaminade & Decety, 2002; Farrer et al., 2003; Ruby & Decety, 2001). Although still a matter of debate, social functions seem to be lateralized within TPJ, with lTPJ being more involved in perspective taking (Samson, Apperly, Chiavarino, & Humphreys, 2004) and anthropomorphism (Chaminade, Hodgins, & Kawato, 2007; Cullen, Kanai, Bahrami, & Rees, 2013; Perner et al., 2006; Zink et al., 2011), and rTPJ being more responsible for discriminating intentional from nonintentional actions (Cavanna & Trimble, 2006; Chaminade et al., 2012; Gallagher et al., 2002; Krach et al., 2008) and reasoning about others’ particular internal states (Costa, Torriero, Oliveri, & Caltagirone, 2008; Gallagher et al., 2000; Saxe, 2006; Saxe & Kanwisher, 2003). The rTPJ also serves as convergence point for social and nonsocial processes (Chang et al., 2013; Krall et al., 2015; Krall et al., 2016; Mitchell, 2008; Scholz, Triantafyllou, Whitfield-Gabrieli, Brown, & Saxe, 2009), and is involved in processing language and semantics (Binder, Desai, Graves, & Conant, 2009). The anterior part of the MS includes areas like the mPFC, vmPFC, and ACC, and is involved in making inferences about others based on enduring dispositions, such as traits or preferences rather than inferring particular internal states on a trial-by-trial basis (Amodio & Frith, 2006; Brothers, 2002; Saxe, 2006; Saxe & Kanwisher, 2003; Saygin et al., 2012; Van Overwalle, 2009). This requires neurons with the ability to

represent behavior over a longer period of time, across different circumstances, and with different social partners, a feature that applies to neurons in the mPFC (Amodio & Frith, 2006; Decety & Chaminade, 2003; U. Frith & Frith, 2001; Gallagher & Frith, 2003; Huey, Krueger, & Grafman, 2006; Leslie, Friedman, & German, 2004; Wood & Grafman, 2003). Activation within mPFC is positively correlated to the degree of background knowledge one possesses about another person (Saxe & Wexler, 2005), as well as to the social relevance ascribed to information about others (Grèzes et al., 2004). In contrast, activity within vmPFC has been associated with reasoning about the emotional states of others (Hynes, Baird, & Grafton, 2006; Völlm et al., 2006). Medial prefrontal areas are also involved in impression formation by providing access to general social knowledge (Mitchell, Macrae, & Banaji, 2006; Szczepanski & Knight, 2014), with activity in mPFC being linked to retrieving stereotypical knowledge about people (Contreras, Banaji, & Mitchell, 2012; Fairhall, Anzellotti, Ubaldi, & Caramazza, 2014), and activity in vmPFC being related to retrieving script-based social knowledge (Ghosh, Moscovitch, Melo Colella, & Gilboa, 2014; van Kesteren, Ruiter, Fernández, & Henson, 2012). Medial prefrontal areas are also related to egocentric mentalizing about similar others (Jenkins, Macrae, & Mitchell, 2008; Mitchell, Macrae, & Banaji, 2004, 2006), and are activated when viewing social scenes containing human versus nonhuman agents (Wagner, Kelley, & Heatherton, 2011). The ACC is specifically activated during interactions that require mentalizing in real time (Gallagher et al., 2000; McCabe, Houser, Ryan, Smith, & Trouard, 2001) and has been suggested as a neural correlate of

mind perception, as human-like agents capable of executing intentional actions activate this brain area more strongly than nonhuman agents (Gallagher et al., 2002).

Previous research has shown that during interactions with others, activity within social brain areas is modulated by the degree to which others are perceived as “having a mind”, with stronger activation for agents believed to have a mind than for those who do not (Gallagher et al., 2002; Krach et al., 2008; Sanfey et al., 2003). For instance, observing the actions of nonhuman agents recruits the APS to a smaller degree than observation of human actions (Kilner et al., 2003; Oberman, McCleery, Ramachandran, & Pineda, 2007; Oztup et al., 2005; Press et al., 2005); the actual degree of activation has been shown to depend on features like physical appearance (Chaminade et al., 2007; Kupferberg et al., 2012), motion kinematics (Bisio et al., 2014), and familiarity (Press, Gillmeister, & Heyes, 2007). Agents failing to trigger mind perception also underactivate the MS, with reduced activation for nonhuman versus human agents, as well as agents who are deprived of their ability of “having a mind” due to dehumanization (Gallagher et al., 2002; Harris & Fiske, 2006; Krach et al., 2008; Özdem et al., 2016; Sanfey et al., 2003; Spunt et al., 2015; Waytz, Morewedge, et al., 2010a; Wykowska et al., 2014).

In addition to activation in social brain areas, mind perception also modulates performance and attitudes during social interactions. For example, mind perception has been shown to influence prosocial behaviors (Bering & Johnson, 2005; Epley, Waytz, Akalis, & Cacioppo, 2008; Graham & Haidt, 2010; Gray, Young, & Waytz, 2012; Shariff & Norenzayan, 2007), reactions to observing negative consequences for others (Cushman, 2008; Gray & Wegner, 2008; Ohtsubo, 2007), and the motivation to

perpetuate moral standards (Haley & Fessler, 2005). Similarly, attitudes and performance in interactions with nonhuman agents can be improved when the agents trigger mind perception by displaying human features or behaviors (Bennewitz, Faber, Joho, Schreiber, & Behnke, 2005; Fussell, Kiesler, Setlock, & Yew, 2008; Huang & Thomaz, 2011; Mutlu, Forlizzi, & Hodgins, 2006; Mutlu, Kanda, Forlizzi, Hodgins, & Ishiguro, 2012; Pfeiffer- Leßmann, Pfeiffer, & Wachsmuth, 2018; Sidner, Kidd, Lee, & Lesh, 2004; Staudte & Crocker, 2011; Wiese, Metta, & Wykowska, 2017; Yamazaki, Yamazaki, Burdelski, Kuno, & Fukushima, 2010). In contrast, agents not triggering mind perception negatively impact performance in social interactions (Caruana et al., 2016; Wiese et al., 2012; Wykowska et al., 2014) and fail to induce social facilitation (Bartneck, 2003; Park & Catrambone, 2007; Riether, Hegel, Wrede, & Horstmann, 2012; Woods, Dautenhahn, & Kaouri, 2005). Specifically, it has been shown that social signals, like changes in gaze direction, are followed to a larger extent when they are believed to reflect the actions of a mind compared to a preprogrammed algorithm (Caruana et al., 2016; Wiese et al., 2012; Wykowska et al., 2014), with faster responses to targets presented at gazed-at locations (gaze-cueing effect; Friesen & Kingstone, 1998).

Aim of study

Prior research indicates that mind perception has the capacity to modulate activation in social brain areas, as well as performance during social-cognitive tasks. However, relations between activation in brain areas related to mind perception and performance during social-cognitive tasks have yet to be established. That is, prior studies have not tested whether within-subject variation in brain activation during mind

perception is related to subsequent variation in performance on social-cognitive tasks and, if so, which brain areas are most closely related to social-cognitive performance. We address this question by relating brain activation during a mind perception task (i.e., judging the likelihood that agents have internal states; Martini, Gonzalez, & Wiese, 2016) to performance on a low-level social-cognitive task (i.e., attentional orienting to gaze cues; Friesen & Kingstone, 1998). These tasks were chosen based on previous studies showing that (a) judgments regarding others' capacity of having internal states require mind perception (Cheetham, Suter, & Jancke, 2014; Hackel, Looser, & Van Bavel, 2014; Looser & Wheatley, 2010; Martini et al., 2016; Waytz, Gray, et al., 2010), and (b) the degree to which others' gaze is followed is linked to mind perception and other more complex social- cognitive processes like mentalizing (Baron-Cohen, Leslie, & Frith, 1985). In both tasks, we used a set of images that varied in their degree of physical humanness and were created by morphing separate images of a human and a robot face into each other in steps of 20%. Manipulating physical humanness via morphing has been used in previous studies as a reliable tool to manipulate the degree to which mind is perceived in others (Cheetham et al., 2014; Hackel et al., 2014; Looser & Wheatley, 2010; Martini et al., 2016; Waytz, Gray, et al., 2010). The mind perception task was performed inside an fMRI scanner to determine the degree to which reasoning about the agents' capability of having internal states elicited activation within the social brain network; the social attention task was performed outside the scanner, and reaction times were collected in order to assess the degree to which the agents' gaze triggered shifts of spatial attention to gazed-at locations.

We first confirmed that the mind perception task activated the social brain network by employing a parametric analysis of the fMRI data utilizing the mind perception ratings as weights. As a second step, to test whether activation in the social brain network was also related to subsequent performance on a social attention task, a parametric analysis of the fMRI data was performed utilizing each participant's variation in gaze cueing across the different levels of physical human-ness as weights. Together, these two parametric analyses of the fMRI data provide insight about the neural regions involved in mind perception and relations with subsequent low-level social-cognitive performance, respectively. Of particular interest was whether any neural regions were activated not only during the mind perception task but also were related to subsequent low-level social-cognitive performance during gaze cueing. An overlap in activity between these two analyses would provide evidence that initial neural activity related to mind perception, for a particular agent, is related to subsequent low-level social-cognitive performance involving that same agent. In line with the notion that mind perception is a prerequisite for low-level social-cognitive processes like social attention, we predicted that overlapping fMRI activation would be identified within the social brain network.

Methods and materials

Participants

Twenty-two undergraduate students (seven female, mean age = 24.36, SD = 4.73) were recruited from George Mason University and paid \$15 per hour for their participation. All were right-handed, had normal or corrected-to-normal vision, had no known neurological deficits, and were not currently taking any medications known to

affect the central nervous system. The office of integrity and assurance approved all procedures, and participants provided informed consent prior to the experiment.

Stimuli

Six agent images were created that varied in their degree of physical humanness (in %) from machine-like (100% robot) to human-like (100% human) and were used both for the mind perception task and the social attention task. Changing the physical appearance of an agent in a parametric fashion has been shown to modulate the degree to which mind is attributed to an agent in previous studies (Hackel et al., 2014; Martini et al., 2016) and to alter activation within social brain areas (e.g., Gao, McCarthy, & Scholl, 2010; Looser & Wheatley, 2010; Waytz, Morewedge, et al., 2010; Wheatley, Weinberg, Looser, Moran, & Hajcak, 2011).

The stimuli were created using FantaMorph, which allows two images to be blended together at specified increments (in %). The images used to create the stimuli were the Meka S2 humanoid robot head and a male human face (Lundqvist, Flykt, & Öhman, 1998). Morphing occurred at 20% increments, yielding a total of six images (0%, 20%, 40%, 60%, 80%, 100% physical humanness; see Fig. 6). Each image was presented on white background in full frontal orientation and subtended 7.8° wide and 8.6° high. For both the mind perception and the social attention tasks, the eyes were centered on the horizontal axis of the screen. In the mind perception task, the pupils always remained centered relative to the vertical axis of the screen, looking straight ahead; in the social attention task, irises and pupils were additionally shifted with Photoshop to deviate 0.4° from direct gaze in order to create the impression of an eye movement.

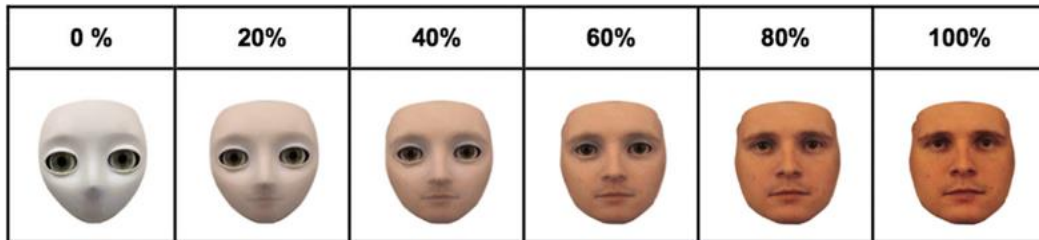


Figure 6. Stimuli used for the mind perception and social attention tasks. The images were created by morphing a robot face (Mekarobot; image on the very left) into a human face (adult male; image on the very right) in steps of 20%.

Tasks

Mind perception task

The mind perception task was performed inside of an fMRI scanner and involved making judgments about the capability of different agents (varying in their degree of physical humanness) of having internal states (see Martini et al., 2016). The sequence of events on a given trial is shown in Fig. 7. Each trial began with the presentation of a question (see Supplementary Table S1) regarding an internal state (e.g., “How likely is it that this agent has a mind?”), followed by a series of images depicting the different morphed images in a randomized order. As each agent image was presented, participants were instructed to rate the agent on the particular question that had just been presented using a Likert scale from 1 (very unlikely) to 8 (very likely). Responses were entered using a pair of fMRI-safe button boxes. Each internal state question was presented for 5 seconds, followed by a screen that contained only a fixation cross for a jittered time

period of 12 to 16 seconds. During the sequence of agent images, each agent was presented for 2 seconds, and participants were given an additional 4 seconds on average (jittered between 2 and 6 seconds) to give a response (i.e., the minimum amount of time for a response was 4 seconds). The mind perception task was divided into four blocks, each consisting of 12 questions and an average of 72 agent presentations (six agents times 12 questions) and lasting approximately 12 minutes each (total task time = 48 minutes). During the task, each of the six distinct agent images (i.e., 0%, to 100% physical humanness in steps of 20%) was presented 72 times, while each of the 24 questions was presented twice.

Social attention task

A gaze-cueing paradigm (Friesen & Kingstone, 1998) was used to measure low-level social-cognitive performance in the current experiment. This task was chosen for two reasons: (1) being able to attend to where others are looking is a prerequisite for mentalizing and other more complex social-cognitive functions and is thus a good proxy for social-cognitive performance (Frischen, Bayliss, & Tipper, 2007; for a review), and (2) the degree to which a mind is perceived in others has been shown to modulate mechanisms of social attention like gaze cueing in previous studies (Teufel et al., 2009; Wiese et al., 2012; Wykowska et al., 2014). In contrast to the mind perception task, the gaze-cueing task was performed outside the fMRI scanner and required participants to respond to the identity of a target letter (F or T) while reaction times were measured. The target either appeared at the location that was looked at by the agent (i.e., valid trial) or opposite of where the face was looking (i.e., invalid trial). Gaze-cueing effects were calculated by subtracting reaction times for valid trials from reaction times for invalid

trials (i.e., difference score). In the current experiment, we used a reversed gaze-cueing task, where targets appeared with a higher likelihood opposite of where the agent was looking (80% of the cases) compared with locations that were looked at by the agent (20% of the cases; see Friesen, Ristic, & Kingstone, 2004). This was done in order to distinguish between bottom-up components of gaze cueing, which are apparent if participants attend to the gazed-at location despite the target being more likely to appear at the uncued location (i.e., shorter reaction times at the cued location), and top-down influences on gaze cueing, which would be apparent if participants orient away from the gaze cue and shift their attention to the uncued location, which is more likely to contain the target (i.e., shorter reaction times at the uncued location).

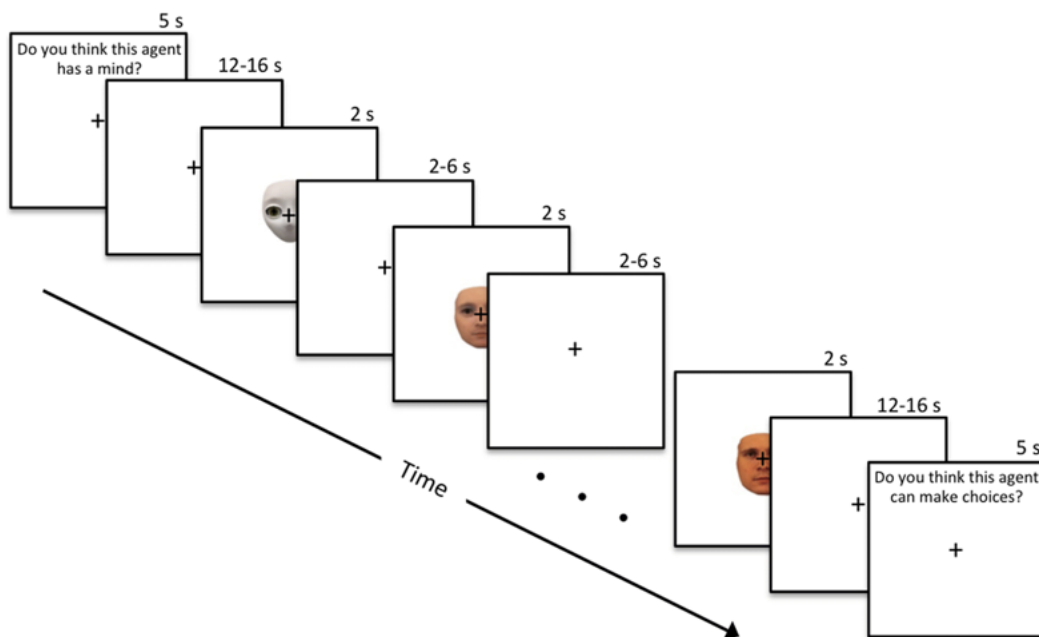


Figure 7. Mind perception task. While inside of an fMRI scanner, participants were presented with theory-of-mind questions and a series of morphed images to judge. Participants rated each image on a 1–8 scale.

The sequence of events on a given trial is shown in Fig. 8. At the beginning of each trial, a black fixation cross appeared on white screen for a jittered time interval of 700 to 1,000 ms, followed by the image of one of the agents displaying a straight gaze. After a jittered time interval of 700 to 1,000 ms, the agent changed gaze direction and looked either to the left or right side of the screen for 400 to 600 ms, followed by the presentation of the target letter (F or T, measuring $.5^\circ$ in width and $.9^\circ$ in height) that either occurred where the face was looking or opposite of where the face was looking. Targets appeared on the horizontal axis of the screen and were located 14.7° from the center of the screen. The image of the agent and the target remained on the screen until the participant gave a response or a time-out criterion was reached (1,200 ms after target presentation), whichever came first. The intertrial interval (ITI) was 680 ms.

Participants used the index finger of each hand to respond to the identity of the target letter by pressing either the key that was marked with BF[^] or BT[^]. For half of the participants, BF[^] was assigned to the BD[^] key, and BT[^] was assigned to the BK[^] key of a regular keyboard, with reversed key assignment for the other half of the participants; key labels were counterbalanced across participants throughout the study. Participants were instructed to maintain fixation on the center of the screen throughout all trials and to respond as quickly and accurately as possible to the target letters. Before the actual experiment started, participants first completed a practice block that mirrored the experimental task but used a different agent stimulus (EDDIE; developed at Technische

Universitaet Muenchen; see Wiese et al., 2012) to avoid priming effects or other response biases. Participants then performed six experimental blocks, with each block employing one of the six agent images; the order in which agents were presented was counterbalanced across participants. Total time for the social attention task was approximately 20 minutes.

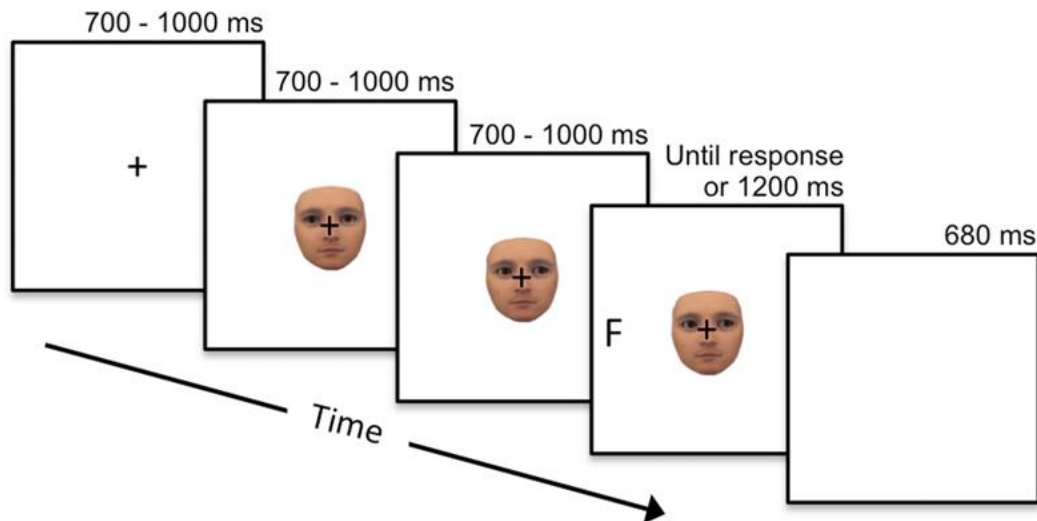


Figure 8. Social attention task. Outside of the fMRI scanner, participants performed the gaze cueing task using the same morphed images from the mind perception task. Participants were required to identify the identity of a target letter, presented in the periphery that was preceded by either a congruent or incongruent looking morph image

Procedure

The experiment started with the mind perception task in the scanner, followed by the social attention task and a series of questionnaires outside the scanner. Participants were screened for fMRI safety and completed a demographic questionnaire

approximately 1 week prior to participating in the experiment. When participants arrived on the day of the experiment, they were first provided with the instructions of the mind perception task and then positioned in the fMRI scanner in order to perform the task. Following the mind perception task, participants exited the scanner and were then provided with the instructions for a social attention task, which took place in a separate room. Critically, the same agents were employed for both the mind perception and the social attention task. After completion of the social attention task, participants filled out questionnaires and were debriefed.

Analysis

Behavioral data

The behavioral data of the mind perception and social attention tasks were analyzed using the lme4 and the Mediation packages in R (version 3.2.4). We first tested if the relationship of the three variables (physical humanness, gaze-cueing behavior, and mind ratings) were linear or nonlinear. To do so, we constructed four mixed-effects regression models (i.e., one linear and three polynomials: quadratic, cubic, and fourth level) to model the data. This step was done for each of our three predictive relationships. In other words, we tested whether the mind ratings could be predicted by physical humanness in a linear or nonlinear method, if gaze-cueing behavior could be predicted by physical humanness, in a linear or nonlinear fashion, and if gaze-cueing behavior could be predicted by mind ratings in a linear or nonlinear way (see Fig. 9). After we examined the linear and nonlinear relationships for all pairs of our three variables (mind ratings and physical humanness, gaze-cueing behavior and physical humanness, gaze-cueing behavior and mind ratings), we compared the linear model to the nonlinear models in a

nested model comparison to determine which of the models represented the data best. Choosing the model of best fit was decided based on a chi-square test that compares more complex, polynomial models (i.e., quadratic, cubic, and fourth level) to a linear reference model (i.e., the simplest model fit; all models with a chi-square test result of $p < .05$ differ significantly from the linear model in terms of model fit). Moreover, the model with the smallest Bayesian information criterion (BIC) constitutes the best, and at the same time most parsimonious, model fit for a given data set (Konishi & Kitagawa, 2008). This step was repeated for each pair of relationships (mind ratings and physical humanness, gaze-cueing behavior and physical humanness, and gaze-cueing behavior and mind ratings).

