The Influence of Intentional Stance on the Neural Encoding of Joint Attention

Kirilee Wagner (Master of Research)

Department of Cognitive Science

Faculty of Human Sciences

Macquarie University, Sydney, Australia

Submission Date: 11.10.19

Table of Contents

Abstract

Adopting an intentional stance towards a social partner is a crucial component of evaluating the success of joint attention, where social interlocuters must represent the mind and perspectives of others. Recent work has established that centroparietal P250 and P350 ERPs are sensitive to whether the gaze shifts of others signal the achievement or avoidance of joint attention, and that this modulation depends on the adoption of an intentional stance (Caruana & McArthur, 2019). The current study attempted to replicate these effects, determining their reliability across testing contexts, and examining the influence of the aesthetic anthropomorphism of the stimuli used. Participants initiated gaze-cued joint attention bids with an on-screen virtual partner, which shifted its eye gaze congruently to respond to joint attention bids on 50% of trials and responded incongruently to avoid joint attention on the remaining trials. Participants were told that in one block their partner was controlled by a human, and in another by a computer program. The aesthetic anthropomorphism of the faces was manipulated between-subjects so that one group interacted with an animated human face (n=21), and the other interacted with a humanoid robot face (n=19). Larger P250 mean amplitudes were measured in response to congruent gaze shifts compared to incongruent, and the opposite pattern was observed for P350 responses. However, these ERPs were not reliably modulated by the adoption of an explicit intentional stance across both stimulus groups. We found tentative evidence to suggest that this unreliability may be explained by individual differences in anthropomorphism tendencies.

Statement of Originality

This work has not previously been submitted for a degree or diploma in any university. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made in the thesis itself.

Date: 11.10.19

Acknowledgements

Firstly, I'd like to thank my thesis supervisor, Dr. Nathan Caruana of the Department of Cognitive Science at Macquarie University. He has been enormously supportive and encouraging throughout this entire year. I would also like to thank Professor Genevieve McArthur, also of the Department of Cognitive Science, as co-supervisor and second reader for this thesis. I am very thankful to her for her valuable comments.

Many thanks also to Dr. Nicholas Badcock (Department of Cognitive Science), for his support and patience in teaching me to use MATLAB. His door was always open whenever I needed help, for which I am extremely grateful.

Finally, a huge thanks to my parents and my two brothers, for their unfailing support and encouragement throughout this year.

Introduction

Joint Attention

Joint attention is the social ability to coordinate attention between two people and an object of interest, so that both individuals are knowingly attending to the same thing (Bruner, 1974; Mundy, 2018). This can be achieved using verbal and non-verbal social cues, such as eye gaze and finger pointing (Yu & Smith, 2013; 2017a; 2017b). In a typical joint attention episode, one individual will initiate joint attention by shifting their gaze or pointing to the object or event of interest. A joint attention response occurs if the second individual recognises the bid as intentional and responds by shifting their attention to the same location (Bruner, 1974). The success of a joint attention episode is then evaluated, typically by the initiator who must determine if their social interlocutor has responded to their bid (Caruana, de Lissa & McArthur, 2015). This final stage is critical, as it enables the initiator to determine whether they need to engage in further attempts to guide their partner's attention.

Gaze shifts are a particularly important cue for joint attention because the eyes are the only sensory organ with the dual function of sensing and signaling visual information (Gobel, Kim & Richardson, 2015). As such, the eye gaze of others can be a rich and constant source of information about their attentional focus and perspective (Cañigueral & Hamilton, 2019). Gaze-cued joint attention relies on a range of social cognitive abilities. For example, one must be able to process eye movements as a biologically-relevant cue (Carlin & Calder, 2013), to then recognise this cue as a representation of another's shifting attention, and to attribute communicative intent to that gaze shift (Baron-Cohen, 1997; Kaplan & Hafner, 2006; Tomasello, 1995). Therefore, joint attention requires interacting individuals to adopt an 'intentional stance' towards each other. This refers to the idea that one represents another entity as a mindful and sentient agent with different perspectives, goals, desires and intentions (Dennett, 1989). By adopting this stance, an individual is ready to represent the

mind and perspectives of their social partner, and to make relevant predictions or evaluations of their behaviour. This is particularly crucial for evaluating the success of joint attention during genuine social interactions.