After testing for which of the relationships were linear and which were nonlinear, we investigated a mediation model that predicted gaze-cueing behavior from physical humanness through mind ratings as a mediator. To avoid overfitting the model and to aid interpretation, we specified a more simplistic mediation model by allowing only linear relations for the mediation analysis, regardless of how well the polynomials predicted the outcome of the nested model comparisons described above.

- (A) $y = b_1 x$
- (B) $y = b_1 x + b_2 x^2$
- (C) $y = b_1 x + b_2 x^2 + b_3 x^3$
- (D) $y = b_1 x + b_2 x^2 + b_3 x^3 + b_4 x^4$

Figure 9. Equations tested in the nested model comparison. Data were modeled using a linear model (a), as well as a quadratic (b), cubic (c), and fourth-level (d) polynomial. The error terms and intercept have been omitted in all of the equations

fMRI data

Image acquisition and preprocessing

We acquired fMRI data using a Siemens Allegra 3T scanner, equipped with a standard one-channel quadrature birdcage head coil. During each run, T2* gradient-echo, echo-planar imaging was acquired, with a TR/TE of 2300/30 ms, flip angle = 90 degrees, 40 interleaved axial slices 3 mm thick/1 mm gap, FOV = 192 mm, and matrix size = 64×64 (in-plane resolution of 3 mm^2). Following fMRI acquisition, a whole-head, T1 structural scan was acquired using a three-dimensional, magnetization-prepared, rapid-acquisition gradient echo (MPRAGE) pulse sequence. During the MPRAGE sequence, 160 1-mm-thick slices (256×256 matrix, field of view = 260, .94 mm voxels) were acquired with a TR/TE of 2300/3 ms.

All analyses of fMRI data were performed using FSL (www.fmrib.ox.ac.uk/fsl). In order to allow the scanner to reach equilibrium magnetization, the first five volumes were removed prior to analysis. The fMRI data were high-pass filtered (128-s cutoff), slice timing corrected (Hanning-windowed sinc interpolation to shift each time series relative to the middle of the TR period), and motion corrected using FMRIB's Linear Registration Tool (MCFLIRT). Prewhitening using FMRIB's Improved Linear Model (FILM) was performed to remove temporal autocorrelation in the fMRI time-series data. Data were smoothed using a 6-mm full-width at half-maximum (FWHM) Gaussian kernel. Coregistration was completed in a two-step process. Functional data were first registered to a high-resolution structural image (MPRAGE) using FMRIB's Linear

Registration Tool (FLIRT) following brain extraction using the Brain Extraction Tool (BET) with the fractional intensity threshold set to .35. Registration to standard space (T1 2-mm MNI template) was then performed using FLIRT.

Neural activation associated with mind perception

The first analysis of the fMRI data sought to identify whether the mind perception task reliably activated regions within the social brain network. To this end, a parametric analysis of the fMRI data was carried out, with a parametric regressor being used to identify neural regions that tracked trial-by-trial variation in mind perception. The initial, a priori analysis of the data employed all four blocks (separate runs of fMRI acquisition). However, while results from this initial analysis yielded a cluster of activation within vmPFC (see Fig. 9), no activations survived a whole-brain correction for multiple comparisons. Due to concerns that the lack of statistical robustness for this initial analysis was the result of habituation and repeated exposure to the same agent images over extended periods of time, we performed a second, post hoc, analysis of these data in which only the first two blocks (separate runs of fMRI acquisition) were employed. This second analysis was performed in an effort to optimize the likelihood of identifying statistically robust neural regions that tracked trial-by-trial variation in mind perception; indeed, as described within the Results section, this post hoc analysis revealed a qualitatively similar cluster of activation within the vmPFC that survived correction for multiple comparisons.

For both the a priori and post hoc analyses of the fMRI data, a parametric regressor modeled the onset of each agent at a magnitude determined by the mean-centered Likert-scale rating provided on each trial, whereas a second task-related

regressor modeled the onset of each agent image at a fixed magnitude. A nuisance regressor was also included to model the onset of each question, using a fixed magnitude. All task-related regressors were convolved with a canonical double-gamma hemodynamic response function (HRF) with no phase delay. Six motion parameters (three translation, three rotation) were also added to the GLM model as confound regressors in order to account for residual motion effects after correction by MCFLIRT (nine regressors total). A second-level analysis was used to average across the first two runs for each participant using a fixed-effects model. Data were then averaged across participants in a third-level analysis, using FMRIB's Local Analysis of Mixed Effects (FLAME1). We then conducted a whole-brain analysis investigating the parametric effect of gaze cueing. The family-wise error rate (FWER) were controlled for at an alpha level of .05, using cluster-based correction following Gaussian random field (GRF) theory and a cluster-defining threshold of $Z = 1.96$.

Association of neural activation during mind perception and social attention

Following the identification of neural regions associated with mind perception, we sought to identify how brain activity elicited by the mind perception task was related to the degree of gaze cueing during the social attention task. To this end, we again performed a parametric analysis; however, in this second analysis, the parametric regressor was modulated based on average gaze-cueing effect values that an individual exhibited for each agent during the social attention task. Therefore, the second parametric analysis allowed us to identify which brain regions that were activated during the mind perception task were directly related to performance in a separate low-level social-cognitive task. Moreover, we identified neural regions that were significantly activated

by both the parametric analysis based on gaze cuing and the parametric analysis based on mind ratings. In line with the a priori analysis of neural activity associated with the mind perception task, all four blocks of fMRI data were also employed for the analysis of relations between neural activity during mind perception and behavior during the social attention task. All other aspects of this second fMRI analysis were identical to those described for the analysis focusing on mind perception (see above).

Results

Behavioral data

Mind perception task

Results of the nested model comparison predicting mind ratings from physical humanness revealed that both the cubic model, $\chi^2(2) = 11.04$, $p = .003$, BIC = 430.05, and the fourth-level polynomial model, $\chi^2(1) = 12.54$, $p < .001$, BIC = 422.39, fit the data significantly better than the linear model; the fourth-level polynomial model constitutes the overall model of best fit based on the BIC estimate (i.e., smallest BIC; see Fig. 10). These results suggest that linear changes in physical human-likeness do not lead to linear changes in ratings of mind perception; in contrast, linear increases in human-likeness were associated, on average, with a nonlinear (fourth-level polynomial) increase in ratings of mind perception.

Social attention task

The nested model comparison of models predicting gaze-cueing behavior from physical humanness showed that only the cubic model fit significantly different better than the linear model, $\chi^2(2) = 4.51$, $p = .03$; however, the cubic model was not the most parsimonious model based on the BIC (i.e., BIC for the cubic model was larger than the

BIC for the linear model; see Table 2). Thus, the linear model constitutes the overall best model fit for the gaze-cueing data (see Fig. 11). This result suggests that linear increases in physical human likeness lead to linear increases in gaze cueing. That is, although gaze cues invalidly cued the target location on 80% of trials, increases in physical humanness led to increased reflexive attentional orienting in direction of the gaze cue (and slower response times at the uncued location).

Link between physical humanness, mind perception, and social attention

Before examining the nested model comparison of models predicting gaze-cueing behavior from mind ratings, we controlled for the agents' physical humanness by adding it as a covariate in the model. After controlling for physical humanness, the nested model comparison of models predicting gaze-cueing behavior showed that none of the polynomial models fit significantly better than the linear model, as indicated by the chi-square test (see Table 3). This indicates that the linear model is the best fit for the relationship between gaze-cueing behavior and mind ratings (after controlling for physical humanness). This finding illustrates that reflexive orienting to gaze cues decreases as mind ratings increase, but that voluntary attentional orienting to predicted target locations increases (see Fig. 12). In other words, after controlling for physical humanness, increases in mind ratings were associated with reductions of the “bottom up” component and enhancements of the “top down” component of gaze cueing, which led to a greater reliance on the predictivity of the gaze cue and faster response times to targets appearing at the uncued location. These data are consistent with the notion that increased mind ratings lead to a greater reliance on higher-level behavioral attributes of the agent (i.e., the predictivity of its gaze direction).

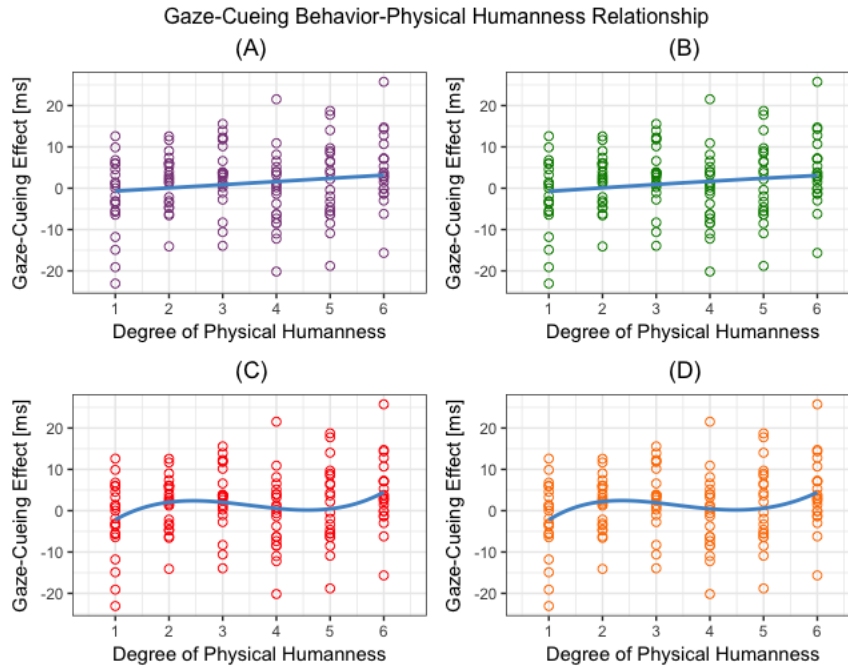


Figure 10. Average mind ratings as a function of physical humanness. 1 = 0% physical humanness; 6 = 100% physical humanness, as modeled by a linear (a), quadratic (b), cubic (c), and fourth-level polynomial (d) model. The fourth-level polynomial model constituted the overall best model fit (see Table 1).

After investigating the models of best fit for all three relationships (mind ratings and physical humanness, gaze-cueing behavior and physical humanness, and gaze-cueing behavior and mind ratings), we tested whether mind ratings partially mediated the relationship between physical humanness and gaze-cueing behavior, despite their having effects of opposite directions on gaze-cueing performance. As indicated before, we only used linear models to avoid overfitting the model with too many parameters, as well as to simplify the interpretation. The mediation analysis revealed a nonsignificant total effect

of ($\beta = .15$, 95% CI $[-.01, .33]$, $p = .08$), a significant positive direct effect ($\beta = .44$, 95% CI $[.14, .73]$, $p < .01$) of physical humanness on gaze-cueing behavior, as well as a significant negative indirect effect of mind perception on gaze cueing ($\beta = -.28$, 95% CI $[-.52, -.05]$, $p < .01$; see Fig. 13). Since physical humanness and mind ratings were highly correlated ($r = .83$), the observed negative indirect relationship between mind perception and gaze-cueing behavior needs to be interpreted with caution, as it could be an artifact due to issues with multicollinearity (Cohen, Cohen, West, & Aiken, 2003); this would mean that we could not be certain of the direction of this effect as the sign (positive or negative) of the weight, could flip. However, since multicollinearity, if anything, decreases the power of detecting an effect and thus decreases the probability of rejecting the null hypothesis (Cohen et al., 2003), it is unlikely that the negative direction of the indirect effect is a mere artifact of multicollinearity, as the indirect effect of mind perception on gaze cueing is statistically significant despite such multicollinearity. What is more likely is that the mediation model is showing a suppression phenomenon; unlike in consistent mediation models (i.e., models that have the same direction for all of their paths), suppression occurs when two variables that are related to each other (i.e., an independent variable and a mediator) cause the dependent variable to move in opposite directions (Mackinnon, Krull, & Lockwood, 2000). Consistent with suppression, we find that adding the mediator (i.e., mind ratings) increases the strength of the relationship between physical humanness and gaze-cueing behavior (β increased from .15 to .44 after including the mediator). Taken together, the data suggest that physical humanness affects social attention performance in two

potentially opposing ways: On the one hand, increases in physical humanness seem to enhance reflexive attentional orienting to gazed-at locations (i.e., increases in gaze-cueing effects) despite the fact that the predictivity of the gaze cue is low (i.e., 20%), suggesting that changes in gaze direction are more automatically followed as the stimulus looks more human-like. On the other hand, physical humanness also exerts an indirect effect on gaze-cueing behavior by increasing mind perception, such that more human-like agents are ascribed a greater degree of mind, which in turn seems to facilitate voluntary shifts of attention away from the gazed-at location toward the likely target location (i.e., 80%).

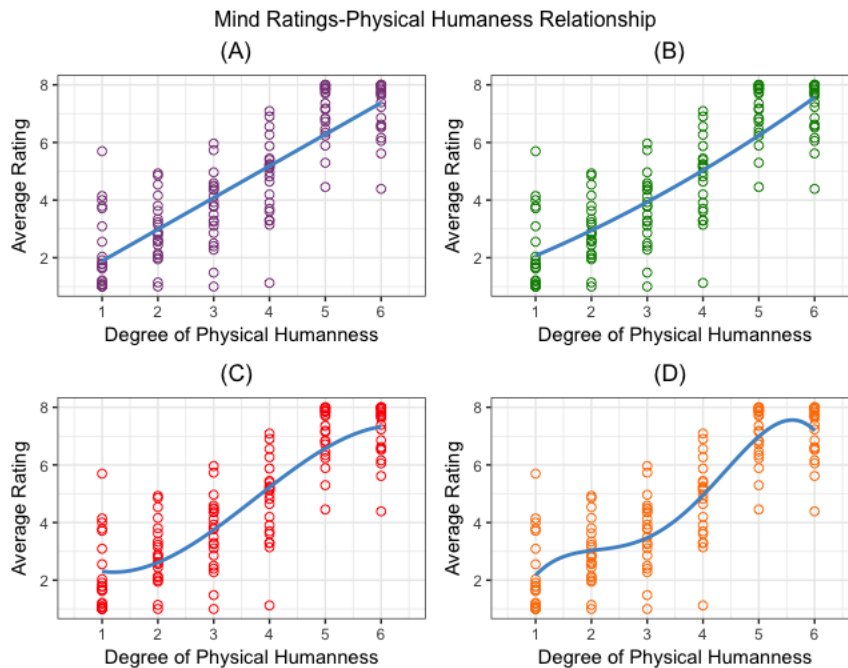


Figure 11. Average gaze-cueing effects as a function of the degree of physical humanness. For the x-axis, 1=0%human, 6=100%human, as modeled by a linear (a), quadratic (b), cubic (c), and fourth-level polynomial (d) model. The linear model constituted the overall best model fit.

fMRI data

Neural activation associated with mind perception

The neural basis of mind perception was investigated using a parametric regressor of agent image onset during the mind perception task, using trial-by-trial mind ratings to weight the regressor. This analysis allowed for testing whether the mind perception task indeed activated the social brain network, and if so, which subdivisions of this network were related to mind perception. The initial, a priori analysis employing all four blocks of the task revealed a cluster of activation within vmPFC, although this cluster of activation did not survive correction for multiple comparisons (see Fig. 14). However, a post hoc analysis that employed only the first two blocks of data, due to concerns over habituation, revealed statistically robust activation within a similar region of vmPFC that indeed survived correction for multiple comparisons. Specifically, this post hoc analysis revealed a significant cluster located primarily within vmPFC, but also extending into more anterior and less ventral subdivisions of the mPFC such as the frontal poles (peak $z = 3.14$; 2, 68, 22; 1,050 voxels). No other effects within the whole-brain analysis survived correction for multiple comparisons (see Fig. 115 for the results of the post hoc parametric analysis).

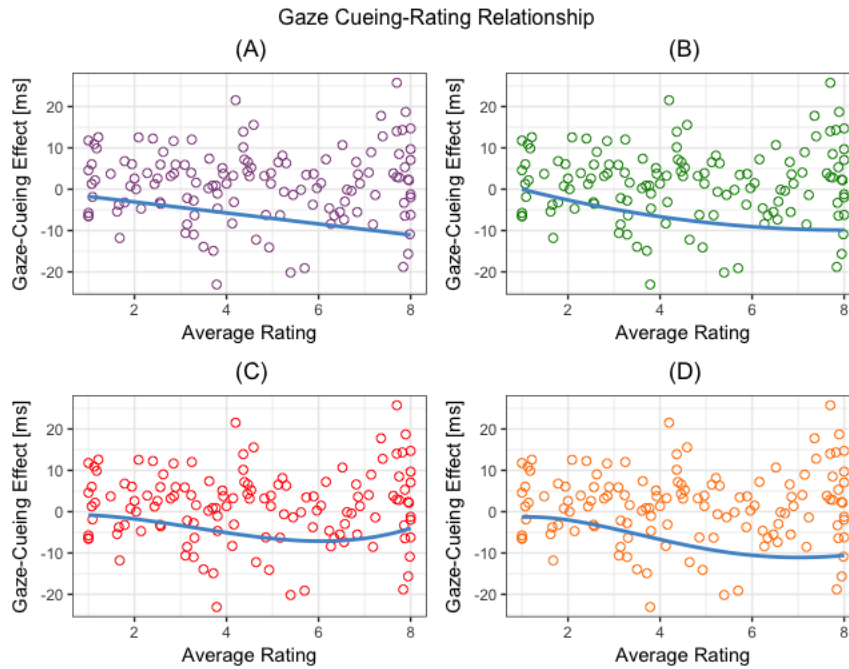


Figure 12. Average gaze-cueing effects as a function of mind ratings. Data were modeled by a linear (a), quadratic (b), cubic (c), and a fourth-level polynomial (d) model. None of the nonlinear models fit significantly better than the linear model, which is evidence that the linear model was the best predictor of gaze-cueing behavior.

Association of neural activation during mind perception and social attention

The relationship between brain activation related to mind perception and gaze-cueing performance was investigated using a parametric regressor of agent image onset during mind perception, using gaze-cueing effects to weight the regressor. This analysis allows for testing whether neural activity during the mind perception task significantly matches the patterns of gaze-cueing behavior with the respective agent. Similar to the post hoc analysis of mind ratings described above, the parametric analysis based on gaze-cueing effects revealed significant activation within the vmPFC (peak $z = 3.59$; $-46, 38, -2$; 982 voxels). Several other neural regions, such as the left TPJ and insula, right

fusiform cortex and middle temporal gyrus, and bilateral occipital cortex were also significant for the parametric analysis based on gaze cueing (see Fig. 16). Most importantly, an overlapping region of the vmPFC was identified for both parametric analyses (see Fig. 16), suggesting that vmPFC may relate not only to mind perception, but also low-level social-cognitive performance during a gaze-cueing task.

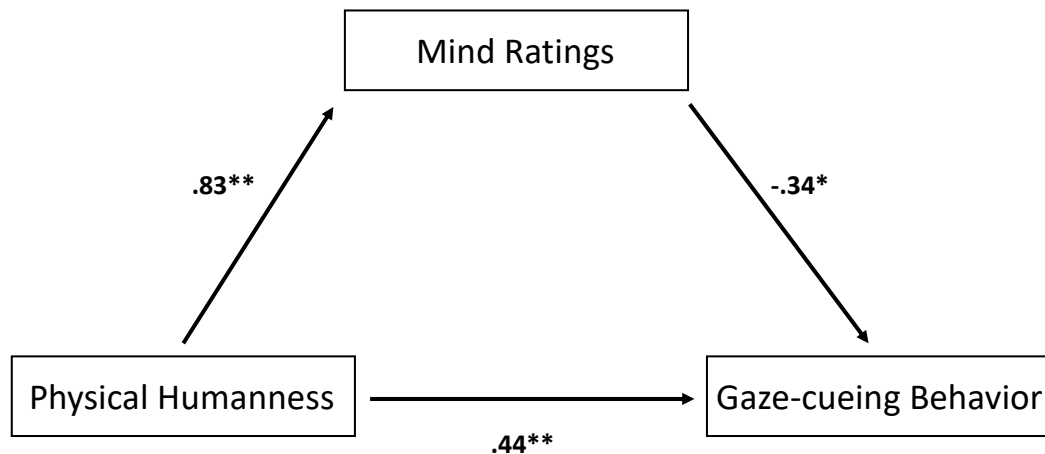


Figure 13. Path diagram illustrating the mediation model. The mediation analysis revealed both a significant and positive direct effect of physical humanness on gaze cueing, as well as a negative indirect effect, as mediated by mind ratings. Values over the directional arrows reflect standardized coefficients produced from each regression model in the mediation. * $p < .05$. ** $p < .01$

Discussion

The goal of the present experiment was to investigate whether within-subject variation in brain activation during mind perception is directly related to variations in

social attention performance, and, if so, which social brain areas are most strongly related to this performance measure. For that purpose, we manipulated the physical appearance of social agents (i.e., on a spectrum from robot to human) and measured the effect of this manipulation on two orthogonal tasks: social judgments regarding the agents' capability of having a mind (i.e., ratings and brain activation), and low-level social-cognitive performance during a social attention task (i.e., gaze-cueing effects). Patterns within the behavioral data (i.e., ratings and gaze-cueing effects) were analyzed using a nested model comparison. We used a mediation model to test the complex relations between physical humanness, mind ratings, and gaze-cueing effects. Moreover, a set of parametric analyses of fMRI data was used to investigate the relations between brain activation, mind perception, and low-level social-cognitive performance. In particular, vmPFC was found to be activated not only during mind perception, but the level of vmPFC activity during mind perception was also directly related to subsequent low-level social-cognitive performance on a separate gaze-cueing task. This pattern of results suggests that initial activity within vmPFC actually influences subsequent social-cognitive behavior. However, future research that measures vmPFC activation not only during mind perception but also during subsequent social interactions, is critical in order to identify whether the vmPFC indeed serves as a direct link between mind perception and subsequent low-level social-cognitive behavior. Moreover, additional work using larger sample sizes and more ecologically valid measures of social interaction will be needed to confirm the exact role of the vmPFC in low-level social-cognitive performance.

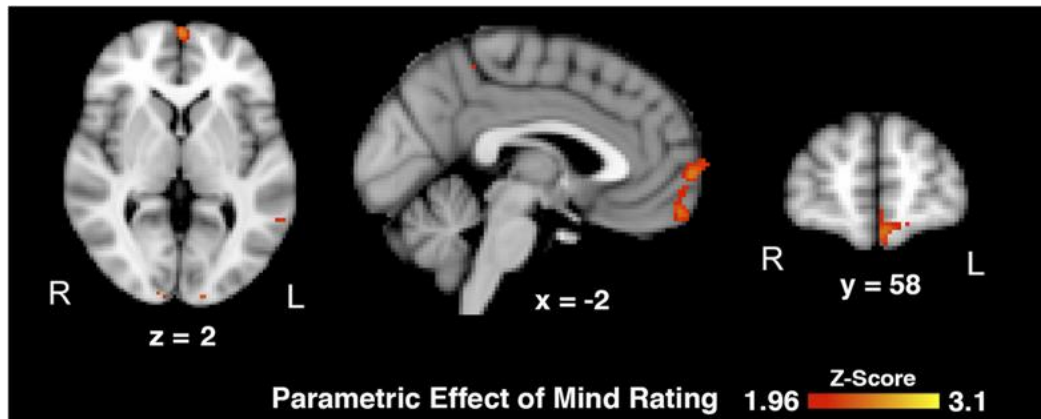


Figure 14. A priori parametric analysis of fMRI activations based on mind ratings. Z maps reflecting onset of the morph images, using mind ratings to weight the parametric regressor. From left to right: coronal ($y = 58$), sagittal ($x = -2$), and axial ($z = 2$) slices; no activations survived correction for multiple comparisons.

The linear mixed models revealed that increasing levels of physical humanness were associated with a general increase in mind ratings (i.e., positive social judgments) and low-level social-cognitive performance (i.e., stronger gaze cueing). This is consistent with prior research, demonstrating that increasing levels of physical humanness are associated with increased mind perception (Cheetham et al., 2014; Hackel et al., 2014; Looser & Wheatley, 2010; Martini et al., 2016) and improved low-level social-cognitive performance (Teufel et al., 2009; Wiese et al., 2012; Wykowska et al., 2014). However, we also found that for the counterpredictive social attention task employed here, mind perception (after controlling for physical humanness) seemed to affect gaze cueing in a different manner than physical humanness; that is, increasing levels of physical humanness directly enhanced reflexive attentional orienting to gazed-at locations (i.e.,

faster reaction times to targets presented at valid compared to invalid locations), suggesting that changes in gaze direction were more automatically followed the more the stimulus looked human-like despite the fact that the gazed-at location was unlikely to contain the target (i.e., counterpredictive cue: 20%). Increasing levels of physical humanness, however, also lead to an increase in mind perception, which seemed to facilitate voluntarily shifts of attention away from the gazed-at location and toward the location that most likely contained the target (i.e., predicted location: 80%). This pattern of results is interesting in the light of previous reports that attentional orienting to gaze cues is hard to suppress given the high social relevance of eye gaze for social learning and the development of close relationships (see Friesen & Kingstone, 1998). It suggests that although increasing physical humanness enhances the reflexive component of gaze cueing, it also leads to higher levels of mind perception, which seems to facilitate voluntary shifts of attention to uncued, but likely target locations (in counterpredictive cueing paradigms). In particular, it is possible that participants who ascribe higher levels of intentionality to the gazing stimulus might pay more attention to contingencies in its behavior, making it more likely that they pick up on the counterpredictivity of the gaze signal (which can potentially be interpreted as negative intention; e.g. “The agent wants to trick me and make me miss the target”), and adjust attentional orienting accordingly (for reports of top-down modulation of gaze cueing, see Bonifacci, Ricciardelli, Lugli, & Pellicano, 2008; Cazzato, Liuzza, Caprara, Macaluso, & Aglioti, 2015; Dalmaso, Edwards, & Bayliss, 2016; Fox, Mathews, Calder, & Yiend, 2007; Graham, Friesen, Fichtenholtz, & LaBar, 2010; Hungr & Hunt, 2012; Tipples, 2006; Wiese, Wykowska, &

Müller, 2014; Wykowska et al., 2014). Although this finding is interesting, since it points at a possible dissociation between perception of intentionality and perception of human appearance, it needs to be interpreted with caution, due to potential issues with multicollinearity in the current experiment, and warrants further investigation.

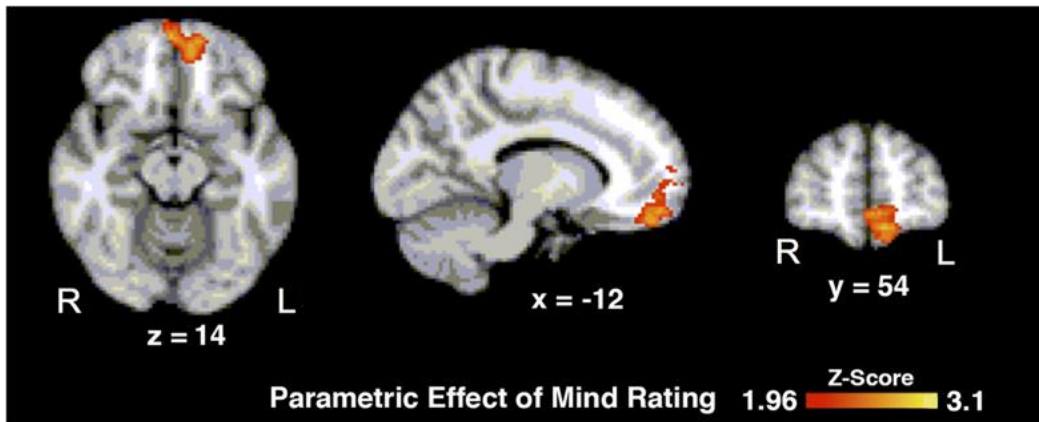


Figure 15. Post hoc parametric analysis of fMRI activations based on mind ratings. Z maps reflecting onset of the morph images, using mind ratings to weight the parametric regressor. From left to right: coronal ($y = 54$), sagittal ($x = -12$), and axial ($z = -14$) slices; cluster corrected ($Z = 1.96$, $p < .05$) at the whole-brain level.

Analysis of the fMRI data provided insight into the neural regions involved in mind perception and explored how brain activation related to mind perception is related to subsequent gaze-cueing performance (as a proxy for social-cognitive performance). We found that mind ratings were associated with vmPFC activation, a finding that is consistent with prior investigations linking perceptions of intentionality to ventromedial pre-frontal areas (Gallagher et al., 2002; Pfeiffer et al., 2014; Sanfey et al., 2003). Activity within vmPFC was also related to low-level social-cognitive performance during gaze cueing, together with a set of other regions including the left TPJ and insula,

right medial temporal gyrus and fusiform cortex, and bilateral occipital cortex. Thus, while social attention was associated with a set of regions involved in gaze perception (Nummenmaa & Calder, 2009) and mentalizing (Van Overwalle, 2009), an overlapping region of vmPFC was associated with both mind perception and social attention, suggesting that the vmPFC might play an important role in linking higher-order social-cognitive processes (like mind perception) and performance on lower-level social-cognitive tasks (like gaze cueing). However, additional research that measures neural activity not only during mind perception but also during social-cognitive tasks within the same study will be required to substantiate claims surrounding the link between mind perception and social interaction within the vmPFC.

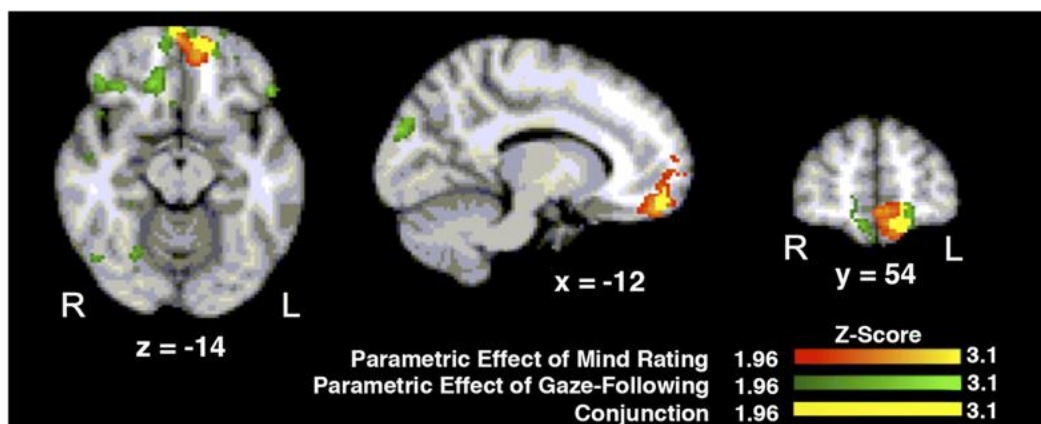


Figure 16. Parametric analysis of fMRI activations based on gaze cueing effects, mind ratings, and their conjunction. Z maps reflecting onset of the morph images, using either mind ratings (orange) or gaze cueing (green) to weight the parametric regressor, along with their conjunction (yellow). From left to right: coronal (y = 54), sagittal (x = -12), and axial (z = -14) slices; cluster corrected (Z = 1.96, $p < .05$) at the whole-brain level.

While prior work has investigated relations between mind perception and social attention (Özdem et al., 2016; Teufel et al., 2009; Wiese et al., 2012; Wykowska et al., 2014), the current study adds to these findings by showing that both mind perception and mechanisms of social attention are related to activation within the vmPFC. This neural region has been associated with mentalizing (Amodio & Frith, 2006; Frith & Frith, 1999; Frith & Frith, 2003; Gallagher et al., 2002), and is involved in impression formation in social situations by providing access to general social knowledge (Mitchell et al., 2006; Szczepanski & Knight, 2014), and retrieving script-based social knowledge (Ghosh et al., 2014; van Kesteren et al., 2012). Moreover, lesions to vmPFC result in impaired mental state understanding (Beer, Heerey, Keltner, Scabini, & Knight, 2003; Stone, Baron-Cohen, & Knight, 1998), emotion recognition (Hornak et al., 2003; Tsuchida & Fellows, 2012), social and moral reasoning (Anderson, Bechara, Damasio, Tranel, & Damasio, 1999), and cognitive empathy (Shamay-Tsoory, Aharon-Peretz, & Perry, 2009). While the vmPFC has previously been shown to modulate higher-order social-cognitive processes involved in economic or strategic decision-making (Gallagher et al., 2002; Sanfey et al., 2003), associations between this neural region and low-level social-cognitive processes like gaze cueing have not previously been reported (to the best of our knowledge).

In addition to the vmPFC literature reviewed above, it is important to note that this neural region has also been shown to track feelings of eeriness toward nonhuman agents in a parametric fashion (Wang & Quadflieg, 2015), and has been suggested as a potential neural correlate of the uncanny valley (i.e., nonhuman agents with human-like

appearance induce feelings of eeriness when being not perfectly human; Mori, 1970). Based on this research, variation of vmPFC activation in the current experiment could be driven by feelings of eeriness toward agents that are ambiguous in terms of their physical human-likeness. This could lead to a general disengagement from gaze cues as the agents' physical human-likeness increases (this is less likely since increases in physical human-likeness were associated with stronger reflexive gaze cueing in the current study), or an impaired ability to control attentional orienting in a top-down manner, since processing “uncanny” stimuli has been shown to consume additional cognitive resources to resolve the conflict of whether one is looking at a human or a nonhuman agent (Weis & Wiese, 2017). Previous studies have also shown that activation in medial prefrontal areas, and in particular bilateral vmPFC, is related to evaluating the predictability of stimuli, leading to higher levels of anthropomorphism if a stimulus is hard to predict (Waytz, Morewedge, et al., 2010). Thus, it is possible that activation within vmPFC reflects one's sensitivity to the predictability of gaze cues, and that the level of vmPFC activation is modulated by the degree to which mind is perceived in the agents. This interpretation is in line with previous neurophysiological studies showing that evaluations of predictability are associated with additional neural effort in frontocentral areas for human versus robot agents (Caruana et al., 2016).

In addition to activation in bilateral vmPFC, gaze cueing performance was also associated with activation in left TPJ and insula, right medial temporal gyrus and fusiform cortex, and bilateral occipital cortex. This is in line with previous studies showing that both prefrontal and temporo-occipital areas like bilateral TPJ and STS are

implicated in social attention (Nummenmaa & Calder, 2009). Previous studies have also related TPJ activation to judgments about another's intentionality, with stronger activation for agents with versus without a mind (Cavanna & Trimble, 2006; Chaminade et al., 2012; Gallagher et al., 2002; Krach et al., 2008), and shown that TPJ serves as convergence point for social and nonsocial processes (Chang et al., 2013; Krall et al., 2015; Krall et al., 2016; Mitchell, 2008; Scholz et al., 2009). Notably, Özdem et al. (2016) have shown that attentional orienting in response to nonpredictive gaze cues is sensitive to the perceived intentionality underlying these cues (i.e., human controlled vs. preprogrammed) and is associated with activation in bilateral TPJ. The question remains, however, why only relations between gaze cueing performance and the left TPJ reached statistical significance in the current experiment, although both left and right TPJ are activated during mentalizing and social judgment tasks. First, it is possible that the lack of significant activation within the right TPJ could simply arise as a result of issues with statistical power. However, we might also suggest that social functions of the TPJ are lateralized and that the functionalities subserved by the left TPJ (i.e., attribution of human-likeness; Perner et al., 2006) might be more important for the current task than the functionalities of the right TPJ (i.e., mentalizing; Costa et al., 2008; Gallagher et al., 2000; Saxe, 2006; Saxe & Kanwisher, 2003). Specifically, right TPJ activation is found during classic mentalizing tasks (Frith & Frith, 2003; Saxe & Wexler, 2005), while left TPJ activation is related to perspective taking (Samson et al., 2004), anthropomorphism (Chaminade et al., 2007; Cullen et al., 2013; Zink et al., 2011), and processing of agent identity from visual information (Van Overwalle, 2009). Cullen et al. (2013) also showed

that gray-matter volume in the left TPJ is related to individual differences in one's willingness to treat nonhuman entities as human-like, and Chaminade et al. (2007) showed that activation in the left TPJ is positively correlated with one's tendency to perceive humanness in motion patterns of nonhuman agents. Both the mind perception task and gaze-cueing task employed in the current study required reasoning about the agents' human- likeness based on visual features, which is expected to trigger different degrees of anthropomorphism (Cheetham, Suter, & Jäncke, 2011, 2014; Martini et al., 2016) and might explain why specifically left TPJ activation was found to be related to social attention performance. Nonetheless, the lateralized function of the TPJ observed in the current experiment will require replication in future work.

Conclusion

In sum, the present study provides evidence that variation in bilateral vmPFC activation, when perceiving the mind of a novel agent, is related to variation in subsequent low-level social-cognitive performance when interacting with that agent, as measured in gaze-cueing performance. Critically, this relationship was identified by recording neural activity upon initial exposure to a set of novel agents, followed by engaging in a separate, orthogonal low-level social-cognitive task with the same agents. The current study adds to previous research by (a) showing that the degree to which an agent is perceived to have a mind is significantly related to low-level social-cognitive performance on an orthogonal task with the respective agent (with the advantage that measuring brain activation related to mind perception is not confounded by behavioral performance during the low-level social-cognitive task), and (b) identifying a potential

neural substrate associated with both mind perception and low-level social-cognitive performance: the bilateral vmPFC. This finding also adds to a growing body of evidence suggesting that mind perception constitutes a source of top-down modulation on attentional orienting, ensuring that more attentional resources are devoted to interactions with agents who are believed to have a mind compared to machine agents without a mind (Krall et al., 2015; Mitchell, 2008; Özdem et al., 2016; Scholz et al., 2009; Wiese et al., 2012; Wykowska et al., 2014). Future research could build on the present results by employing a network perspective, probing the functional or structural connectivity of the vmPFC with other neural regions involved in social cognition and attention.

STUDY 3.

Introduction

People are seeing a day-to-day increase in the number of robot agents in their lives that could assist them in various domains (Tapus & Matarić, 2006). Although evidence of positive outcomes of human–robot interactions exist (Basteris et al., 2014; Mubin, Stevens, Shahid, Mahmud, & Dong, 2013; Scassellati, Admoni, & Matarić, 2012), designing social robots that elicit natural human responses can be challenging owing to people’s negative perceptions about having social robots be a part of their everyday life (Bartneck & Reichenbach, 2005; Scopelliti, Giuliani, & Fornara, 2005). To remedy this and design robots that are able to elicit social responses, we must understand how the human brain processes social information in interactions with others and whether non-human agents are able to activate these networks to a similar extent to human interaction partners (and if so, under which conditions). Following this approach, we can identify physical and behavioral agent features that reliably activate brain areas involved in social-cognitive processes and investigate whether robots that elicit activation in these networks lead to more acceptance and trust, as well as improved performance in human–robot interaction (Wiese et al., 2017).

Meaningful social interactions require the ability to infer internal states of others, such as intentions (i.e. mentalizing) and emotions (i.e. empathizing) (Baron-Cohen, 1997), and to use this information to predict future behavior (Frith & Frith, 2006b). For that purpose, the human brain is equipped with neural networks specialized in processing

information relevant to social interactions (i.e. social brain; (Adolphs, 2009; Frith & Frith, 2006b; Van Overwalle & Baetens, 2009)) which involve posterior areas like the temporo-parietal junction (TPJ), the superior temporal sulcus (STS) and the fusiform gyrus (FG), as well as anterior areas like the anterior cingulate cortex (ACC) and the ventromedial and dorsolateral prefrontal cortex (vmPFC, dlPFC) (Amodio & Frith, 2006; Bzdok et al., 2013; Frith & Frith, 2006b; Gallagher & Frith, 2003; R Saxe, Carey, & Kanwisher, 2004; R. P. Spunt & Lieberman, 2012; Robert P. Spunt & Adolphs, 2014).

TPJ is involved in inferring higher-order action goals (Chaminade & Decety, 2002; Farrer et al., 2003; Gallagher et al., 2000; Grèzes, 2004; Grèzes, Berthoz, & Passingham, 2006; Ohnishi et al., 2004; R. Saxe & Kanwisher, 2003; Rebecca Saxe & Powell, 2006) and mental and spatial perspective taking (Chaminade & Decety, 2002; Farrer et al., 2003; Ruby & Decety, 2001), while STS and FG are involved in processing bio- logical motion and face identity, respectively (Frith & Frith, 2006b; R Saxe et al., 2004). PFC is involved in making inferences about enduring dispositions such as preferences or beliefs (Amodio & Frith, 2006; R. Saxe & Kanwisher, 2003; Rebecca Saxe, 2006; Van Overwalle, 2009), and activation in medial PFC is positively correlated with the amount of back- ground knowledge we have about others (Grèzes et al., 2006; Rebecca Saxe & Wexler, 2005) . ACC is activated in social interactions requiring mentalizing in real- time (Gallagher et al., 2000; McCabe, Houser, Ryan, Smith, & Trouard, 2001), and is more strongly activated when interacting with human versus machine agents (i.e. computers; (Gallagher, Jack, Roepstroff, & Frith, 2002)).

Critically for human – robot interaction, most current social robot platforms underactivate the social brain network (Gallagher et al., 2002; Harris & Fiske, 2011; Özdem et al., 2016), which negatively impacts social (e.g. joint attention), emotional (e.g. empathy) and cognitive (e.g. trust) processes that would be essential in order for humans to socially interact with robot agents. Fortunately for social roboticists, social brain activation depends on the degree to which an interaction partner is perceived as ‘having a mind’ (i.e. mind perception (Robert P. Spunt, Meyer, & Lieberman, 2015)), with the general capability of making changes in the environment (i.e. agency; (Gray et al., 2007)), and experiencing internal states, such as emotions and intentions (i.e. experience), and as such can presumably be triggered by design. Whereas mind is easily perceived in other human agents (Epley, Waytz, & Cacioppo, 2007), the degree to which non-human entities like robots trigger mind perception depends on whether their physical and behavioral characteristics are perceived as sufficiently human-like (Castelli, Happé, Frith, & Frith, 2013; DiSalvo & Gemperle, 2003; Heider & Simmel, 1944; Kiesler et al., 2008). Mind perception is in fact a highly effortless process that activates social brain networks in a bottom-up or reflexive fashion (Gao, McCarthy, & Scholl, 2010; Looser & Wheatley, 2010; Schein & Gray, 2015; Wheatley, Weinberg, Looser, Moran, & Hajcak, 2011), triggered by human-like facial features and relations (i.e. spatial arrangement of eye – nose – mouth configurations) (Balas & Tonsager, 2014; Deska, Lloyd, & Hugenberg, 2016; Looser & Wheatley, 2010; Maurer, Grand, & Mondloch, 2002; Schein & Gray, 2015), as well as biological motion and/or predictable behavior (Waytz et al., 2010).

Owing to its reflexive nature, mind perception allows observers to differentiate intentional from non-intentional agents within a few hundred milliseconds (Looser & Wheatley, 2010; Wheatley et al., 2011), and even just passively viewing stimuli that trigger mind perception is sufficient in order to induce activation in a wide range of social brain networks (Wagner, Kelley, & Heatherton, 2011), which varies parametrically as a function of the agent's physical human-likeness (i.e. increases as the agent's face or body becomes more human-like in appearance) (Krach et al., 2008; Wiese, Buzzell, Abubshait, & Beatty, 2018). Brain areas related to variations in mind perception involve anterior social brain areas, such as the left ACC (Gallagher et al., 2000; Sanfey, Rilling, Aronson, Nystrom, & Cohen, 2003; Robert P. Spunt & Adolphs, 2014), as well as posterior areas, such as the left TPJ (Cullen, Kanai, Bahrami, & Rees, 2013). Left ACC is activated when others are treated as intentional agents (Gallagher et al., 2002; Sanfey et al., 2003), and responds more strongly during social decision-making tasks that involve intentional versus non-intentional agents (Gallagher et al., 2002; Sanfey et al., 2003); activation within left TPJ is associated with attributing human-likeness and intentionality to non-human agents (Chaminade, Hodgins, & Kawato, 2007; Perner, Aichhorn, Kronbichler, Staffen, & Ladurner, 2006; Zink et al., 2011), and grey matter volume in left TPJ corresponds to individual differences in anthropomorphizing non-human entities, in particular animals (Cullen et al., 2013). The degree to which agents trigger mind perception not only modulates activation in social brain areas, it also determines how we feel about the agents (Cehajic, Brown, & Gonzalez, 2009; Gutsell & Inzlicht, 2012; Harris & Fiske, 2011; Saygin, Chaminade, Ishiguro, Driver, & Frith, 2012), behave

towards them (Bering & Johnson D, 2005; Gallagher et al., 2002; Harris & Fiske, 2011; Hertz & Wiese, 2017; Sanfey et al., 2003) and interact with them (Hertz & Wiese, 2017; Riether, Hegel, Wrede, & Horstmann, 2012; Short, Hart, Vu, & Scassellati, 2010; Waytz et al., 2010), and as such has the potential to impact acceptance, trust and performance in human – robot interactions. Mind perception affects higher-order social-cognitive processes like prosociality, morality and economic decision-making (Bering & Johnson D, 2005; Waytz et al., 2010), and also modulates low-level social-cognitive processes, such as face perception (K. Takahashi & Watanabe, 2013) and social attention (Abubshait & Wiese, 2017; Caruana, de Lissa, & McArthur, 2017; Wiese, Wykowska, Zwickel, & Müller, 2012; Wykowska, Wiese, Prosser, & Müller, 2014). The effect of mind perception on social cognition is so profound that the mere belief that observed behavior may reflect the actions of an agent ‘with a mind’ makes people interpret pre-programmed behaviors as intentional and motivates them to attune their actions accordingly (Epley, Waytz, Akalis, & Cacioppo, 2008; Wiese et al., 2012; Wykowska et al., 2014).