Understanding the neural mechanisms of joint attention has become a focus of empirical investigation because it is a core social skill used in everyday interactions, and because it is central to socio-cognitive development and language acquisition (Tomasello, 1995). Delays in joint attention development are also associated with subsequent delays in social cognition (e.g. "mentalising" – making inferences about other minds) and communication (Charman et al. 2003; Dawson et al. 2004). Further, delayed joint attention development is a characteristic behavioural marker of autism (Adamson, Bakeman, Suma & Robins, 2019; Bruinsma, Koegel & Koegel, 2004; Mundy, 2018; Pelphrey, Morris & McCarthy, 2005). However, studying the neural mechanisms of joint attention is difficult because this demands experimental paradigms of joint attention behaviour which can simultaneously simulate ecologically valid social interactions whilst maintaining experimental control and objectivity in the measurement of corresponding brain activity (Caruana, McArthur, Brock & Woolgar, 2017; Schilbach et al. 2013). Specifically, in order to be ecologically-valid, the paradigm must enable a reciprocal, and seemingly genuine, social interaction, such that the participant adopts an intentional stance towards their partner. Furthermore, the paradigm must also control all aspects of a social partner's behaviour, such as facial expression, or head and body movements. Given these competing demands, investigating the neural mechanisms of joint attention – like most aspects of interactive cognition – has challenged the field of social neuroscience.

Pre-recorded stimuli. Balancing ecological validity and experimental control for investigating joint attention is particularly difficult within the constraints of neuroimaging research environments, such as functional magnetic resonance imaging (fMRI). Early fMRI

studies probing the neural correlates of joint attention attempted to simulate social scenarios by showing participants a video of an actor. For example, to investigate the final, evaluative stage of joint attention, Williams, Waiter, Perra, Perret and Whiten (2005) developed a paradigm in which participants saw a video of an actor who focussed on a moving red dot at the bottom of the screen. Participants were instructed to focus on the red dot in two conditions. In a "joint attention" condition, the actor's gaze congruently followed the motion of the dot. In the "no joint attention" condition, the location of the dot was manipulated so that the actor appeared to always avoid looking at the dot. This meant that the participant's gaze coincided with the actor's gaze during the joint attention condition but was misaligned to the actor's gaze during the no joint attention condition. The authors reported increased activation in the right ventromedial and left anterior frontal cortices during 'joint attention' compared to 'no joint attention'. Given that these regions have also been implicated in mentalising tasks, the authors argued that these findings highlight the role of mental state attribution during joint attention (Frith & Frith, 2000). However, the paradigm in this study is limited in its ecological validity since it lacks the reciprocity and interactivity inherent to social situations. The paradigm is not interactive because the actor did not respond to participants' eye gaze. Further, given that the paradigm was pre-recorded, the actor was not a 'present social partner' so participants did not need to adopt an intentional stance towards him or to think about his mental states ("mentalise"). As such, it is possible that the observed effects reflect self-relevance processing, rather than the representation of 'other' minds, given that the medial prefrontal cortex has been well-established to be implicated during self-reflection and representation (e.g., Heatherton, McCrae & Kelley, 2004; Heatherton et al., 2006). It is therefore uncertain whether these effects reflect evaluations about whether another mind has aligned with our own focus of attention (i.e., joint attention) or whether it

reflects a more domain-general evaluation about whether a stimulus or event is congruent with our own current perspective.