Although the neural link between mind perception and social brain activation (Van Overwalle, 2009; Wiese et al., 2018), as well as the behavioral link between mind perception and social-cognitive processes, are relatively well understood (Wiese et al., 2017), research has just begun to examine how mind perception modulates social cognition on a neural level (i.e. activation of which brain areas modulates social cognition as a function of mind perception) (Özdem et al., 2016). To address this important issue, we use transcranial direct current stimulation (tDCS) to investigate the

link between activation of social brain areas implicated in mind perception and low-level social cognitive processes, such as social attention (i.e. the degree to which changes in gaze direction are followed) (Friesen & Kingstone, 1998). tDCS is a non-invasive electrical stimulation technique that can be applied without disrupting the participant during task execution, and has been proven to be effective in modulating a wide range of cognitive processes in previous studies (e.g. memory, attention, decision-making, perception; (Antal, Nitsche, & Paulus, 2001; Coffman, Clark, & Parasuraman, 2014; Cohen Kadosh, Soskic, Iuculano, Kanai, & Walsh, 2010; Jacobson, Koslowsky, & Lavidor, 2012). Social attention, or the tendency to follow changes in others' gaze direction, was chosen for the present study, as it is a basic, yet essential, social-cognitive mechanism that allows for initiating and coordinating communication (Adams & Kleck, 2005; Blakemore, Winston, & Frith, 2004; Nummenmaa & Calder, 2009), establishing joint attention between two interaction partners and an object of interest in the environment (Frischen, Bayliss, & Tipper, 2007), and an important precursor for developing a theory of mind (Frischen et al., 2007).

Social attention can be examined using a gaze-cueing paradigm, where a face-like stimulus is presented centrally on a screen that first looks straight and then changes gaze direction to the left or right side of the screen. This so-called gaze cue is followed by the presentation of a target stimulus (e.g. F or T) that appears at either the cued location (i.e. valid trial) or an uncued location (i.e. invalid trial) and triggers shifts of the observer's attention to the gazed-at location. As a result, reaction times on valid trials are usually shorter than reaction times on invalid trials, with the difference in reaction times between

invalid and valid conditions constituting the gaze- cueing effect (Friesen & Kingstone, 1998). Although it is widely accepted that social attention has a strong bottom-up component (i.e. attention is shifted reflexively to the cued location; (Frischen et al., 2007), there is accumulating evidence that it can be top-down controlled by higher-order social-cognitive processes when context information is available that increases the social relevance of observed gaze signals (e.g. gazer is similar or known to the observer; (Bonifacci, Ricciardelli, Lugli, & Pellicano, 2008; Cazzato, Liuzza, Caprara, Macaluso, & Aglioti, 2015; Ciardo, Marino, Actis-Grosso, Rossetti, & Ricciardelli, 2014; Mario Dalmaso, Edwards, & Bayliss, 2016; Fox, Snyder, Vincent, & Raichle, 2007; Hungr & Hunt, 2012; Kawai, 2011; Porciello et al., 2014; Ristic & Kingstone, 2005; Tipples, 2006; Wykowska et al., 2014). With particular relevance to the current study, manipulating the degree to which a gazer is perceived as an intentional being ‘with a mind’ has been shown to modulate social attention, such that gaze-cueing effects are larger in response to human (i.e. intentional) versus machine (i.e. pre-programmed) gazers (Caruana, McArthur, Woolgar, & Brock, 2017; Wiese, Wykowska, et al., 2014; Wiese et al., 2012; Wykowska et al., 2014). The notion that this top-down modulation is specifically related to mind perception is supported by experiments showing that although individuals on the autism spectrum reflexively attend to gaze signals, they do not show the reported enhancement of gaze-cueing effects for human agents but an enhancement in attentional orienting in response to robot gaze cues (potentially owing to difficulties with inferring internal states underlying human gaze behavior but an increased interest in machine behavior; (Wiese, Wykowska, et al., 2014).

While modulatory effects of mind perception on social attention have been examined behaviorally (Bering & Johnson D, 2005; Riether et al., 2012), how mind perception modulates lower-level social-cognitive mechanisms like gaze cueing at the neuronal level still remains an open question. In particular, it is unclear whether anterior or posterior parts of the social brain network are more strongly involved in modulating social cognition via mind perception. On the one hand, Özdem et al. showed that believing that an agent's eye movements are human-controlled (i.e. intentional) as opposed to pre-programmed (i.e. non-intentional) modulated activation in bilateral TPJ (but not prefrontal networks), and enhanced attentional orienting to gaze cues (Özdem et al., 2016). On the other hand, Wiese et al. showed that although variations in social attention in response to human versus robot gazers correlated with activation in anterior (i.e. vmPFC) and posterior social brain areas (i.e. TPJ), only activation in bilateral vmPFC correlated with both variations in mind perception and social attention, suggesting that vmPFC might be involved in the top-down control of social attention via mind perception (Wiese et al., 2018). Taken together, these findings show that although there is convincing evidence that social attention can be modulated via mind perception, the exact source of this modulatory effect still needs to be identified.

To address inconsistencies of previous studies and to examine whether anterior and/or posterior areas of the social brain network implicated in mind perception are causally involved in modulating social attention, in the current study we compare gaze-cueing effects induced by agents 'with a mind' (i.e. human) to agents 'without a mind' (i.e. robot), with and without tDCS stimulation to left prefrontal and left temporo-parietal

areas. This manipulation was chosen for two reasons: first, previous studies showed that sophisticated minds are attributed to agents with human-like appearance but not to agents with a robot-like appearance (i.e. mind ratings increase as a function of physical human-likeness (Bartneck & Reichenbach, 2005; Gallagher et al., 2002; Jack & Robbins, 2012; Rosenthal-Von Der Pütten & Krämer, 2014), which allows us to experimentally manipulate mind perception via physical human-likeness. Second, because mind perception increases activation in social brain networks (Carter, Hodgins, & Rakison, 2011; Gobbini et al., 2011; Krach et al., 2008; H. Takahashi et al., 2014) and has been shown to modulate social-cognitive processes like gaze cueing on a behavioral level (Gobel, Tufft, & Richardson, 2017; Wiese et al., 2012), using tDCS in the context of a social attention paradigm seems suitable to investigate the outlined research goal. Left prefrontal and left temporo- parietal areas were chosen as sites for tDCS stimulation for the following reasons: first, both areas have been implicated in mind perception in previous studies (Gallagher et al., 2002; Özdem et al., 2016; Sanfey et al., 2003; Van Overwalle, 2009; Wiese et al., 2018), and activation in both areas has been shown to correlate with variations in social attention (i.e. gaze cueing) (Özdem et al., 2016; Wiese et al., 2018). Second, both areas are located distant enough from each other to minimize the risk of accidental stimulation of the respective other site during stimulation (i.e. prefrontal versus temporo-parietal). Third, previous studies suggest that active stimulation is superior to sham stimulation as a control, as it controls for side effects of stimulation that are not directly associated with brain functionality like stimulation sensations, and allows for drawing specific inferences regarding the origin of the

modulation in terms of location, as permitting by the low spatial resolution of tDCS targeting (Parkin, Ekhtiari, & Walsh, 2015; Polanía, Nitsche, & Ruff, 2018).

Methods

Participants

Eighty-two undergraduate students from George Mason University (57 females; $M_{\text{age}} = 19.84$ years, $SD = 2.35$, range = 18 – 29 years) participated in the current study for course credit. All participants were right-handed, had normal or corrected-to-normal vision, had no known neurological or psychological deficits, were not taking any medications known to affect the central nervous system at the time of the experiment, and had no history of migraines, seizures, or head injuries. All participants provided written consent to participation and were debriefed at the end of the study. Collection and handling of participant data were in accordance with the Internal Review Board guidelines (obtained prior to data collection). Data of participants whose accuracy rate in the gaze-cueing task during baseline was below 85% (six participants), who did not follow the instructions properly (two participants), or did not complete the study (two participants) were excluded from data analysis (13% of participants in total). The remaining 72 participants were quasi-randomly assigned to the experimental condition, that is: active stimulation of the left prefrontal cortex (PFA; $n = 36$, 23 females), or stimulation of the left temporo-parietal cortex (TPA; $n = 36$, 27 females). To determine the sample size needed for the study, an a priori power analysis was conducted in G*power for a mixed-effects repeated measures design (i.e. one between factor and one within factor). Since mixed-effects repeated measures designs in G*power can only

handle two-factorial designs, the sample size was determined for an ANOVA with a 2-level between factor and a 4-level within factor (i.e. the two within variables were combined). The analysis was based on an alpha of ($\alpha = 0.05$), the power set to ($1 - \beta = 0.95$), and the assumption of a small-to-medium effect size (Cohen's $d = 0.17$). The analysis showed that an approximate sample size of 76 participants would be sufficient for both experimental groups. Since the current version of G*power cannot test for a three-way factorial design but only a two-way factorial design, another power analysis was conducted using R for the general linear model, which is a more generic test. The analysis was done for a $2 \times 2 \times 2$ factorial model that contains three main effects and their interactions (i.e. three two-way interactions and one three-way interaction) using the same alpha ($\alpha = 0.05$), beta ($1 - \beta = 0.95$) and effect size (Cohen's $d = 0.17$) as the previous analysis. The analysis for the general linear model revealed that the study needed a total of 83 participants.

Apparatus

Stimuli were presented on a Dell 1703FP monitor with the refresh rate set at 85 Hz. Reaction time measures were based on standard keyboard responses. Participants were seated approximately 57 cm from the monitor, and the experimenter ensured that participants were centred with respect to the monitor. The experiment was controlled by Experiment Builder (SR Research Ltd, Ontario, Canada).

tDCS stimulation

An ActivaDose II, ActivaTek system was used to administer the 2 mA stimulation. Although there is not a universal and ideal level of current stimulation for

any one montage, we selected 2 mA because it is a commonly used level that is within human safety limits (Bikson, Datta, & Elwassif, 2009). Electrode placement followed the 10–5 electroencephalogram (EEG) system (Oostenveld & Praamstra, 2001), and tDCS was delivered via two 5 x 5 cm saline-soaked sponges in rubber housing (resulting in a sponge contact area of 3 x 3 cm). For PFA stimulation, the anode was placed on the scalp over F9 and the cathode was placed over Fz. For TPA stimulation, the anode was placed on the scalp over P5 and the cathode was placed extracephalically on the right (Blumberg et al., 2014; Falcone, Coffman, Clark, & Parasuraman, 2012). This montage was chosen to minimize erroneous stimulation of non-targeted cortical areas based on brain modelling.

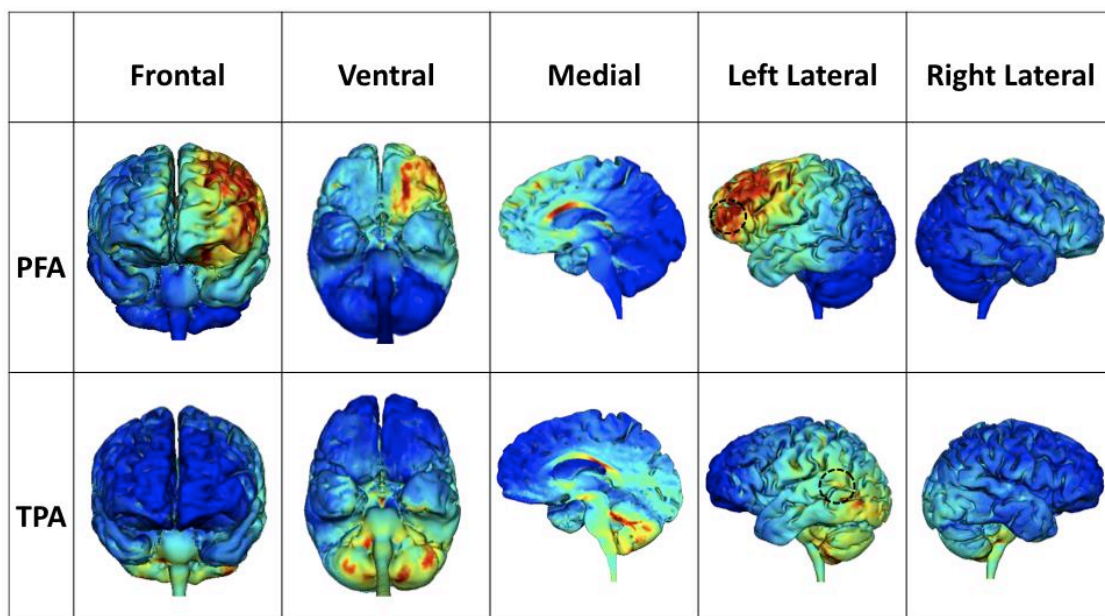


Figure 17. Brain models illustrate the field intensities of stimulation in V/m^{21} per mA. Darker red areas illustrate higher stimulation intensities compared with blue areas. The circled regions in the ‘left lateral’ pane show the PFA region and the TPA region, respectively.

Brain modelling

Stimulation parameters were modelled using Finite-Element Models (FEMs) within the HD Explore brain modelling software (Soterix Medical, NY, USA). Brain models were obtained based on conductivities from Datta et al. (2012) The brain models for PFA and TPA stimulation were created for 5 x 5 cm sponges (as required by Soterix software). However, because the sponges were reduced to 3 x 3 cm owing to the use of sponge housings, we used a 160-channel EEG cap to measure the number of electrodes that were not targeted by the 3 x 3 cm sponges in order to correct the initial brain models by reducing the number of electrodes stimulated in the brain modelling software to provide the best approximation of the stimulation intensity at each area. PFA stimulation reached multiple prefrontal structures, including the left medial prefrontal cortex (lmpFC), left dorso- lateral prefrontal cortex (ldlPFC), left dorsomedial prefrontal cortex (ldmPFC), left ventromedial prefrontal cortex (lvmPFC) and left anterior cingulate cortex (lACC); Soterix showed an average peak stimulation intensity of 0.34 V/m^{21} per 1 mA. For TPA stimulation, structures that received stimulation included the left temporo-parietal junction (lTPJ), left superior temporal sulcus (lSTS), left precuneus (lPRC) and left fusiform face area (lFFA): Soterix showed an average peak stimulation intensity of 0.33 V.m^{21} per 1 mA (see figure 17 for brain models and electronic supplementary

materials for exact stimulation intensities of each structure). Although multiple models were evaluated, the electrode montages for PFA and TPA stimulation chosen demonstrated the greatest peak stimulation intensity over the targeted brain regions, with the least amount of stimulation to non-target brain regions. Specifically, FEM modelling illustrated that using an extracephalic cathodal electrode reduces erroneous stimulation for non-targeted cortical areas in the TPA stimulation condition. It is important to note that although recent evidence suggests that an extracephalic electrode may reduce the magnitude of the stimulation effect when compared with scalp placements (Moliadze, Antal, & Paulus, 2010), this is not the case in the current experiment as FEM models suggest that similar stimulation intensities were achieved for PFA and TPA stimulation (0.34 V/m^{21} versus 0.33 V/m^{21} per 1 mA).



Figure 18. Human and robot stimuli. The human agent is represented by a female face taken from the Karolinska Directed Emotional Faces (KDEF) data- base (F07; written informed consent from the Karolinska

Institute was received to use the photograph for experimental investigations and illustrations). The robot agent is the robot EDDIE (developed at the Technical University of Munich, Germany).

Stimuli

Gaze cueing requires participants to detect, locate or identify targets that are looked at or looked away from by a gazer (Friesen & Kingstone, 1998), which in the current study is either a human (mind perception high) or a robot (mind perception low). In the human condition, the digitized photo of a female face was used as a gazer, which can be found in the Karolinska Directed Emotional Faces database, while in the robot condition the photo of a humanoid robot was used as a gazer (EDDIE; developed by the Technical University of Munich, Germany). The gazing stimuli were 6.48 wide and 10.08 high, depicted on a white background and presented in full frontal orientation with eyes positioned on the central horizontal axis of the screen (figure 18). For left- and rightward gaze, irises and pupils of the human and the robot gazer were shifted with Photoshop and deviated 0.48 from direct gaze. The target stimulus was a black capital letter (F or T; measuring 0.88 in width and 1.38 in height), which participants had to discriminate by pressing assigned keys on a regular keyboard. The target letters appeared on the horizontal axis of the screen and were located 6.08 left or right from the centre of the screen.

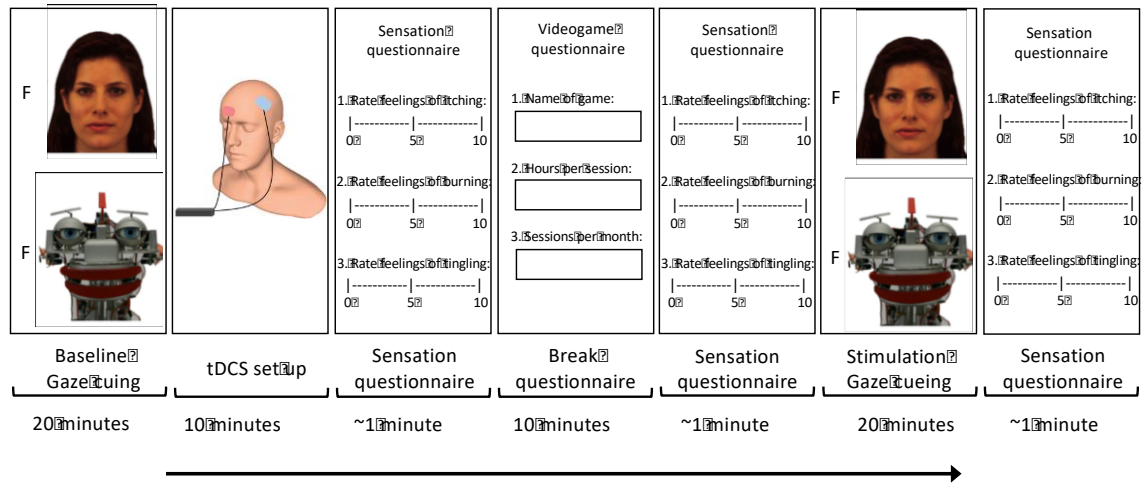


Figure 19. Timing of the experimental procedure. The experiment started with participants completing the baseline gaze-cueing task. Next the researcher set up the tDCS machine and participants completed a questionnaire about their sensations. The participants then completed a decoy survey, which asked about their video- game experience. Participants then completed a second sensation questionnaire followed by a gaze-cueing task under stimulation. After the stimulation gaze cueing task was completed, a final sensation questionnaire was administered and the tDCS stimulation was stopped.

Procedure

At the beginning of the experiment, participants gave informed consent and completed the Snellen near-sightedness exam to test their eye vision. They then answered a set of questionnaires to assess their perception of robots from the Godspeed measure (Bartneck, Croft, & Kulic, 2008), attitudes towards robots from the Negative Attitude Towards Robots questionnaire (Nomura, Kanda, & Suzuki, 2006) and autistic traits from the Autism quotient (Baron-Cohen & Wheelwright, 2004). Upon completion of the questionnaires, participants were instructed how to perform the gaze- cueing task, and were informed that the experiment consisted of two parts: a first part where they would perform the gaze- cueing task without stimulation, and a second part, where they would

perform the gaze-cueing task under stimulation, with a break for setting up the tDCS equipment in between.

After completing the baseline block, which took 20 min, the researcher started setting up the tDCS equipment, which took about 10 min. As soon as the current reached its maximum value of 2.0 mA, participants completed a sensation questionnaire to monitor their comfort levels. They were given an unrelated videogame experience questionnaire following the first sensation questionnaire. The unrelated video game questionnaire was administered to ensure that the timing of the stimulation was similar to previous studies that successfully modulated cognitive processes using tDCS (Berryhill, 2014; Blumberg et al., 2014; Boggio, Rocha, da Silva, & Fregni, 2008; Ferrucci et al., 2008; Javadi, Cheng, & Walsh, 2012). This was an important step as there is no unified standard for the use of tDCS modulation of cognitive tasks (Nitsche et al., 2008). After completing the unrelated videogame questionnaire, participants completed a second sensation questionnaire. Next, subjects started the stimulation gaze-cueing block, which also took 20 min. After the stimulation gaze-cueing block, a third sensation questionnaire was administered, the electrodes were removed, the GSM and NARS questionnaire were administered again, and the participants were debriefed. The timing of the experiment can be viewed in figure 19.

The sequence of events on a given trial of gaze cueing is illustrated in figure 20. The beginning of each trial was signaled by a fixation cross at the center of the screen. Between 700 and 1000 ms later, one of the agents (i.e. human or robot) appeared on the screen looking straight (and with the fixation cross remaining in its position). After a

random time interval of 700 – 1000 ms, the gazer shifted its gaze either left- or right-wards, which constituted the gaze cue. After a stimulus onset asynchrony (SOA; i.e. time interval between gaze cue and target onset) of 400–600 ms, a target letter (F or T) appeared on the left or the right side of the screen, and participants were asked to respond as fast and accurately as possible to the identity of the target (Friesen & Kingstone, 1998). The gaze cue and target letter remained on the screen until a response was given or after a timeout of 1200 ms was reached, whichever came first. The next trial started after an inter-trial interval (ITI) of 680 ms.

Participants were instructed to fix their gaze on the fixation cross at the beginning of a trial, and to not make any eye movements during the trial. They were also instructed that after the fixation cross the image of a social agent would appear in the center of the screen, which would first look at them (mutual gaze), and then after some time make an eye movement to look to the left or right side of the screen (averted gaze). Participants were further advised that the change in gaze direction would be followed by the appearance of a target letter (F or T), which would appear either at the gazed-at location or at the opposite of the gazed-at location. Participants were asked to indicate as quickly and accurately as possible whether ‘F’ or ‘T’ was shown on the screen by pressing the respective response key: for one half of the participants ‘F’ was assigned to the ‘D’ key and ‘T’ to the ‘K’ key on a regular keyboard, while for the other half of participants stimulus– response mapping was reversed. The original key labels were covered with a sticker to prevent interference effects with the actual letters on the keyboard. All instructions were given in written form.

Gaze direction (left, right), target location (left, right), target identity (F, T) and agent type (human, robot) were selected pseudo-randomly; every combination appeared with equal frequency throughout the experiment. Gaze direction was manipulated orthogonally to target location: in half of the trials, the target was validly cued, and in the other half of the trials, the target was invalidly cued. Each experimental session was composed of 220 trials, with a block of 20 practice trials preceding two gaze-cueing blocks of 100 trials of gaze cueing each (i.e. 220 trials for the stimulation baseline and 220 trials for the stimulation block). Participants first completed one block of gaze cueing without stimulation, and were then assigned to either the group that received active stimulation to left PFA or the group that received active stimulation to left TPA. Stimulation was applied for 30min, during which participants completed the second block of gaze cueing.

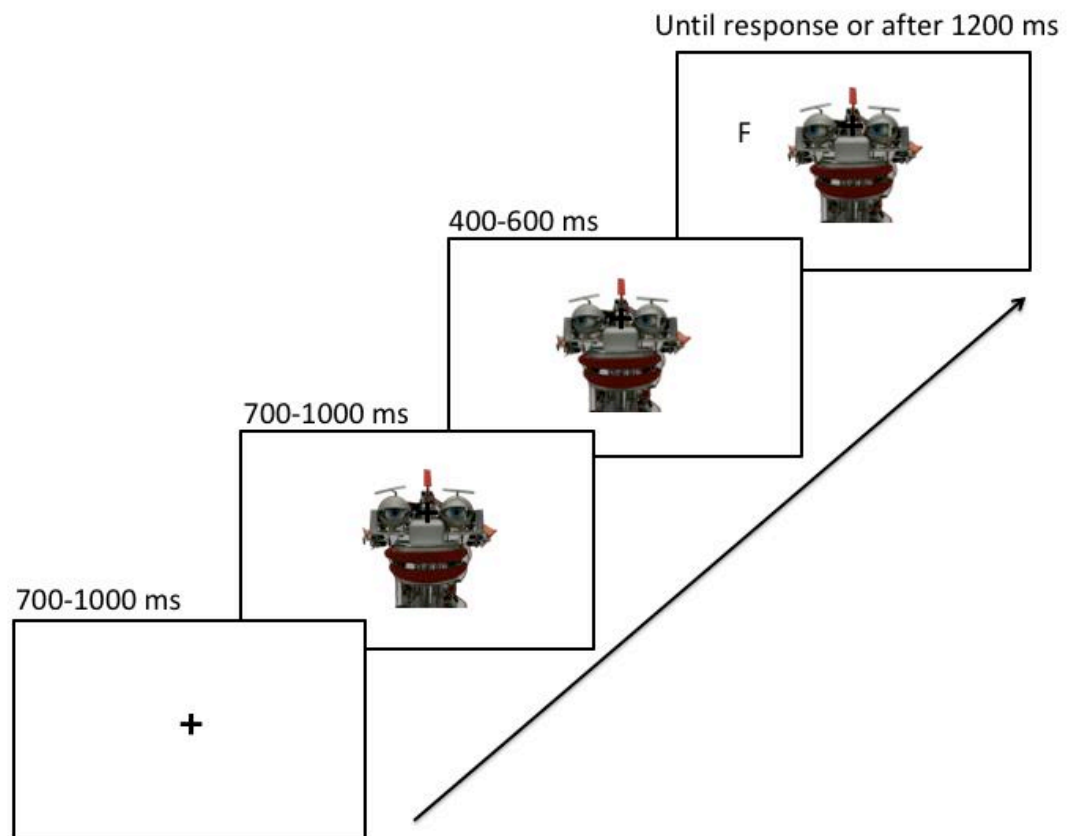


Figure 20. Sequence of events on a trial of gaze cueing. Participants first fixated on a fixation cross for 700–1000 ms and were then presented with the gazing agent (human versus robot) looking straight for 700 – 1000 ms, followed by a change in gaze direction (to either the left or right side of the screen). After an SOA of 400–600 ms, the target letter (F or T) appeared either where the face was looking or opposite to where the face was looking. The target remained on the screen until a response was given or a timeout of 1200 ms was reached.

Results

Questionnaires

We used the Godspeed measure (GSM; (Bartneck et al., 2008) and the Negative Attitude Towards Robots Scale (NARS; (Nomura et al., 2006) to determine whether participants in the two different stimulation conditions differed in their perception of and

attitudes towards robots. Both questionnaires were administered at the beginning of the experiment (to capture potential a priori individual differences), as well as after the completion of the experiment (to assess whether attitudes changed). The short version of the Autism Quotient (AQ-short; (Baron-Cohen, Wheelwright, Skinner, Martin, & Clubley, 2001)) was administered once, at the beginning of the experiment, to measure participants' autistic traits. A 10-point sensation questionnaire was also administered three times to monitor participants' comfort levels during the experiment (Blumberg et al., 2014; Falcone et al., 2012). Participants were told that reporting a '7' would illustrate unbearable sensations. The questionnaire was administered as soon as stimulation started, before the start of the stimulation block, and after completing the stimulation block. Participants in the PFA and the TPA stimulation did not differ in terms of their perception of robots (GSM; $F_{1,69} = 0.09$, $p = 0.75$, $\eta^2 < 0.001$), attitudes towards robots (NARS; $F_{1,69} = 0.2$, $p = 0.65$, $\eta^2 < 0.01$) or autistic traits (AQ score; $t_{69} = 21.66$, $p = 0.09$). Comparison of the GSM and NARS pre – post ratings did not reveal differences between the two stimulation conditions (GSM: $F_{1,69} = 1.19$, $p = 0.27$, $\eta^2 < 0.01$; NARS: $F_{1,69} = 0.04$, $p = 0.84$, $\eta^2 < 0.001$); see the electronic supplementary material, tables S1 and S2 for the full report of the inference statistics. Two participants indicated sensation levels of above '7', which indicated that they were uncomfortable and did not continue the experiment.

Behavioural data

To determine whether active PFA and/or TPA stimulation had a modulatory effect on gaze cueing, we conducted a 2 x 2 x 2 mixed ANOVA on gaze-cueing effects (i.e. mean reaction times for invalid–valid trials) with Agent type (human versus robot)

and Session (baseline versus stimulation) as within-participants factors, and Brain site (left PFA versus TPA) as between- participants factors. The descriptive statistics of this analysis are depicted in figure 21.

Before examining the results of the statistical models, we tested the normality of the residuals using the Kolmogorov– Smirnov test (and not the more commonly used Shapiro – Wilk test as is it not recommended for larger sample sizes; (Park, 2013; Rani Das & Imon, Rahmatullah, 2016). The Kolmogorov – Smirnov test revealed a non-significant effect ($D = 0.07$, $p = 0.42$), showing that the residuals of our data did not differ significantly from a normal distribution (i.e. the normality of the residuals assumption of parametric tests was not violated).

The ANOVA revealed no main effect of Agent type ($F_{1,70} = 0.2$, $p = 0.64$, $\eta^2 < 0.001$), indicating that across stimulation sites and sessions, there were no significant differences in gaze-cueing effects for the human agent compared with the robot agent (human: 8.22 ms versus robot: 7.20 ms). The main effect of Session ($F_{1,70}$, 0.001, $p = 0.98$, $\eta^2 < 0.001$) was also not significant, showing that across stimulation sites and agent types gaze-cueing effects did not differ between baseline and stimulation (baseline: 7.69 ms versus stimulation: 7.74 ms). The main effect of Brain site ($F_{1,70}$, 0.001, $p = 0.98$, $\eta^2 < 0.001$) was also not significant, showing that across agent types and sessions, no differences in gaze-cueing effects were found between the two stimulation sites (PFA: 7.69 ms versus TPA: 7.74 ms). All two-way interactions were also not significant (Agent type x Session: $F_{1,70} = 0.61$, $p = 0.43$, $\eta^2 < 0.001$; Agent type x Brain site: ($F_{1,70}$ 2.24, $p = 0.13$, $\eta^2 < 0.01$); Session x Brain site: ($F_{1,70} = 0.17$, $p = 0.67$, $\eta^2 < 0.001$). Most

importantly, however, the three-way interaction of Agent type x Session x Brain site was significant ($F_{1,70} = 4.50$, $p = 0.03$, $\eta^2 = 0.12$), indicating that tDCS stimulation affected gaze-cueing effects differently for the human versus the robot condition under left PFA but not left TPA stimulation.

To examine the significant three-way interaction effect further, two 2×2 post hoc ANOVAs with Agent type (Human versus Robot) and Session (Baseline versus Stimulation) were conducted, one for left PFA stimulation and one for left TPA stimulation. The ANOVA for the PFA stimulation condition showed no main effects of Agent type ($F_{1,35} = 2.21$, $p = 0.14$, $\eta^2 < 0.01$) or Session ($F_{1,35} = 0.08$, $p = 0.77$, $\eta^2 < 0.001$), but a significant Agent type x Session interaction ($F_{1,35} = 5.51$, $p = 0.02$, $\eta^2 = 0.02$). Post hoc paired t-tests revealed that there was no significant difference in gaze-cueing effects between the human and robot agents at baseline (human: 7.47 ms versus robot: 8.66 ms; $p = 0.75$), but significantly larger cueing effects for the human versus the robot gazer under stimulation (human: 12.25 ms versus robot: 2.34 ms; $p = 0.02$). By contrast, the ANOVA for the TPA stimulation condition revealed neither main effects of Agent type ($F_{1,35} = 0.47$, $p = 0.49$, $\eta^2 < 0.01$) or Session ($F_{1,35} = 0.09$, $p = 0.76$, $\eta^2 < 0.01$), nor a significant Agent type x Session interaction ($F_{1,35} = 0.72$, $p = 0.39$, $\eta^2 < 0.01$). This suggests that while PFA stimulation modulated gaze-cueing effects with significantly larger gaze-cueing effects for the human versus the robot agent under stimulation, TPA stimulation did not have such a modulatory effect on gaze cueing (i.e. no differences in gaze cueing at baseline and under stimulation). Post hoc t-tests were corrected using the false discovery rate (FDR) procedure.

A separate 2 x 2 x 2 mixed ANOVA was conducted on accuracy ratings of participants. The ANOVA revealed no significant main effects or interaction effects. See the ‘Behavioral Results’ section of the electronic supplementary materials for inference statistics. All data and stimuli can be publicly viewed on <https://osf.io/s8ewg/>.

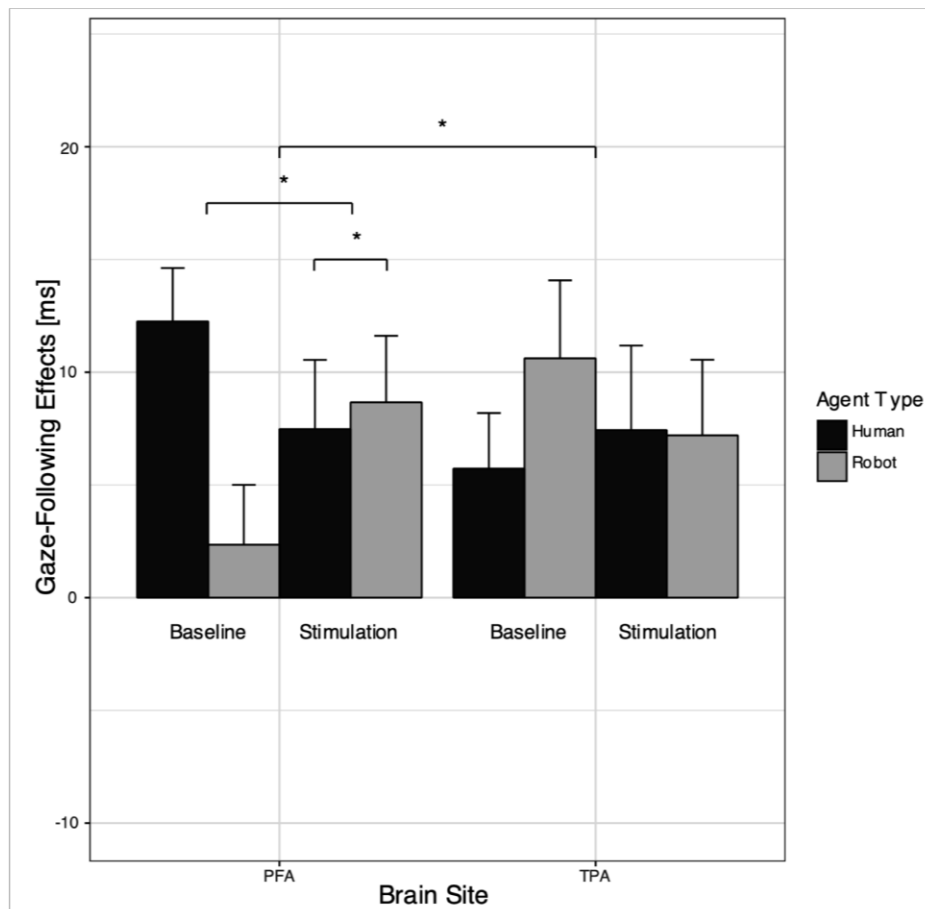


Figure 21. Gaze-cueing effects (in ms) as a function of Brain site (left PFA, left TPA), Session (baseline, stimulation) and Agent type (human, robot). There was a significant change in gaze cueing for active PFA stimulation, with no differences in gaze cueing between human and robot at baseline, but significantly larger gaze- cueing effects for the human versus the robot agent under stimulation. Active TPA stimulation did not have significant effects on gaze cueing (*p , 0.05).

Discussion

Previous studies have shown that activation in left prefrontal (Chaminade et al., 2012; Gallagher et al., 2002; Sanfey et al., 2003; Schurz, Radua, Aichhorn, Richlan, & Perner, 2014; Van Overwalle, 2009; Wiese et al., 2018) and temporo-parietal (Cullen et al., 2013; Gallagher et al., 2000; Sanfey et al., 2003; Van Overwalle, 2009) areas is related to mind perception and the modulation of low-level social-cognitive processes like gaze cueing (Özdem et al., 2016; Wiese et al., 2018, 2017). The goal of the current experiment was to examine the causal involvement of left prefrontal and left temporo-parietal areas in the top-down modulation of social attention via mind perception. To address this issue, we asked participants to perform a gaze-cueing task with a human and a robot agent (i.e. manipulation of mind perception via physical human-likeness) while applying tDCS to left prefrontal and left temporo-parietal areas. The findings show that stimulating left prefrontal had a modulatory effect on social attention, such that gaze cues of intentional agents (i.e. human) were followed significantly more strongly than gaze cues of machine agents (i.e. robot) under stimulation of prefrontal areas. Left temporo-parietal stimulation, in contrast, did not significantly modulate gaze-cueing effects for either of the two agents. These results are in line with previous studies showing that experimental manipulations of mind perception enhance the degree to which gaze signals are followed (Özdem et al., 2016; Wiese et al., 2012; Wykowska et al., 2014). In particular, it was shown that interpreting gaze signals as intentional or human-controlled augments sensory processing of stimuli presented at gazed-at locations

(Wykowska et al., 2014), and increases the social relevance of observed gaze signals (Caruana, de Lissa, et al., 2017; Caruana, de Lissa, & McArthur, 2015; Wiese, Wykowska, et al., 2014, 2014). The results are also in line with studies showing a correlational relationship between activation in left prefrontal areas related to mind perception and modulation of social attention when performing an orthogonal gaze-cueing task outside the fMRI scanner (Wiese et al., 2018), as well as tDCS studies showing that temporo-parietal areas are causally involved in social-cognitive processes like imitation and perspective taking but might not be causally involved in mind perception (Hogeveen et al., 2015; Santiesteban, Banissy, Catmur, & Bird, 2012; Santiesteban et al., 2015).

The current experiment adds to previous findings by localizing the source of top-down modulation of social attention via mind perception to left prefrontal areas, including areas like the ACC and vmPFC. These areas are associated with mentalizing (Amodio & Frith, 2006; Gallagher et al., 2002), and involved in impression formation in social interaction (J. P. Mitchell, 2004; Szczepanski & Knight, 2014). In particular, activation in mPFC is linked to retrieving stereotypical knowledge about other people (Contreras, Banaji, & Mitchell, 2012; Fairhall, Anzellotti, Ubaldi, & Caramazza, 2014; Simmons, Reddish, Bellgowan, & Martin, 2010), and associated with retrieving script-based social knowledge (Ghosh, Moscovitch, Melo Colella, & Gilboa, 2014; Van Kesteren, Ruiter, Fernández, & Henson, 2012). Medial prefrontal areas are also more strongly activated when mentalizing about the internal states of similar than dissimilar others (Jenkins, Macrae, & Mitchell, 2008; J. P. Mitchell, 2004; Jason P. Mitchell,

Banaji, & Macrae, 2005; Jason P. Mitchell, Macrae, & Banaji, 2006), as well as when viewing social scenes that contain human versus non-human agents (Wagner et al., 2011). The current experiment adds to the literature by indicating that prefrontal areas might not only be involved in the modulation of higher-order social-cognitive processes like decision-making (Gallagher & Frith, 2003; Gallagher et al., 2002, p. 2; Sanfey et al., 2003), but might also exert modulatory effects on low-level social cognitive processes like social attention.

Note that the current experiment does not show the previously reported difference in gaze cueing between human and robot gazers at baseline (Wykowska et al., 2014). This could be due to several reasons. First, in previous experiments, the gazers were introduced as ‘human’ versus ‘robot’ via instruction, which provided participants with explicit labels as to how to treat them in terms of mind perception (i.e. ‘human/has a mind’ and ‘robot/ has no mind’). By contrast, in the current experiment, the gazers were introduced more neutrally as ‘agents’ and their mind status had to be inferred from physical appearance, which makes the mind perception manipulation more implicit. Although this certainly increases the external validity of the experimental manipulation, it is possible that participants did not pay enough attention to the gazers’ mind status, which could wash out effects between the two agents at baseline. Second, because mind perception was manipulated via physical appearance in the current experiment, it is also possible that individual differences in anthropomorphism (Cullen et al., 2013; Hackel, Looser, & Van Bavel, 2014) [54,134] attenuated differences in gaze cueing between the human and robot gazers at baseline. However, because baseline effects are comparable

for both stimulation conditions and because the current paper is mainly interested in the modulation of social attention via mind perception, insignificant differences in gaze cueing at baseline should not have impacted the reported findings. Nevertheless, in order to validate the robustness of the reported findings, future experiments should be conducted to determine to what degree they might be influenced by individual differences in anthropomorphism.

The question remains why stimulation of left temporo- parietal networks did not significantly modulate low-level mechanisms of social cognition despite previous reports showing a correlational relationship between mechanisms of social attention and bilateral TPJ activation (Özdem et al., 2016). One explanation for the lack of a significant effect of left temporo-parietal stimulation on gaze cueing is that it is possible that processes related to mind perception and social attention are not sufficiently interconnected at the level of the left TPJ in order to exert a top-down modulatory effect on attentional orienting to gaze cues. This interpretation would be in line with previous reports showing that social functions within the TPJ are lateralized (Perner et al., 2006), and that an overlap between attentional orienting and mentalizing is found within the right but not left TPJ (Bzdok et al., 2013; Krall et al., 2015; Kubit & Jack, 2013). By contrast, left TPJ lesions have been shown to cause selective deficits in false belief reasoning (Apperly, Samson, Chiavarino, & Humphreys, 2004; Hackel et al., 2014), which does require mentalizing but no orientation of social attention. In support of this notion, it has been shown that early posterior ERP components like the N170 are sensitive to the intentionality of an agent without being responsive to the congruency or social outcome of its gaze cues (Caruana

et al., 2015), whereas later anterior ERP components like the P350 are sensitive to both an agent's intentionality and the congruency of gaze cues (Caruana, McArthur, et al., 2017) (i.e. significant difference in P350 amplitudes for invalid versus valid trials for human versus computer-controlled conditions), suggesting that the integration of mind perception related processes and social attention might be instantiated in prefrontal (but not temporo-parietal) areas.

Another possible explanation is that prefrontal and temporo- parietal brain regions might process information about an agent's mind on different levels, with TPJ activation being related to inferring particular internal states from observed behaviors (Bzdok et al., 2013; Rebecca Saxe & Powell, 2006; Schurz et al., 2014; Van Overwalle & Baetens, 2009) (e.g. observing an agent smile leads to the inference that the agent is currently in a state of happiness), and mPFC and vmPFC activation being related to reasoning based on stereotypical assumptions regarding general traits associated with intentional agents (Wang & Quadflieg, 2015) (e.g. agent that looks like a child might like toys). It could be possible that studies that manipulate mind perception via instruction of particular beliefs (i.e. 'eye movements are intentional') engage the posterior mind perception network (Wykowska et al., 2014), while studies that manipulate mind perception via physical appearance (i.e. the agent looks human) more strongly activate stereotypical assumptions about human behavior, thereby engaging the anterior mind perception network involving the mPFC and vmPFC (Contreras et al., 2012; Fairhall et al., 2014; Simmons et al., 2010). This interpretation is particularly plausible given that attentional orienting to gaze cues is a fast-acting process (Friesen & Kingstone, 1998), which requires information

from a readily available source in order to be top-down controlled. Since stereotypical information about an agent is more readily available than the outcomes of mentalizing processes about particular internal states, a stronger involvement of prefrontal areas in the top-down modulation of fast processes like social attention seems tenable.

A third explanation is that the observed modulation of gaze cueing is not specific to mind perception, but due to other (related) functions associated with the left prefrontal cortex. One aspect of the experimental design that could have affected social attention in addition to the gazer's physical appearance is the unpredictability of gaze cues in the current experiment (i.e. targets appear with equal frequency at validly and invalidly cued locations). Previous studies have shown that stimuli whose behavior is hard to predict are more likely anthropomorphized, and that evaluating unpredictable stimuli is associated with increased activation in medial prefrontal areas, and specifically the vmPFC and ACC (Waytz et al., 2010). In consequence, it is possible that one's sensitivity to the predict-ability of gaze cues had an impact on the degree to which the gazer was anthropomorphized and prefrontal brain areas were activated during gaze cueing. For the current experiment, this means that stimulating prefrontal areas may have specifically enhanced the social relevance of human gaze cues, leading to longer processing times on invalid trials (Caruana, McArthur, et al., 2017) and larger gaze-cueing effects (i.e. difference in reaction times between invalid and valid trials), while temporo-parietal stimulation may not have affected the perceived social relevance of human gaze cues.²

Alternatively, the vmPFC has been shown to track feelings of eeriness towards non-human agents in a parametric fashion (Wang & Quadflieg, 2015) and has been

labelled the potential neural correlate of the uncanny valley (Mori, MacDorman, & Kageki, 2012) (i.e. non-human agents with human-like appearance induce feelings of eeriness if they are not perfectly human). If that were the case, stimulation of prefrontal areas could have enhanced feelings of eeriness towards the robotic agent, leading to a disengagement from robot gaze cues together with an increased engagement in attending to human gaze cues (i.e. eeriness of robot cues made human gaze cues more ‘desirable’). Effects related to non-social pre- frontal functions such as working memory, executive functioning or abstract reasoning are less likely to have influenced gaze cueing, because one would expect comparable effects of stimulation for human and non-human agents. Whether prefrontal stimulation modulated gaze cueing directly via mind perception or via processes affected by mind perception, such as perception of uncertainty (Caruana, McArthur, et al., 2017) or emotional reactions to uncanny agents (Quadflieg, Mason, & Macrae, 2004), cannot ultimately be answered based on the current data and requires follow-up studies. It can also not be clearly determined—owing to the lack of spatial specificity of tDCS—which prefrontal brain area(s) ultimately caused the observed top-down modulation of social attention (i.e. areas directly implicated in mind perception such as the left ACC, or areas that are indirectly involved in mind perception such as the left mPFC, vmPFC or dlPFC).