Gordon, Eilbott, Feldman, Pelphrey and Vander Wyk (2013) partially addressed the problem of interactivity by using eye-tracking technology to make their social stimulus gaze contingent and thus increasing the subjective sense of reciprocity between the participant and the observed agent. Subjects were asked to interact with a pre-recorded video showing an actor's head and neck. Images of identical human silhouettes were placed to the left and right of the video frame. Participants made eye contact with the actor and shifted their gaze to one of the target locations. During the first two blocks of testing, the video was programmed to contingently follow the participants' eye movements to simulate responsive joint attention. During the following blocks, however, the actor responded congruently only 50% of the time, simulating successful and unsuccessful joint attention bids. The authors reported that observing congruent responses compared to incongruent responses resulted in increased activation in the anterior cingulate cortex, right fusiform gyrus, amygdala, striatum and parahippocampal regions. Incongruent responses, however, were associated with greater activation in the right temporoparietal junction compared to congruent responses made by the actor. The regions activated after congruent responses align with regions previously implicated in social reward processing. As such, the authors argue that these mechanisms may reflect the hedonic reinforcing process of evaluating that a social partner has aligned with our own focus of attention. However, as with Williams et al. (2005), the strength of this suggestion is obscured by the fact that pre-recorded stimuli do not simulate social interactions, which are characterised by sentient agents who actively attend to the same thing at the same time. Participants were aware that the stimulus was pre-recorded, so there was no need for them to adopt an intentional stance by representing a social partner's perspective. Therefore, it is uncertain to what extent the neural mechanisms

identified in this study reflect the true evaluation of joint attention, since an explicit intentional stance was not adopted.

Live interactive paradigms. To increase reciprocity and to ensure participants adopted an intentional stance towards their social partner, Lachat, Hugueville, Lemaréchal, Conty & George (2012) designed an elegant, live, dyadic interaction paradigm which still offered a certain level of experimental control. Participants were asked to interact through a circular window. Each participant could see a set of four lights which surrounded the window. In a "social" condition, one participant was instructed to choose and direct the other's attention to one of the LED lights, while the other was instructed either to follow this gaze cue (joint attention) or to look at an alternative LED light (no joint attention). In a "non-social" condition, participants were told to attend to a particular coloured light. These light cues were used to either create or avoid instances of incidental joint attention. Participants were either both instructed to look at a single LED so that their gaze coincided, or they were instructed separately to look at different LEDs, so that they avoided each other's gaze. In doing this, this study attempted to tease apart the neural processing associated with incidental and deliberate joint attention. This study found that irrespective of whether joint attention is achieved deliberately or coincidentally, joint attention achievement was associated with increased suppression of alpha-mu oscillations (i.e., between 11-13 Hz) measured at centro-parietal and parieto-occipital regions. This alpha-mu suppression has been associated with theory of mind processing (Pineda & Hecht, 2009) and social co-ordination (Naeem, Prasad, Watson, & Kelso, 2012). The authors interpreted this as evidence that participants experienced the paradigm as "social", regardless of the relevant instruction for completing each task. However, because this study used frequency domain measures of neural activity – which lack temporal resolution – it is not clear the extent to which this reflects the evaluative stage of joint attention only. Another limitation of the study

is that the interaction – although genuine – occurred in a highly-structured context in which the gaze behaviour of participants (even in the social condition) was heavily cued and contrived. Whereas in social interactions, we typically achieve joint attention in order to service a specific goal (e.g., to signal or share some goal-relevant information). An interactive context which demands the evaluation of goals and goal-driven behaviours in others is likely to differentially demand mentalising processes during the evaluation of joint attention. Therefore, the impact of these processes on the neural encoding of joint attention remain unclear.