Conclusion

Previous studies have shown that the degree to which we attend to social cues depends on the degree to which we perceive mind in the entity sending the cues. The neural correlates of mind perception have been localized to prefrontal and temporo-

parietal structures in previous studies, but the causal involvement of these areas in the modulation of low-level social-cognitive processes like gaze cueing has not been determined yet. The current study shows that stimulation to prefrontal areas increases the degree to which human gaze is followed compared with the degree to which robot gaze is followed, while stimulation to temporo-parietal regions does not seem to have a measurable modulatory effect on gaze cueing. Since the effect of prefrontal stimulation is only observable for human gazers, it is tenable that prefrontal stimulation does not simply lead someone to perceive ‘more’ mind in others, but rather seems to enhance the social relevance of signals coming from agents ‘with a mind’. In other words, prefrontal stimulation does not seem to make participants perceive more human-likeness in non-human agents, which makes it unlikely that the observed effect is related to anthropomorphism. Instead, prefrontal stimulation seems to help discriminate agents ‘with a mind’ from agents ‘without a mind’, as evidenced by an increased difference in gaze cueing between the human and the robot gazer under stimulation, indicating that stimulation of prefrontal areas enhances the importance of social signals coming from human agents. Taken together, this study shows a causal link between prefrontal stimulation and mechanisms of social attention, and dissociation between the anterior and posterior part of the social brain network in terms of top-down modulation of social-cognitive processes. Whether the effect is specific to mind perception or related to processes indirectly affected by mind perception needs to be determined in future studies.

GENERAL DISCUSSION/SYNOPSIS

The purpose of this dissertation is to causally link mentalizing brain structures that are associated to mind perception to socio-cognitive processing. Specifically, we were interested in how manipulations of mind perception through physical appearance can affect social attention as predicted by activation in the mentalizing brain network. By doing so, we are able to understand how top-down modulators such as mind perception can exert influence in a social attention task.

In Study 1 we found that physical appearance was related to subjective mind perception ratings but could not predict gaze-cueing performance, which could be due to physical appearance's dependence on observed mind perception ratings. While previous studies have shown links between mind perception and the mentalizing network of the brain (Gallagher et al., 2002; Harris & Fiske, 2011; Krach et al., 2008; Sanfey et al., 2003), Study 2 established a link between mind perception and gaze-cueing performance through activation in the mentalizing region, which is a consistent finding related to previous studies. In the last study, we examined a causal link between activation in prefrontal structures that were identified by Study 2 and gaze behavior in a social attention task. The findings of the last study illustrated how stimulation to the vmPFC modulated behavioral responses to only the human faces but not the robots, which suggests that top-down influence of mind perception from vmPFC could be related to participants' prior experiences with agents with a mind (i.e., other humans).

The findings of these studies are consistent with previous literature that illustrate how manipulating physical appearance can influence ratings of mind (Hackel et al., 2014; Martini, Gonzalez, & Wiese, 2016), and that the vmPFC is associated with judgements of intentionality (Gallagher et al., 2002; Pfeiffer et al., 2014; Sanfey et al., 2003; Van Overwalle, 2009). Since the vmPFC is related to knowledge based intentionality (Van Overwalle, 2009), it is not surprising that participants performed better for only the human and not the robot. This provides important considerations for studies that examine the effects of long-term experiences between human and robot interaction partners.

Together these findings allow us to understand how the social brain reacts when interacting with nonhuman agents that are equipped with the means to trigger mind perception, namely robots that are equipped with human-like appearance. The findings suggest that robotic designers should use human physical appearance as a mind trigger as it has shown to be successful in eliciting social gaze behavior. However, designers should approach with caution as studies have also shown that some robots who look too similar to humans can provoke feelings of uncanniness (Mathur & Reichling, 2016; Mori, 1970). Another major implication of studies exploring the effects of mind triggers on socio-cognitive processes in human-robot interaction is providing robotic designers with a basis for building robots that successfully engage the social brain network.

Although these studies help with identifying the source of top-down modulation of physical appearance on gaze behavior, future studies should focus on individual differences in how people recruit brain structures differently. For example, individuals diagnosed with the Autism Spectrum Disorder (i.e., ASD) have shown differences in

gaze performance compared to individuals who are undiagnosed (Wiese, Müller, & Wykowska, 2014). Future directions should also focus on investigating the timing effects of these modulations. Understanding the timing effects can help with pinpointing event related potentials (i.e., ERPs) that are related to mind attribution as it relates to human-robot interactions. Identifying different ERP components can inform designing adaptive robots and automation that dynamically attune their behavior based on the interacting human's brain activity. Although neuroimaging techniques are imperative for identifying the locus of these modulations in the brain, ERPs are a more suitable brain measurement as it does not introduce delays that are associated with neuroimaging imaging techniques such as fMRI and fNIRS.

REFERENCES

- Abell, F., Happé, F., and Frith, U. (2000). Do triangles play tricks? Attribution of mental states to animated shapes in normal and abnormal development. *Cogn. Dev.* 15, 1–16. doi: 10.1016/S0885-2014(00)00014-9
- Abubshait, A., & Wiese, E. (2017). You look human, but act like a machine: Agent appearance and behavior modulate different aspects of human–robot interaction. *Frontiers in Psychology*, 8. <https://doi.org/10.3389/fpsyg.2017.01393>
- Adams, R. B., & Kleck, R. E. (2005). Effects of direct and averted gaze on the perception of facially communicated emotion. *Emotion*, 5(1), 3–11. <https://doi.org/10.1037/1528-3542.5.1.3>
- Admoni, H., Bank, C., Tan, J., Toneva, M., and Scassellati, B. (2011). “Robot gaze does not reflexively cue human attention,” in *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*, eds L. Carlson, C. Hölscher, and T. Shipley (Austin, TX: Cognitive Science Society), 1983–1988.
- Adolphs, R. (1999). Social cognition and the human brain. *Trends in Cognitive Sciences*, 3(12), 469–479. [https://doi.org/10.1016/s1364-6613\(99\)01399-6](https://doi.org/10.1016/s1364-6613(99)01399-6)
- Adolphs, R. (2009). The social brain: Neural basis of social knowledge. *Annual Review of Psychology*, 60(1), 693–716. <https://doi.org/10.1146/annurev.psych.60.110707.163514>

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: The medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, 7(4), 268–277.

<https://doi.org/10.1038/nrn1884>

Anderson, S. W., Bechara, A., Damasio, H., Tranel, D., & Damasio, A. R. (1999).

Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience*, 2(11), 1032–1037.

<https://doi.org/10.1038/14833>

Antal, A., Nitsche, M. A., & Paulus, W. (2001). External modulation of visual perception in humans. *Neuroreport*, 12(16), 3553–3555. <https://doi.org/10.1097/00001756-200111160-00036>

Apperly, I. A., Samson, D., Chiavarino, C., & Humphreys, G. W. (2004). Frontal and Temporo-Parietal Lobe Contributions to Theory of Mind: Neuropsychological Evidence from a False-Belief Task with Reduced Language and Executive Demands. *Journal of Cognitive Neuroscience*, 16(10), 1773–1784.

<https://doi.org/10.1162/0898929042947928>

Balas, B., & Tonsager, C. (2014). Face animacy is not all in the eyes: Evidence from contrast chimeras. *Perception*, 43(5), 355–367. <https://doi.org/10.1068/p7696>

Baron-Cohen, S. (1997). *Mindblindness: An essay on autism and theory of mind*. MIT press.

Baron-Cohen, S., Leslie, A. M., & Frith, U. (1985). Does the autistic child have a “theory of mind”? *Cognition*, 21(1), 37–46. [https://doi.org/10.1016/0010-0277\(85\)90022-](https://doi.org/10.1016/0010-0277(85)90022-8)

- Baron-Cohen, S., & Wheelwright, S. (2004). The empathy quotient: An investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *Journal of Autism and Developmental Disorders*, 34(2), 163–175.
- Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): Evidence from asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, 31(1), 5–17.
<https://doi.org/10.1023/A:1005653411471>
- Bartneck, C., Croft, E., & Kulic, D. (2008). Measuring the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *Metrics for HRI Workshop, Technical Report*, 37–44.
- Bartneck, C., & Reichenbach, J. (2005). Subtle emotional expressions of synthetic characters. *International Journal of Human-Computer Studies*, 62(2), 179–192.
<https://doi.org/10.1016/j.ijhcs.2004.11.006>
- Bartneck, C. (2003). “Interacting with an embodied emotional character,” in *Proceedings of the Design for Pleasurable Products Conference*, Pittsburgh, PA, 55–60.
- Basteris, A., Nijenhuis, S. M., Stienen, A. H., Buurke, J. H., Prange, G. B., & Amirabdollahian, F. (2014). Training modalities in robot-mediated upper limb rehabilitation in stroke: A framework for classification based on a systematic review. *Journal of NeuroEngineering and Rehabilitation*, 11(1), 111.
<https://doi.org/10.1186/1743-0003-11-111>

- Bayliss, A. P., Frischen, A., Fenske, M. J., and Tipper, S. P. (2007). Affective evaluations of objects are influenced by observed gaze direction and emotional expression. *Cognition* 104, 644–653. doi: 10.1016/j.cognition.2006.07.012
- Bayliss, A. P., and Tipper, S. P. (2006). Predictive gaze cues and personality judgments: should eye trust you? *Psychol. Sci.* 17, 514–520. doi: 10.1111/j.1467-9280.2006.01737.x
- Becchio, C., Adenzato, M., & Bara, B. G. (2006). How the brain understands intention: Different neural circuits identify the componential features of motor and prior intentions. *Consciousness and Cognition*, 15(1), 64–74.
<https://doi.org/10.1016/j.concog.2005.03.006>
- Beer, J. S., Heerey, E. A., Keltner, D., Scabini, D., & Knight, R. T. (2003). The regulatory function of self-conscious emotion: Insights from patients with orbitofrontal damage. *Journal of Personality and Social Psychology*, 85(4), 594–604. <https://doi.org/10.1037/0022-3514.85.4.594>
- Bennewitz, M., Faber, F., Joho, D., Schreiber, M., and Behnke, S. (2005). “Towards a humanoid museum guide robot that interacts with multiple persons,” in *Proceedings of 2005 5th IEEE-RAS International Conference on Humanoid Robots* (Piscataway, NJ: IEEE), 418–423.
- Bering, J. M., & Johnson D. (2005). “O lord... You perceive my thoughts from afar”: Recursiveness and the evolution of supernatural agency. *Journal of Cognition and Culture*, 5(1), 118–142. <https://doi.org/10.1163/1568537054068679>

- Berryhill, M. E. (2014). Hits and misses: Leveraging tDCS to advance cognitive research. *Frontiers in Psychology*, 5. <https://doi.org/10.3389/fpsyg.2014.00800>
- Bikson, M., Datta, A., & Elwassif, M. (2009). Establishing safety limits for transcranial direct current stimulation. *Clinical Neurophysiology*, 120(6), 1033–1034. <https://doi.org/10.1016/j.clinph.2009.03.018>
- Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex* (New York, NY), 19(12), 2767–2796. <https://doi.org/10.1093/cercor/bhp055>
- Bisio, A., Sciutti, A., Nori, F., Metta, G., Fadiga, L., Sandini, G., & Pozzo, T. (2014). Motor contagion during human-human and human-robot interaction. *PLOS ONE*, 9(8), e106172. <https://doi.org/10.1371/journal.pone.0106172>
- Blakemore, S. J., Winston, J., & Frith, U. (2004). Social cognitive neuroscience: where are we heading? *Trends Cogn Sci*, 8(5), 216–222. <https://doi.org/10.1016/j.tics.2004.03.012> [pii]
- Blumberg, E. J., Foroughi, C. K., Scheldrup, M. R., Peterson, M. S., Boehm-Davis, D. A., & Parasuraman, R. (2014). Reducing the disruptive effects of interruptions with noninvasive brain stimulation. *Human Factors*, 57(6), 1051–1062. <https://doi.org/10.1177/0018720814565189>
- Boggio, P. S., Rocha, R. R., da Silva, M. T., & Fregni, F. (2008). Differential modulatory effects of transcranial direct current stimulation on a facial expression go-no-go

task in males and females. *Neuroscience Letters*, 447(2–3), 101–105.

<https://doi.org/10.1016/j.neulet.2008.10.009>

Bonifacci, P., Ricciardelli, P., Lugli, L., & Pellicano, A. (2008). Emotional attention:

Effects of emotion and gaze direction on overt orienting of visual attention.

Cognitive Processing, 9(2), 127–135. <https://doi.org/10.1007/s10339-007-0198-3>

Breazeal, C., Kidd, C. D., Thomaz, A. L., Hoffman, G., & Berlin, M. (2005). Effects of

nonverbal communication on efficiency and robustness in human-robot

teamwork. *2005 IEEE/RSJ International Conference on Intelligent Robots and*

Systems, 708–713. <https://doi.org/10.1109/IROS.2005.1545011>

Brothers, L. (2002). The social brain: A project for integrating primate behavior and

neurophysiology in a new domain. In J. T. Cacioppo (Ed.), *Foundations in social*

neuroscience (pp. 367–385). Cambridge, MA: MIT Press.

Bzdok, D., Langner, R., Schilbach, L., Engemann, D. A., Laird, A. R., Fox, P. T., &

Eickhoff, S. B. (2013). Segregation of the human medial prefrontal cortex in

social cognition. *Frontiers in Human Neuroscience*, 7.

<https://doi.org/10.3389/fnhum.2013.00232>

Carter, E. J., Mistry, M. N., Carr, G. P. K., Kelly, B. A., and Hodgins, J. K. (2014).

“Playing catch with robots: Incorporating social gestures into physical

interactions,” in *Proceedings of the IEEE International Symposium on Robot and*

Human Interactive Communication Edinburgh, 231–236.

- Carter, E. J., Hodgins, J. K., & Rakison, D. H. (2011). Exploring the neural correlates of goal-directed action and intention understanding. *NeuroImage*, 54(2), 1634–1642.
<https://doi.org/10.1016/j.neuroimage.2010.08.077>
- Caruana, N., de Lissa, P., & McArthur, G. (2015). The neural time course of evaluating self-initiated joint attention bids. *Brain and Cognition*, 98, 43–52.
<https://doi.org/10.1016/j.bandc.2015.06.001>
- Caruana, N., de Lissa, P., & McArthur, G. (2016). Beliefs about human agency influence the neural processing of gaze during joint attention. *Social Neuroscience*, 12(2), 194–206. <https://doi.org/10.1080/17470919.2016.1160953>
- Caruana, N., McArthur, G., Woolgar, A., & Brock, J. (2017). Simulating social interactions for the experimental investigation of joint attention. *Neuroscience & Biobehavioral Reviews*, 74, 115–125.
<https://doi.org/10.1016/j.neubiorev.2016.12.022>
- Castelli, F., Happé, F., Frith, U., & Frith, C. (2013). Movement and mind: A functional imaging study of perception and interpretation of complex intentional movement patterns. In *Social Neuroscience: Key Readings* (pp. 155–170).
<https://doi.org/10.4324/9780203496190>
- Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: a review of its functional anatomy and behavioural correlates. *Brain*, 129(3), 564–583.
- Cazzato, V., Liuzza, M. T., Caprara, G. V., Macaluso, E., & Aglioti, S. M. (2015). The attracting power of the gaze of politicians is modulated by the personality and ideological attitude of their voters: A functional magnetic resonance imaging

- study. *European Journal of Neuroscience*, 42(8), 2534–2545.
<https://doi.org/10.1111/ejn.13038>
- Cehajic, S., Brown, R., & Gonzalez, R. (2009). What do I care? Perceived ingroup responsibility and dehumanization as predictors of empathy felt for the victim group. *Group Processes and Intergroup Relations*, 12(6), 715–729.
<https://doi.org/10.1177/1368430209347727>
- Chaminade, T., & Decety, J. (2002). Leader or follower? Involvement of the inferior parietal lobule in agency. *Neuroreport*, 13(15), 1975–1978.
<https://doi.org/10.1097/00001756-200210280-00029>
- Chaminade, T., Hodgins, J., & Kawato, M. (2007). Anthropomorphism influences perception of computer-animated characters' actions. *Social Cognitive and Affective Neuroscience*, 2(3), 206–216. <https://doi.org/10.1093/scan/nsm017>
- Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutchter, 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.
<https://doi.org/10.3389/fnhum.2012.00103>
- Chang, C.-F., Hsu, T.-Y., Tseng, P., Liang, W.-K., Tzeng, O. J. L., Hung, D. L., & Juan, C.-H. (2013). Right temporoparietal junction and attentional reorienting. *Human Brain Mapping*, 34(4), 869–877. <https://doi.org/10.1002/hbm.21476>
- Chong, T. T.-J., Cunnington, R., Williams, M. A., Kanwisher, N., & Mattingley, J. B. (2008). fMRI adaptation reveals mirror neurons in human inferior parietal cortex.

- Current Biology: CB, 18(20), 1576–1580.
<https://doi.org/10.1016/j.cub.2008.08.068>
- Cheetham, M., Suter, P., and Jancke, L. (2014). Perceptual discrimination difficulty and familiarity in the uncanny valley: more like a “Happy Valley”. *Front. Psychol.* 5:1219. doi: 10.3389/fpsyg.2014.01219
- Ciardo, F., Marino, B. F. M., Actis-Grosso, R., Rossetti, A., & Ricciardelli, P. (2014). Face age modulates gaze following in young adults. *Scientific Reports*, 4, 4746.
<https://doi.org/10.1038/srep04746>
- Coffman, B. A., Clark, V. P., & Parasuraman, R. (2014). Battery powered thought: Enhancement of attention, learning, and memory in healthy adults using transcranial direct current stimulation. *NeuroImage*, 85, 895–908.
<https://doi.org/10.1016/j.neuroimage.2013.07.083>
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S., (2003). Applied multiple regression/correlation analysis for the behavioral sciences. New York, NY: Routledge.
- Cohen Kadosh, R., Soskic, S., Iuculano, T., Kanai, R., & Walsh, V. (2010). Modulating neuronal activity produces specific and long-lasting changes in numerical competence. *Current Biology*, 20(22), 2016–2020.
<https://doi.org/10.1016/j.cub.2010.10.007>
- Contreras, J. M., Banaji, M. R., & Mitchell, J. P. (2012). Dissociable neural correlates of stereotypes and other forms of semantic knowledge. *Social Cognitive and Affective Neuroscience*, 7(7), 764–770. <https://doi.org/10.1093/scan/nsr053>

- Costa, A., Torriero, S., Oliveri, M., & Caltagirone, C. (2008). Prefrontal and temporo-parietal involvement in taking others' perspective: TMS evidence. *Behavioural Neurology*, 19(1/2), 71–74.
- Cross, E. S., Riddoch, K. A., Pratts, J., Titone, S., Chaudhury, B., & Hortensius, R. (2019). A neurocognitive investigation of the impact of socializing with a robot on empathy for pain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1771), 20180034. <https://doi.org/10.1098/rstb.2018.0034>
- Cui, G., Zhang, S., & Geng, H. (2014). The impact of perceived social power and dangerous context on social attention. *PLoS ONE*, 9(12), 1–15. <https://doi.org/10.1371/journal.pone.0114077>
- Cullen, H., Kanai, R., Bahrami, B., & Rees, G. (2013). Individual differences in anthropomorphic attributions and human brain structure. *Social Cognitive and Affective Neuroscience*, 9(9), 1276–1280. <https://doi.org/10.1093/scan/nst109>
- Cushman, F. (2008). Crime and punishment: distinguishing the roles of causal and intentional analyses in moral judgment. *Cognition* 108, 353–380. doi: 10.1016/j.cognition.2008.03.006
- Dalmaso, Mario, Edwards, S. G., & Bayliss, A. P. (2016). Re-encountering individuals who previously engaged in joint gaze modulates subsequent gaze cueing. *Journal of Experimental Psychology: Learning Memory and Cognition*, 42(2), 271–284. <https://doi.org/10.1037/xlm0000159>

- Dalmaso, M., Galfano, G., and Castelli, L. (2015). The impact of same- and other- race gaze distractors on the control of saccadic eye movements. *Perception* 44, 1020–1028. doi: 10.1177/0301006615594936
- Dalmaso, M., Galfano, G., Coricelli, C., and Castelli, L. (2014). Temporal dynamics underlying the modulation of social status on social attention. *PLoS ONE* 9:e93139. doi: 10.1371/journal.pone.0093139
- Dalmaso, M., Pavan, G., Castelli, L., and Galfano, G. (2012). Social status gates social attention in humans. *Biol. Lett.* 8, 450–452. doi: 10.1098/rsbl.2011. 0881
- Decety, J., & Chaminade, T. (2003). When the self represents the other: A new cognitive neuroscience view on psychological identification. *Consciousness and Cognition*, 12(4), 577–596. [https://doi.org/10.1016/S1053-8100\(03\)00076-X](https://doi.org/10.1016/S1053-8100(03)00076-X)
- Deaner, R. O., Shepherd, S. V., & Platt, M. L. (2007). Familiarity accentuates gaze cuing in women but not men. *Biology Letters*, 3(1), 64–67. <https://doi.org/10.1098/rsbl.2006.0564>
- Deska, J. C., Lloyd, E. P., & Hugenberg, K. (2016). Advancing Our Understanding of the Interface Between Perception and Intergroup Relations. *Psychological Inquiry*, 27(4), 286–289. <https://doi.org/10.1080/1047840X.2016.1215208>
- Dinstein, I., Hasson, U., Rubin, N., & Heeger, D. J. (2007). Brain areas selective for both observed and executed movements. *Journal of Neurophysiology*, 98(3), 1415–1427. <https://doi.org/10.1152/jn.00238.2007>

- DiSalvo, C., & Gemperle, F. (2003). From seduction to fulfillment: The use of anthropomorphic form in design. *International Conference on Designing Pleasurable Products and Interfaces*, 67–72.
- Dodd, M. D., Hibbing, J. R., and Smith, K. B. (2011). The politics of attention: gaze-cueing effects are moderated by political temperament. *Attent. Percept. Psychophys.* 73, 24–29. doi: 10.3758/s13414-010-0001-x
- Dodd, M. D., Hibbing, J. R., and Smith, K. B. (2016). “The politics of attention: differences in visual cognition between liberals and conservatives,” in *Psychology of Learning and Motivation*, Vol. 65, ed. R. Brian (Cambridge, MA: Academic Press), 277–309.
- Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neurosci. Biobehav. Rev.* 24, 581–604. doi: 10.1016/S0149-7634(00)00025-7
- Fong, T., Nourbakhsh, I., and Dautenhahn, K. (2003). A survey of socially interactive robots: concepts, design, and applications. *Rob. Auton. Syst.* 42, 143–166. doi: 10.1016/S0921-8890(02)00372-X
- Epley, N., Waytz, A., Akalis, S., & Cacioppo, J. T. (2008). When we need a human: Motivational determinants of anthropomorphism. *Social Cognition*, 26(2), 143–155. <https://doi.org/10.1521/soco.2008.26.2.143>
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886. <https://doi.org/10.1037/0033-295X.114.4.864>

- Fairhall, S. L., Anzellotti, S., Ubaldi, S., & Caramazza, A. (2014). Person- and place-selective neural substrates for entity-specific semantic access. *Cerebral Cortex*, 24(7), 1687–1696. <https://doi.org/10.1093/cercor/bht039>
- Falcone, B., Coffman, B. A., Clark, V. P., & Parasuraman, R. (2012). Transcranial Direct Current Stimulation Augments Perceptual Sensitivity and 24-Hour Retention in a Complex Threat Detection Task. *PLoS ONE*, 7(4), e34993. <https://doi.org/10.1371/journal.pone.0034993>
- Farrer, C., Franck, N., Georgieff, N., Frith, C. D., Decety, J., & Jeannerod, M. (2003). Modulating the experience of agency: A positron emission tomography study. *NeuroImage*, 18(2), 324–333. [https://doi.org/10.1016/S1053-8119\(02\)00041-1](https://doi.org/10.1016/S1053-8119(02)00041-1)
- Ferrucci, R., Mameli, F., Guidi, I., Mrakic-Sposta, S., Vergari, M., Marceglia, S., ... Priori, A. (2008). Transcranial direct current stimulation improves recognition memory in Alzheimer disease. *Neurology*, 71(7), 493–498. <https://doi.org/10.1212/01.wnl.0000317060.43722.a3>
- Fox, E., Calder, A. J., and Yiend, J. (2007). Anxiety and sensitivity to gaze direction in emotionally expressive faces. *Emotion* 7, 478–486. doi: 10.1037/1528-3542.7.3.478
- Fox, M. D., Snyder, A. Z., Vincent, J. L., & Raichle, M. E. (2007). Intrinsic fluctuations within cortical systems account for intertrial variability in human behavior. *Neuron*, 56(1), 171–184. <https://doi.org/10.1016/j.neuron.2007.08.023>

- Frishen, A., Bayliss, A. P., & Tipper, S. P. (2007). Gaze cueing of attention: Visual attentionn, social cognition, and individual differences. *Psychological Bulletin*, 133(4), 694–724. <https://doi.org/10.1037/0033-2909.133.4.694>.Gaze
- Frishen, A., and Tipper, S. P. (2006). Long-term gaze cueing effects: evidence for retrieval of prior states of attention from memory. *Vis. Cogn.* 14, 351–364. doi: 10.1080/13506280544000192
- Frishen, A., and Tipper, S. P. (2004). Orienting attention via observed gaze shift evokes longer term inhibitory effects: implications for social interactions, attention, and memory. *J. Exp. Psychol.* 133, 516–533. doi: 10.1037/0096-3445. 133.4.516
- Friesen, C. K., Ristic, J., and Kingstone, A. (2004). Attentional effects of counterpredictive gaze and arrow cues. *J. Exp. Psychol.* 30, 319–329. doi: 10.1037/0096-1523.30.2.319
- Friesen, C. K., & Kingstone, A. (1998). The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. *Psychonomic Bulletin & Review*, 5(3), 490–495. <https://doi.org/10.3758/BF03208827>
- Frith, C. D., & Frith, U. (2006a). How we predict what other people are going to do. *Brain Research*, 1079, 36–46. <https://doi.org/10.1016/j.brainres.2005.12.126>
- Frith, U., & Frith, C. D. (2003). Development and neurophysiology of mentalizing. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, 358(1431), 459–473. <https://doi.org/10.1098/rstb.2002.1218>
- Frith, C. D., & Frith, U. (1999). Interacting minds—A biological basis. *Science* (New York, N.Y.), 286(5445), 1692–1695.

- Fussell, S. R., Kiesler, S., Setlock, L. D., and Yew, V. (2008). “How people anthropomorphize robots,” in *Proceedings of the 3rd ACM/IEEE International Conference on Human Robot Interaction* (New York, NY: ACM), 145–152.
- Gallagher, H. L., & Frith, C. D. (2003). Functional imaging of ‘theory of mind.’ *Trends in Cognitive Sciences*, 7(2), 77–83.
- Gallagher, H. L., Happé, F., Brunswick, N., Fletcher, P. C., Frith, U., & Frith, C. D. (2000). Reading the mind in cartoons and stories: An fMRI study of “theory of mind” in verbal and nonverbal tasks. *Neuropsychologia*, 38(1), 11–21.
[https://doi.org/10.1016/S0028-3932\(99\)00053-6](https://doi.org/10.1016/S0028-3932(99)00053-6)
- Gallagher, H. L., Jack, A. I., Roepstroff, A., & Frith, C. D. (2002). Imaging the intentional stance in a competitive game. *NeuroImage*, 16, 814–821.
<https://doi.org/10.1006/nimg.2002.1117>
- Gallese, V., Keysers, C., & Rizzolatti, G. (2004). A unifying view of the basis of social cognition. *Trends in Cognitive Sciences*, 8(9), 396–403.
<https://doi.org/10.1016/j.tics.2004.07.002>
- Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain: A Journal of Neurology*, 119(Pt. 2), 593–609.
- Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological Science*, 21(12), 1845–1853. <https://doi.org/10.1177/0956797610388814>
- Ghosh, V. E., Moscovitch, M., Melo Colella, B., & Gilboa, A. (2014). Schema Representation in Patients with Ventromedial PFC Lesions. *Journal of*

Neuroscience, 34(36), 12057–12070. <https://doi.org/10.1523/JNEUROSCI.0740-14.2014>

Gilbert, D. T., Lieberman, M. D., Morewedge, C. K., and Wilson, T. D. (2004). The peculiar longevity of things not so bad. *Psychol. Sci.* 15, 14–19. doi: 10.1111/j.0963-7214.2004.01501003.x

Gobbini, M. I., Gentili, C., Ricciardi, E., Bellucci, C., Salvini, P., Laschi, C., ... Pietrini, P. (2011). Distinct neural systems involved in agency and animacy detection. *Journal of Cognitive Neuroscience*, 23(8), 1911–1920. <https://doi.org/10.1162/jocn.2010.21574>

Gobel, M. S., Tufft, M. R. A., & Richardson, D. C. (2017). Social beliefs and visual attention: How the social relevance of a cue influences spatial orienting. *Cognitive Science*, 42, 161–185. <https://doi.org/10.1111/cogs.12529>

Graham, J., & Haidt, J. (2010). Beyond beliefs: Religions bind individuals into moral communities. *Personality and Social Psychology Review: An Official Journal of the Society for Personality and Social Psychology*, 14(1), 140–150. <https://doi.org/10.1177/1088868309353415>

Graham, R., Friesen, C. K., Fichtenholtz, H. M., and LaBar, K. S. (2010). Modulation of reflexive orienting to gaze direction by facial expressions. *Vis. Cogn.* 18, 331–368. doi: 10.1080/13506280802689281

Gray, H. M., Gray, K., & Wegner, D. M. (2007). Dimensions of mind perception. *Science*, 315(5812), 619–619. <https://doi.org/10.1126/science.1134475>

- Gray, K., and Wegner, D. M. (2008). The sting of intentional pain. *Psychol. Sci.* 19, 1260–1262. doi: 10.1111/j.1467-9280.2008.02208.x
- Gray, K., Young, L., & Waytz, A. (2012). Mind perception is the essence of morality. *Psychological Inquiry*, 23(2), 101–124. <https://doi.org/10.1080/1047840X.2012.651387>
- Grèzes, J. (2004). Brain mechanisms for inferring deceit in the actions of others. *Journal of Neuroscience*, 24(24), 5500–5505. <https://doi.org/10.1523/JNEUROSCI.0219-04.2004>
- Grèzes, J., Berthoz, S., & Passingham, R. E. (2006). Amygdala activation when one is the target of deceit: Did he lie to you or to someone else? *NeuroImage*, 30(2), 601–608. <https://doi.org/10.1016/j.neuroimage.2005.09.038>
- Gutsell, J. N., & Inzlicht, M. (2012). Intergroup differences in the sharing of emotive states: Neural evidence of an empathy gap. *Social Cognitive and Affective Neuroscience*, 7(5), 596–603. <https://doi.org/10.1093/scan/nsr035>
- Hackel, L. M., Looser, C. E., & Van Bavel, J. J. (2014). Group membership alters the threshold for mind perception: The role of social identity, collective identification, and intergroup threat. *Journal of Experimental Social Psychology*, 52, 15–23. <https://doi.org/10.1016/j.jesp.2013.12.001>
- Haley, K. J., & Fessler, D. M. T. (2005). Nobody's watching?: Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior*, 26(3), 245–256. <https://doi.org/10.1016/j.evolhumbehav.2005.01.002>

- Harris, L. T., & Fiske, S. T. (2011). Perceiving humanity or not: A social neuroscience approach to dehumanized perception. In *Social Neuroscience: Toward Understanding the Underpinnings of the Social Mind*.
<https://doi.org/10.1093/acprof:oso/9780195316872.003.0008>
- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2), 243–259.
<https://doi.org/10.1017/CBO9781107415324.004>
- Hertz, N., & Wiese, E. (2017). Social facilitation with non-human agents: Possible or not? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 61(1), 222–225. <https://doi.org/10.1177/1541931213601539>
- Hietanen, J. K., and Leppänen, J. M. (2003). Does facial expression affect attention orienting by gaze direction cues? *J. Exp. Psychol.* 29:1228–1243. doi: 10.1037/0096-1523.29.6.1228
- Hogeveen, J., Obhi, S. S., Banissy, M. J., Santiesteban, I., Press, C., Catmur, C., & Bird, G. (2015). Task-dependent and distinct roles of the temporoparietal junction and inferior frontal cortex in the control of imitation. *Social Cognitive and Affective Neuroscience*, 10(7), 1003–1009. <https://doi.org/10.1093/scan/nsu148>
- Hood, B. M., Willen, J., and Driver, J. (1998). Adult's eyes trigger shifts of visual attention in human infants. *Psychol. Sci.* 9, 131–134. doi: 10.1111/1467-9280.00024

- Hori, E., Tazumi, T., Umeno, K., and Kamachi, M. (2005). Effects of facial expression on shared attention mechanisms. *Physiol. Behav.* 84, 397–405. doi: 10.1016/j.physbeh.2005.01.002
- Hornak, J., Bramham, J., Rolls, E. T., Morris, R. G., O'Doherty, J., Bullock, P. R., & Polkey, C. E. (2003). Changes in emotion after circumscribed surgical lesions of the orbitofrontal and cingulate cortices. *Brain: A Journal of Neurology*, 126(Pt. 7), 1691–1712. <https://doi.org/10.1093/brain/awg168>
- Huang, C., and Thomaz, A. L. (2011). “Effects of responding to, initiating and ensuring joint attention in human-robot interaction,” in *Proceedings of the 20th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN '11)*, Atlanta, GA, 65–71.
- Huey, E. D., Krueger, F., & Grafman, J. (2006). Representations in the human prefrontal cortex. *Current Directions in Psychological Science*, 15(4), 167–171.
- Hungr, C. J., & Hunt, A. R. (2012). Physical self-similarity enhances the gaze-cueing effect. *Quarterly Journal of Experimental Psychology*, 65(7), 1250–1259. <https://doi.org/10.1080/17470218.2012.690769>
- Hynes, C. A., Baird, A. A., & Grafton, S. T. (2006). Differential role of the orbital frontal lobe in emotional versus cognitive perspective-taking. *Neuropsychologia*, 44(3), 374–383. <https://doi.org/10.1016/j.neuropsychologia.2005.06.011>
- Iacoboni, M. (2005). Neural mechanisms of imitation. *Current Opinion in Neurobiology*, 15(6), 632–637. <https://doi.org/10.1016/j.conb.2005.10.010>

- Jack, A. I., & Robbins, P. (2012). The phenomenal stance revisited. *Review of Philosophy and Psychology*, 3(3), 383–403. <https://doi.org/10.1007/s13164-012-0104-5>
- Jacobson, L., Koslowsky, M., & Lavidor, M. (2012). TDCS polarity effects in motor and cognitive domains: A meta-analytical review. *Experimental Brain Research*, 216(1), 1–10. <https://doi.org/10.1007/s00221-011-2891-9>
- Javadi, A. H., Cheng, P., & Walsh, V. (2012). Short duration transcranial direct current stimulation (tDCS) modulates verbal memory. *Brain Stimulation*, 5(4), 468–474. <https://doi.org/10.1016/j.brs.2011.08.003>
- Jenkins, A. C., Macrae, C. N., & Mitchell, J. P. (2008). Repetition suppression of ventromedial prefrontal activity during judgments of self and others. *Proceedings of the National Academy of Sciences of the United States of America*, 105(11), 4507–4512. <https://doi.org/10.1073/pnas.0708785105>
- Jones, B. C., DeBruine, L. M., Main, J. C., Little, A. C., Welling, L. L. M., Feinberg, D. R., et al. (2010). Facial cues of dominance modulate the short-term gaze-cuing effect in human observers. *Proc. R. Soc. B* 277, 617–624. doi: 10.1098/rspb.2009.1575
- Kanda, T., Ishiguro, H., and Ishida, T. (2001). “Psychological analysis on human-robot interaction,” in *Proceedings of the 2001 IEEE International Conference on Robotics and Automation*, Seoul, 4166–4173.
- Kätsyri, J., Förger, K., Mäkräinen, M., and Takala, T. (2015). A review of empirical evidence on different uncanny valley hypotheses: support for perceptual

mismatch as one road to the valley of eeriness. *Front. Psychol.* 6:390. doi:
10.3389/fpsyg.2015.00390

Kawai, N. (2011). Attentional shift by eye gaze requires joint attention: Eye gaze cues are unique to shift attention1: Social attention by the gaze cues. *Japanese Psychological Research*, 53(3), 292–301. <https://doi.org/10.1111/j.1468-5884.2011.00470.x>

Keysers, C., & Perrett, D. I. (2004). Demystifying social cognition: A Hebbian perspective. *Trends in Cognitive Sciences*, 8(11), 501–507.
<https://doi.org/10.1016/j.tics.2004.09.005>

Kiesler, S., Powers, A., Fussell, S. R., & Torrey, C. (2008). Anthropomorphic interactions with a robot and robot-like agent. *Social Cognition*, 26(2), 169–181.

Kilner, J. M., Neal, A., Weiskopf, N., Friston, K. J., & Frith, C. D. (2009). Evidence of mirror neurons in human inferior frontal gyrus. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 29(32), 10153–10159.
<https://doi.org/10.1523/JNEUROSCI.2668-09.2009>

Konishi, S., & Kitagawa, G. (2008). Information criteria and statistical modeling. New York, NY: Springer Science & Business Media. Retrieved from
https://books.google.com/books?hl=en&lr=&id=3I9ZJusaYh0C&oi=fnd&pg=PA1&dq=Konishi+%26+Kitagawa,+2008&ots=Y_S1YyHleU&sig=6HHmYgfK5xgLLn0FrOqLohteqOA

- 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). <https://doi.org/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. <https://doi.org/10.1007/s00429-014-0803-z>
- Krall, S. C., Volz, L. J., Oberwelland, E., Grefkes, C., Fink, G. R., & Konrad, K. (2016). The right temporoparietal junction in attention and social interaction: A transcranial magnetic stimulation study. *Human Brain Mapping*, 37(2), 796–807.
- 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. <https://doi.org/10.3389/fnhum.2013.00323>
- Kupferberg, A., Huber, M., Helfer, B., Lenz, C., Knoll, A., & Glasauer, S. (2012). Moving just like you: Motor interference depends on similar motility of agent and observer. *PLOS ONE*, 7(6), e39637. <https://doi.org/10.1371/journal.pone.0039637>
- Langton, S. R. H., and Bruce, V. (1999). Reflexive visual orienting in response to the social attention of others. *Vis. Cogn.* 6, 541–567. doi: 10.1080/135062899394939

- Leslie, A. M., Friedman, O., & German, T. P. (2004). Core mechanisms in Btheory of mind.^ Trends in Cognitive Sciences, 8(12), 528–533.
<https://doi.org/10.1016/j.tics.2004.10.001>
- Liuzza, M. T., Cazzato, V., Vecchione, M., Crostella, F., Caprara, G. V., & Aglioti, S. M. (2011). Follow my eyes: The gaze of politicians reflexively captures the gaze of ingroup voters. *PLoS ONE*, 6(9), e25117.
<https://doi.org/10.1371/journal.pone.0025117>
- Looije, R., Neerincx, M. A., & Cnossen, F. (2010). Persuasive robotic assistant for health self-management of older adults: Design and evaluation of social behaviors. *International Journal of Human Computer Studies*, 68(6), 386–397.
<https://doi.org/10.1016/j.ijhcs.2009.08.007>
- Looser, C. E., & Wheatley, T. (2010). The tipping point of animacy: How, when, and where we perceive life in a face. *Psychological Science*, 21(12), 1854–1862.
<https://doi.org/10.1177/0956797610388044>
- Lundqvist, D., Flykt, A., and Öhman, A. (1998). *The Karolinska Directed Emotional Faces (KDEF)*. Stockholm: Karolinska Institute.
- MacDorman, K. F., and Ishiguro, H. (2006). The uncanny advantage of using androids in social and cognitive science research. *Interact. Stud.* 7, 361–368. doi: 10.1075/is.7.3.10
- Mackinnon, D., Krull, J., & Lockwood, C. (2000). Equivalence of the mediation, confounding, and suppression effect. *Prevention Science*, 1(4), 173–181.
<https://doi.org/10.1023/A1026595011371>

- Mandell, A. R., Smith, M. A., Martini, M. C., Shaw, T. H., and Wiese, E. (2015). “Does the presence of social agents improve cognitive performance on a vigilance task?,” in *Social Robotics. Lecture Notes in Computer Science*, eds A. Tapus, E. André, J. C. Martin, F. Ferland, and M. Ammi (Cham: Springer), 421–430.
- Mandell, A., Smith, M., and Wiese, E. (2017). “Mind Perception in humanoid agents has negative effects on cognitive processing,” in *Proceedings of Human Factors and Ergonomics Society*, Santa Monica, CA.
- Martini, M., Buzzell, G., and Wiese, E. (2015). “Agent appearance modulates mind attribution and social attention in human-robot interaction,” in *Social Robotics. Lecture Notes in Computer Science*, Vol. 9388, eds A. Tapus, E. André, J. C. Martin, F. Ferland, and M. Ammi (Cham: Springer), 431–439.
- Martini, M. C., Gonzalez, C. A., & Wiese, E. (2016). Seeing minds in others - Can agents with robotic appearance have human-like preferences? *PLoS ONE*, *11*(1), 1–23.
<https://doi.org/10.1371/journal.pone.0146310>
- Mathur, M. B., & Reichling, D. B. (2016). Navigating a social world with robot partners: A quantitative cartography of the Uncanny Valley. *Cognition*, *146*, 22–32.
<https://doi.org/10.1016/j.cognition.2015.09.008>
- Maurer, D., Grand, R. L., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*(6), 255–260.
[https://doi.org/10.1016/S1364-6613\(02\)01903-4](https://doi.org/10.1016/S1364-6613(02)01903-4)
- McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of*

the National Academy of Sciences, 98(20), 11832–11835.