An fMRI study by Redcay, Kleiner & Saxe (2012) utilised a more goal-oriented paradigm, in which participants interacted with a partner to complete a cooperative task using their eye-gaze. Due to space constraints, it is difficult to create face to face interactions in studies using fMRI. However, as demonstrated by this study, a live interaction can be facilitated using video feeds as an alternative to face to face interactions. In this paradigm, participants were asked to establish joint attention with their partner to catch a mouse hiding behind one of four blocks of cheese located in each corner of the screen, whilst a video frame of their partner – a member of the research team – was positioned in the centre of the screen. If the participant observed the hiding mouse, it was their task to direct their partner's gaze to it. If they did not observe the mouse, they were instructed to look at their partner and follow their gaze cue. For the baseline, “solo attention” condition, the experimenter closed their eyes while the participant located the mouse by themselves. The study found that compared to the “solo attention” condition, joint attention was associated with activation in the right superior temporal sulcus and right temporoparietal junction. This study is an example of a goal-oriented paradigm, thus offering a more ecologically valid simulation of a social interaction within the constraints of neuroimaging. However, a limitation of the live video feed approach used by Redcay et al. (2012) is that it does not control for neural

processes associated with stimulus features, such as facial expressions and head movements. These may confound or obscure neural processes associated specifically with joint attention.

Addressing this, an fMRI study by Saito et al. (2010) presented participants with a live video feed portraying just the eyes of an interactive partner. The video of each partner's eyes was displayed above two red circles located on the left- and right-hand side of the screen. Participants underwent three conditions, which each started with participants establishing eye-contact. In the first, one participant saw one circle change from red to blue. This participant had to look at this circle, whilst the other had to either follow their gaze cue to look at the same location (establishing joint attention) or were instructed to look elsewhere (avoiding joint attention). During the second condition, both participants observed one circle change colour, and had to gaze at it before re-establishing eye contact, thus either establishing (if the same circle changed colour for each participant) or avoiding incidental joint attention (if different circles changed colour for the two participants). The third condition was the baseline, in which neither of the circles changed colour, and no gaze shift was required from either participant. Inter-subject neural synchronisation (correlated changes in brain activation across the interacting participants) was found in the right inferior frontal gyrus. The authors proposed that this synchronisation reflects the process of understanding another's intentions and perspectives. This interpretation is possible, given that the dyadic interaction enabled participants to adopt an intentional stance towards their partner. However, whilst this paradigm supported genuine interactions, the ecological validity of these interactions was reduced by presenting eyes in isolation. This study, once again, highlights the challenge of achieving experimental control in neuroimaging contexts without compromising reciprocity and the adoption of an intentional stance towards the social stimulus under investigation.

Virtual interactions. One methodology that addresses limitations associated with live video feeds and pre-recorded stimuli is virtual reality, in which a computer-based environment is used to simulate the experience of a social interaction. During virtual joint attention interactions, participants are asked to interact with social partners represented onscreen by avatars (animated human characters) which are programmed to respond to the participant's social behaviour (e.g., eye movements). Participants can engage with these virtual characters in a manner that closely resembles a real, reciprocal, interaction, whilst experimental control is maintained over the virtual character's behaviour and appearance (Bohil, Alicea & Biocca, 2011). This approach can be used to isolate and manipulate specific social cues, and has also been utilised to investigate a number of aspects of social cognition (see Georgescu, Kuzmanovic, Roth, Bente & Vogeley, 2014 for a review), such as facial expression (Carter & Pelphrey, 2008), social exclusion (Kassner et al. 2012; Wesselmann et al. 2012; Williams, 2007), racial implicit biases (Banakou, Hanumanthu & Slater, 2016; Peck, Seinfeld, Aglioti & Slater, 2012) and body perception (Slater & Sanchez-Vives, 2014).