<https://doi.org/10.1073/pnas.211415698>

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

Mitchell, J. P., Macrae, C. N., & Banaji, M. R. (2004). Encoding-Specific Effects of Social Cognition on the Neural Correlates of Subsequent Memory. *Journal of Neuroscience*, 24(21), 4912–4917. <https://doi.org/10.1523/JNEUROSCI.0481-04.2004>

Mitchell, Jason P., Banaji, M. R., & Macrae, C. N. (2005). General and specific contributions of the medial prefrontal cortex to knowledge about mental states. *NeuroImage*, 28(4), 757–762. <https://doi.org/10.1016/j.neuroimage.2005.03.011>

Mitchell, Jason P., Macrae, C. N., & Banaji, M. R. (2006). Dissociable Medial Prefrontal Contributions to Judgments of Similar and Dissimilar Others. *Neuron*, 50(4), 655–663. <https://doi.org/10.1016/j.neuron.2006.03.040>

Moliadze, V., Antal, A., & Paulus, W. (2010). Electrode-distance dependent after-effects of transcranial direct and random noise stimulation with extracephalic reference electrodes. *Clinical Neurophysiology*, 121(12), 2165–2171. <https://doi.org/10.1016/j.clinph.2010.04.033>

Morewedge, C. K. (2009). Negativity bias in attribution of external agency. *J. Exp. Psychol. Gen.* 138, 535–545. doi: 10.1037/a0016796

- Mori, M. (1970). The uncanny valley: The original essay by masahiro mori. Retrieved from Energy website: <http://spectrum.ieee.org/automaton/robotics/humanoids/the-uncanny-valley>
- Mori, M., MacDorman, K., & Kageki, N. (2012). The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine*, 19(2), 98–100.
<https://doi.org/10.1109/MRA.2012.2192811>
- Mubin, O., Stevens, C. J., Shahid, S., Mahmud, A. A., & Dong, J.-J. (2013). A review of the applicability of robots in education. *Technology for Education and Learning*, 1(1). <https://doi.org/10.2316/Journal.209.2013.1.209-0015>
- Mukamel, R., Ekstrom, A. D., Kaplan, J., Iacoboni, M., & Fried, I. (2010). Single-neuron responses in humans during execution and observation of actions. *Current Biology: CB*, 20(8), 750–756. <https://doi.org/10.1016/j.cub.2010.02.045>
- Mutlu, B., Forlizzi, J., and Hodgins, J. (2006). “A storytelling robot: Modeling and evaluation of human-like gaze behavior,” in *Proceedings of the 2006 6th IEEE-RAS International Conference on Humanoid Robots* (Genova: IEEE), 1–6.
- Mutlu, B., Kanda, T., Forlizzi, J., and Ishiguro, H. (2012). Conversational gaze mechanisms for human-like robots. *ACM Trans. Interact. Intell. Syst.* 1, 12. doi: 10.1145/2070719.2070725
- Mutlu, B., Yamaoka, F., Kanda, T., Ishiguro, H., and Hagita, N. (2009). “Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior,” in *Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction*, La Jolla, CA, 69–76. doi: 10.1145/1514095.1514110

- Nitsche, M. A., Cohen, L. G., Wassermann, E. M., Priori, A., Lang, N., Antal, A., ... Pascual-Leone, A. (2008). Transcranial direct current stimulation: State of the art 2008. *Brain Stimulation*, 1(3), 206–223. <https://doi.org/10.1016/j.brs.2008.06.004>
- Nomura, T., Kanda, T., & Suzuki, T. (2006). Experimental investigation into influence of negative attitudes toward robots on human-robot interaction. *AI and Society*, 20(2), 138–150. <https://doi.org/10.1007/s00146-005-0012-7>
- Nummenmaa, L., & Calder, A. J. (2009). Neural mechanisms of social attention. *Trends in Cognitive Sciences*, 13(3), 135–143. <https://doi.org/10.1016/j.tics.2008.12.006>
- Oberman, L. M., McCleery, J. P., Ramachandran, V. S., & Pineda, J. A. (2007). EEG evidence for mirror neuron activity during the observation of human and robot actions: Toward an analysis of the human qualities of interactive robots. *Neurocomputing*, 70(13), 2194–2203. <https://doi.org/10.1016/j.neucom.2006.02.024>
- Ohlsen, G., van Zoest, W., and van Vugt, M. (2013). Gender and facial dominance in gaze cuing: emotional context matters in the eyes that we follow. *PLoS ONE* 8:4. doi: 10.1371/journal.pone.0059471
- Ohnishi, T., Moriguchi, Y., Matsuda, H., Mori, T., Hirakata, M., Imabayashi, E., ... Uno, A. (2004). The neural network for the mirror system and mentalizing in normally developed children: An fMRI study. *NeuroReport*, 15(9), 1483–1487. <https://doi.org/10.1097/01.wnr.0000127464.17770.1f>

- Ohtsubo, Y. (2007). Perceived intentionality intensifies blameworthiness of negative behaviors: blame-praise asymmetry in intensification effect 1. *Jpn. Psychol. Res.* 49, 100–110. doi: 10.1111/j.1468-5884.2007. 00337.x
- Oostenveld, R., & Praamstra, P. (2001). The five percent electrode system for high-resolution EEG and ERP measurements. *Clinical Neurophysiology*, 7.
- Özdem, C., Wiese, E., Wykowska, A., Müller, H., Brass, M., & Van Overwalle, F. (2016). 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. <https://doi.org/10.1080/17470919.2016.1207702>
- Oztop, E., Franklin, D. W., Chaminade, T., & Cheng, G. (2005). Human– humanoid interaction: Is a humanoid robot perceived as a human? *International Journal of Humanoid Robotics*, 2(4), 537–559. <https://doi.org/10.1142/S0219843605000582>
- Pak, R., Fink, N., Price, M., Bass, B., and Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics* 55, 1059–1072. doi: 10.1080/00140139. 2012.691554
- Park, Y. G. (2013). Comments on Statistical Issues in January 2013. *Korean Journal of Family Medicine*, 34(1), 64. <https://doi.org/10.4082/kjfm.2013.34.1.64>
- Parkin, B. L., Ekhtiari, H., & Walsh, V. F. (2015). Non-invasive Human Brain Stimulation in Cognitive Neuroscience: A Primer. *Neuron*, 87(5), 932–945. <https://doi.org/10.1016/j.neuron.2015.07.032>

- Pavan, G., Dalmasso, M., Galfano, G., & Castelli, L. (2011). Racial group membership is associated to gaze-mediated orienting in Italy. *PLoS ONE*, 6(10).
<https://doi.org/10.1371/journal.pone.0025608>
- Perner, J., Aichhorn, M., Kronbichler, M., Staffen, W., & Ladurner, G. (2006). Thinking of mental and other representations: The roles of left and right temporo-parietal junction. *Social Neuroscience*, 1(3–4), 245–258.
<https://doi.org/10.1080/17470910600989896>
- Pfeiffer, U. J., Timmermans, B., Bente, G., Vogeley, K., and Schilbach, L. (2011). A non-verbal turing test: differentiating mind from machine in gaze-based social interaction. *PLoS ONE* 6:11. doi: 10.1371/journal.pone.002 7591
- Pfeiffer, U. J., Schilbach, L., Timmermans, B., Kuzmanovic, B., Georgescu, A. L., Bente, G., & Vogeley, K. (2014). Why we interact: On the functional role of the striatum in the subjective experience of social interaction. *NeuroImage*, 101, 124–137.
<https://doi.org/10.1016/j.neuroimage.2014.06.061>
- Pfeiffer-Lessmann, N., Pfeiffer, T., and Wachsmuth, I. (2012). “An operational model of joint attention-Timing of gaze patterns in interactions between humans and a virtual human,” in *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, eds N. Miyake, D. Peebles, and R. P. Cooper (Austin, TX: Cognitive Science Society).
- Pobric, G., & de Hamilton, A. F. C. (2006). Action understanding requires the left inferior frontal cortex. *Current Biology: CB*, 16(5), 524–529.
<https://doi.org/10.1016/j.cub.2006.01.033>

- Polanía, R., Nitsche, M. A., & Ruff, C. C. (2018). Studying and modifying brain function with non-invasive brain stimulation. *Nature Neuroscience*, 21(2), 174–187.
<https://doi.org/10.1038/s41593-017-0054-4>
- Porciello, G., Holmes, B. S., Liuzza, M. T., Crostella, F., Aglioti, S. M., & Bufalari, I. (2014). Interpersonal multisensory stimulation reduces the overwhelming distracting power of self-gaze: psychophysical evidence for ‘engazement.’ *Scientific Reports*, 4(6669), 1–7. <https://doi.org/10.1038/srep06669>
- Press, C., Bird, G., Flach, R., & Heyes, C. (2005). Robotic movement elicits automatic imitation. *Brain Research. Cognitive Brain Research*, 25(3), 632–640.
<https://doi.org/10.1016/j.cogbrainres.2005.08.020>
- Press, C., Gillmeister, H., & Heyes, C. (2007). Sensorimotor experience enhances automatic imitation of robotic action. *Proceedings of the Royal Society B: Biological Sciences*, 274(1625), 2509–2514.
<https://doi.org/10.1098/rspb.2007.0774>
- Quadflieg, S., Mason, M. F., & Macrae, C. N. (2004). The owl and the pussycat: Gaze cues and visuospatial orienting. *Psychonomic Bulletin & Review*, 11(5), 826–831.
<https://doi.org/10.3758/BF03196708>
- Rani Das, K., & Imon, Rahmatullah. (2016). A Brief Review of Tests for Normality. *American Journal of Theoretical and Applied Statistics*, 5(1), 5.
<https://doi.org/10.11648/j.ajtas.20160501.12>
- Riether, N., Hegel, F., Wrede, B., & Horstmann, G. (2012). Social facilitation with social robots? *Proceedings of the Seventh Annual ACM/IEEE International Conference*

on Human-Robot Interaction - HRI '12, 41.

<https://doi.org/10.1145/2157689.2157697>

Ristic, J., & Kingstone, A. (2005). Taking control of reflexive social attention. *Cognition*, 94(3), B55–B65. <https://doi.org/10.1016/j.cognition.2004.04.005>

Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169–192. <https://doi.org/10.1146/annurev.neuro.27.070203.144230>

Rosenthal-Von Der Pütten, A. M., & Krämer, N. C. (2014). How design characteristics of robots determine evaluation and uncanny valley related responses. *Computers in Human Behavior*, 36, 422–439. <https://doi.org/10.1016/j.chb.2014.03.066>

Ruby, P., & Decety, J. (2001). Effect of subjective perspective taking during simulation of action: a PET investigation of agency. *Nature Neuroscience*, 4(5), 546–550. <https://doi.org/10.1038/87510>

Samson, D., Apperly, I. A., Chiavarino, C., & Humphreys, G. W. (2004). Left temporoparietal junction is necessary for representing someone else's belief. *Nature Neuroscience*, 7(5), 499–500.

Sanfey, A. G., Rilling, J. K., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2003). The neural basis of economic decision-making in the Ultimatum Game. *Science*, 300(5626), 1755–1758. <https://doi.org/10.1126/science.1082976>

Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2012). Enhancing Social Ability by Stimulating Right Temporoparietal Junction. *Current Biology*, 22(23), 2274–2277. <https://doi.org/10.1016/j.cub.2012.10.018>

- Santiesteban, I., Banissy, M. J., Catmur, C., & Bird, G. (2015). Functional lateralization of temporoparietal junction - imitation inhibition, visual perspective-taking and theory of mind. *European Journal of Neuroscience*, 42(8), 2527–2533.
<https://doi.org/10.1111/ejn.13036>
- Saxe, R., Carey, S., & Kanwisher, N. (2004). Understanding other minds: linking developmental psychology and functional neuroimaging. *Annual Review of Psychology*, 55, 87–124.
<https://doi.org/10.1146/annurev.psych.55.090902.142044>
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in “theory of mind.” *NeuroImage*, 1835–1842.
- Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*, 16(2), 235–239. <https://doi.org/10.1016/j.conb.2006.03.001>
- Saxe, Rebecca, & 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.
<https://doi.org/10.1111/j.1467-9280.2006.01768.x>
- Saxe, R., & Wexler, A. (2005). Making sense of another mind: The role of the right temporo-parietal junction. *Neuropsychologia*, 43(10), 1391–1399.
<https://doi.org/10.1016/j.neuropsychologia.2005.02.013>
- Saygin, A. P. (2007). Superior temporal and premotor brain areas necessary for biological motion perception. *Brain: A Journal of Neurology*, 130(Pt. 9), 2452–2461. <https://doi.org/10.1093/brain/awm162>

- Saygin, A. P., Chaminade, T., Ishiguro, H., Driver, J., & Frith, C. (2012). The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *SCAN*, 7, 413–422. <https://doi.org/10.1093/scan/nsr025>
- Saygin, A. P., Wilson, S. M., Hagler, D. J., Bates, E., & Sereno, M. I. (2004). Point-light biological motion perception activates human premotor cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, 24(27), 6181–6188. <https://doi.org/10.1523/JNEUROSCI.0504-04.2004>
- Scassellati, B., Admoni, H., & Matarić, M. (2012). Robots for Use in Autism Research. *Annual Review of Biomedical Engineering*, 14, 275–294. <https://doi.org/10.1146/annurev-bioeng-071811-150036>
- Schein, C., & Gray, K. (2015). The unifying moral dyad: Liberals and conservatives share the same harm-based moral template. *Personality and Social Psychology Bulletin*, 41(8), 1147–1163. <https://doi.org/10.1177/0146167215591501>
- Scholz, J., Triantafyllou, C., Whitfield-Gabrieli, S., Brown, E. N., & Saxe, R. (2009). Distinct regions of right temporo-parietal junction are selective for theory of mind and exogenous attention. *PLOS ONE*, 4(3), e4869.
- Schurz, M., Radua, J., Aichhorn, M., Richlan, F., & Perner, J. (2014). Fractionating theory of mind: A meta-analysis of functional brain imaging studies. *Neuroscience and Biobehavioral Reviews*, 42, 9–34. <https://doi.org/10.1016/j.neubiorev.2014.01.009>

- Schweinberger, S. R., Kloth, N., & Jenkins, R. (2007). Are you looking at me? Neural correlates of gaze adaptation. *Neuroreport*, 18(7), 693–696.
<https://doi.org/10.1097/WNR.0b013e3280c1e2d2>
- Scopelliti, M., Giuliani, M. V., & Fornara, F. (2005). Robots in a domestic setting: A psychological approach. *Universal Access in the Information Society*, 4(2), 146–155. <https://doi.org/10.1007/s10209-005-0118-1>
- Shariff, A. F., & Norenzayan, A. (2007). God is watching you: Priming God concepts increases prosocial behavior in an anonymous economic game. *Psychological Science*, 18(9), 803–809. <https://doi.org/10.1111/j.1467-9280.2007.01983.x>
- Shamay-Tsoory, S. G., Aharon-Peretz, J., & Perry, D. (2009). Two systems for empathy: A double dissociation between emotional and cognitive empathy in inferior frontal gyrus versus ventromedial pre-frontal lesions. *Brain: A Journal of Neurology*, 132(Pt. 3), 617–627. <https://doi.org/10.1093/brain/awn279>
- Shepherd, S. V., Deaner, R. O., & Platt, M. L. (2006). Social status gates social attention in monkeys. *Current Biology*, 16(R), 119–120.
- Short, E., Hart, J., Vu, M., & Scassellati, B. (2010). No fair!! An interaction with a cheating robot. *2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, 219–226. <https://doi.org/10.1109/HRI.2010.5453193>
- Sidner, C. L., Kidd, C. D., Lee, C., and Lesh, N. (2004). “Where to look: a study of human-robot engagement,” in *Proceedings of the 9th International Conference on Intelligent User Interfaces*, (New York, NY: ACM Press), 78–84.

- Simmons, W. K., Reddish, M., Bellgowan, P. S. F., & Martin, A. (2010). The selectivity and functional connectivity of the anterior temporal lobes. *Cerebral Cortex*, 20(4), 813–825. <https://doi.org/10.1093/cercor/bhp149>
- Spunt, R. P., & Lieberman, M. D. (2012). Dissociating modality-specific and supramodal neural systems for action understanding. *Journal of Neuroscience*, 32(10), 3575–3583. <https://doi.org/10.1523/JNEUROSCI.5715-11.2012>
- Spunt, Robert P., & Adolphs, R. (2014). Validating the Why/How contrast for functional MRI studies of Theory of Mind. *NeuroImage*, 99, 301–311. <https://doi.org/10.1016/j.neuroimage.2014.05.023>
- Spunt, Robert P., Meyer, M. L., & Lieberman, M. D. (2015). The default mode of human brain function primes the intentional stance. *Journal of Cognitive Neuroscience*, 27(6), 1116–1124. https://doi.org/10.1162/jocn_a_00785
- Staudte, M., and Crocker, M. W. (2011). Investigating joint attention mechanisms through spoken human-robot interaction. *Cognition* 120, 268–291. doi: 10.1016/j.cognition.2011.05.005
- Stone, V. E., Baron-Cohen, S., & Knight, R. T. (1998). Frontal lobe contributions to theory of mind. *Journal of Cognitive Neuroscience*, 10(5), 640–656.
- Süßenbach, F., & Schönbrodt, F. (2014). Not afraid to trust you: Trustworthiness moderates gaze cueing but not in highly anxious participants. *Journal of Cognitive Psychology*, 26(September 2015), 1–9. <https://doi.org/10.1080/20445911.2014.945457>

- Szczepanski, S. M., & Knight, R. T. (2014). Insights into Human Behavior from Lesions to the Prefrontal Cortex. *Neuron*, 83(5), 1002–1018.
<https://doi.org/10.1016/j.neuron.2014.08.011>
- Takahashi, H., Terada, K., Morita, T., Suzuki, S., Haji, T., Kozima, H., ... Naito, E. (2014). Different impressions of other agents obtained through social interaction uniquely modulate dorsal and ventral pathway activities in the social human brain. *Cortex*, 58, 289–300. <https://doi.org/10.1016/j.cortex.2014.03.011>
- Takahashi, K., & Watanabe, K. (2013). Gaze cueing by pareidolia faces. *I-Perception*, 4(8), 490–492. <https://doi.org/10.1068/i0617sas>
- Takao, S., & Ariga, A. (2016). General Trust is Correlated with Attentional Orientation Triggered by Gaze Direction. *8th International Conference on Knowledge and Smart Technology*, 287–290. IEEE.
- Tapus, A., & Matarić, M. (2006). Towards socially assistive robots. *Journal of Robotics Society of Japan*, 14(5), 576–578. <https://doi.org/10.7210/jrsj.24.576>
- Teufel, C., Alexis, D. M., Todd, H., Lawrance-Owen, A. J., Clayton, N. S., and Davis, G. (2009). Social cognition modulates the sensory coding of observed gaze direction. *Curr. Biol.* 19, 1274–1277. doi: 10.1016/j.cub.2009. 05.069
- Tipples, J. (2006). Fear and fearfulness potentiate automatic orienting to eye gaze. *Cognition and Emotion*, 20(2), 309–320.
<https://doi.org/10.1080/02699930500405550>
- Tsuchida, A., & Fellows, L. K. (2012). Are you upset? Distinct roles for orbitofrontal and lateral prefrontal cortex in detecting and distinguishing facial expressions of

- emotion. *Cerebral Cortex* (New York, N.Y.: 1991), 22(12), 2904–2912.
<https://doi.org/10.1093/cercor/bhr370>
- Tung, F. (2011). “Influence of gender and age on the attitudes of children towards humanoid robots,” in *Proceedings of the 14th International Conference on Human-Computer Interaction: Users and Applications*, Orlando, FL, 637–646.
doi: 10.1007/978-3-642-21619-0_76
- Umiltà, M. A., Kohler, E., Gallese, V., Fogassi, L., Fadiga, L., Keysers, C., & Rizzolatti, G. (2001). I know what you are doing. a neuro- physiological study. *Neuron*, 31(1), 155–165.
- Van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, 35(4), 211–219. <https://doi.org/10.1016/j.tins.2012.02.001>
- Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, 30(3), 829–858. <https://doi.org/10.1002/hbm.20547>
- Van Overwalle, F., & Baetens, K. (2009). Understanding others’ actions and goals by mirror and mentalizing systems: A meta-analysis. *NeuroImage*, 48(3), 564–584.
<https://doi.org/10.1016/j.neuroimage.2009.06.009>
- Vecera, S. P., and Rizzo, M. (2006). Eye gaze does not produce reflexive shifts of attention: evidence from frontal-lobe damage. *Neuropsychologia* 44, 150–159.
doi: 10.1016/j.neuropsychologia.2005.04.010
- Völlm, B. A., Taylor, A. N. W., Richardson, P., Corcoran, R., Stirling, J., McKie, S., . . . Elliott, R. (2006). Neuronal correlates of theory of mind and empathy: A

- functional magnetic resonance imaging study in a nonverbal task. *NeuroImage*, 29(1), 90–98. <https://doi.org/10.1016/j.neuroimage.2005.07.022>
- Wagner, D. D., Kelley, W. M., & Heatherton, T. F. (2011). Individual differences in the spontaneous recruitment of brain regions supporting mental state understanding when viewing natural social scenes. *Cerebral Cortex*, 21(12), 2788–2796. <https://doi.org/10.1093/cercor/bhr074>
- Wang, Y., & Quadflieg, S. (2015). In our own image? Emotional and neural processing differences when observing human-human vs human-robot interactions. *Social Cognitive and Affective Neuroscience*, 10(11), 1515–1524. <https://doi.org/10.1093/scan/nsv043>
- Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, 14(8), 383–388. <https://doi.org/10.1016/j.tics.2010.05.006>
- Waytz, A., Heafner, J., and Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *J. Exp. Soc. Psychol.* 52, 113–117. doi: 10.1016/j.jesp.2014.01.005
- Waytz, A., Morewedge, C. K., Epley, N., Monteleone, G., Gao, J.-H., and Cacioppo, J. T. (2010). Making sense by making sentient: effectance motivation increases anthropomorphism. *J. Pers. Soc. Psychol.* 99, 410–435. doi: 10.1037/a0020240
- Wheatley, T., Weinberg, A., Looser, C., Moran, T., & Hajcak, G. (2011). Mind perception: Real but not artificial faces sustain neural activity beyond the

N170/VPP. *PLoS ONE*, 6(3), e17960.

<https://doi.org/10.1371/journal.pone.0017960>

Weis, P., and Wiese, E. (2017). “Cognitive conflict as possible origin of the uncanny valley,” in *Proceedings of Human Factors and Ergonomics Society*, Santa Monica, CA.

Wiese, E., Buzzell, G. A., Abubshait, A., & Beatty, P. J. (2018). Seeing minds in others: Mind perception modulates low-level social-cognitive performance and relates to ventromedial prefrontal structures. *Cognitive, Affective, & Behavioral Neuroscience*, 18(5), 837–856. <https://doi.org/10.3758/s13415-018-0608-2>

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots as intentional agents: Using neuroscientific methods to make robots appear more social. *Frontiers in Psychology*, 8, 1663. <https://doi.org/10.3389/fpsyg.2017.01663>

Wiese, E., Müller, H. J., & Wykowska, A. (2014). Using a gaze-cueing paradigm to examine social cognitive mechanisms of individuals with autism observing robot and human faces. *International Conference on Social Robotics*, 8755, 370–379. https://doi.org/10.1007/978-3-319-11973-1_38

Wiese, E., Wykowska, A., & Müller, H. J. (2014). What we observe is biased by what other people tell us: Beliefs about the reliability of gaze behavior modulate attentional orienting to gaze cues. *PLoS ONE*, 9(4), e94529. <https://doi.org/10.1371/journal.pone.0094529>

- 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. <https://doi.org/10.1371/journal.pone.0045391>
- Wiese, E., Zwickel, J., & Müller, H. J. (2013). The importance of context information for the spatial specificity of gaze cueing. *Attention, Perception, & Psychophysics*, 75(5), 967–982. <https://doi.org/10.3758/s13414-013-0444-y>
- Wood, J. N., & Grafman, J. (2003). Human prefrontal cortex: processing and representational perspectives. *Nature Reviews Neuroscience*, 4(2), 139–147.
- Woods, S., Dautenhahn, K., and Kaouri, C. (2005). “Is someone watching me? Consideration of social facilitation effects in human-robot interaction experiments,” in *Proceedings of 2015 IEEE International Symposium on Computational Intelligence in Robotics and Automation*, Espoo, 53–60.
- 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. <https://doi.org/10.1371/journal.pone.0094339>
- Wykowska, A., Kajopoulos, J., Obando-Leitón, M., Chauhan, S. S., Cabibihan, J. J., and Cheng, G. (2017). Humans are well tuned to detecting agents among non-agents: examining the sensitivity of human perception to behavioral characteristics of intentional systems. *Int. J. Soc. Robot.* 7, 767–781. doi: 10.1007/ s12369- 015-0299- 6

- Yamazaki, A., Yamazaki, K., Burdelski, M., and Kuno, Y. (2010). Coordination of verbal and non-verbal actions in human-robot interaction at museums and exhibitions. *J. Pragmat.* 42, 2398–2414. doi: 10.1016/j.pragma.2009. 12.023
- Zink, C. F., Kempf, L., Hakimi, S., Rainey, C. A., Stein, J. L., & Meyer-Lindenberg, A. (2011). Vasopressin modulates social recognition-related activity in the left temporoparietal junction in humans. *Translational Psychiatry*, 1(4), e3–e3. <https://doi.org/10.1038/tp.2011.2>

BIOGRAPHY

Abdulaziz Abubshait received his Bachelor of Arts in Psychology from George Mason University with Honors in 2015. He received his Master of Arts in Psychology with a concentration in Human Factors and Applied Cognition from George Mason University in 2017.