To study the neural correlates of evaluating joint attention, Schilbach et al. (2010) used a virtual interaction paradigm to engage participants – who were laying in an MRI scanner – with an anthropomorphic avatar, whom they believed was controlled by a confederate outside the scanner. This deception allowed participants to adopt an intentional stance towards their virtual partner, when in fact, the avatar was controlled by a gaze-contingent computer algorithm that used input from an eye-tracking camera focussed on the participants to program a contingent response (see Wilms et al., 2010 for more details). This allowed participants to establish eye contact with the avatar before focussing on one of three grey blocks that was situated above, to the left, and to right of the avatar. These blocks turned blue upon fixation. During the task, participants were instructed to do one of three things: (1) initiate joint attention, (2) respond to the avatar's joint attention bid congruently by

looking at the cued location or (3) respond to the avatar's gaze cue incongruently by making an anti-saccade towards a different location. When the participant initiated joint attention, the avatar would respond either congruently to successfully achieve joint attention with the participant or incongruently, to avoid joint attention. This study found that a successful self-initiated joint attention episode resulted in increased activation in the ventral striatum, compared both to unsuccessful self-initiated joint attention, and to joint attention bids (whether successful or unsuccessful) initiated by the virtual partner. Since the ventral striatum has been implicated in social reward processing (Liu et al. 2007; Izuma, Saito & Sadato, 2008), the authors suggest that this reflects a unique hedonic response when one's own joint attention bid is successful. This was the first study to use a reciprocal, yet experimentally-controlled paradigm to investigate the neural correlates of achieving self-initiated joint attention when participants adopted an intentional stance towards their social partner.

The virtual reality paradigm introduced by Schilbach et al. (2010) represents an important step forward towards an experimentally controlled, yet still ecologically valid context for studying social interaction. Nevertheless, as with the paradigms developed by Lachat et al. (2012), Saito et al. (2012) and Williams et al. (2005), the paradigm was limited in that joint attention was not achieved in a social context which was goal-oriented, context-driven, or intuitive. These should be important criterion when designing an ecologically-valid joint attention paradigm given that joint attention behaviours often emerge when a person is focussed on the objective of an interaction with a partner whose behaviour and intentions are processed subconsciously (Tomasello, 2005). Contrastingly, Schilbach et al. (2010) cued participants on each trial about their social role (i.e., initiator or responder) and whether to engage in joint attention or not – thus removing the ability for participants to intuitively engage in social evaluation processes typically engaged in face-to-face

interactions. Furthermore, the achievement of – or failure to achieve – joint attention was evaluated in a context where either outcome was inconsequential given the absence of a goal-directed task.

Caruana and colleagues attempted to address these limitations by extending Schilbach et al.'s (2010) interactive approach using a goal-oriented task, called the Burglar Game (Caruana, Brock & Woolgar, 2015). In their fMRI study, participants were presented with two rows of houses with closed doors, and an interactive human-like avatar. Participants were told that the avatar was controlled by a member of the research team (called 'Alan') located outside the scanner, thus ensuring that participants adopted an intentional stance. As in the Schilbach paradigm described earlier (Schilbach et al. 2010), the avatar was actually controlled by a gaze-contingent algorithm which used the online recordings of participant's eye movements to program contingent responses displayed by the avatar. During the "social" condition, participants performed a cooperative search task with the virtual character, coordinating their attention to locate a burglar hidden in one of the houses. If the participant found the burglar in the houses they were allocated to search, they would establish eye contact with the avatar and direct its gaze to the correct location. If they did not find the burglar during the search phase, they were to wait for the virtual character to make eye contact and follow the avatar's subsequent saccade. For a control condition, the avatar's eye gaze cues were replaced by a dynamic arrow, which participants were aware was controlled by a computer program. The goal of each joint attention episode was to search for and cooperatively catch a burglar. Thus, the concerns of ecological validity raised about the abstract task presented to participants in the Schilbach et al. (2010) paradigm are addressed, by offering participants a more intuitive social context in which to experience joint attention.

This study found that self-initiated joint attention resulted in increased activation in the right medial, superior and inferior frontal gyri, and right temporoparietal junction, compared both to a successful joint attention initiated by a social partner, and to neural activation during the non-social condition, when participants interacted with arrow cues instead of eye gaze. This corresponds with neural activation identified by Redcay, Kleiner & Saxe (2012) and Gordon et al. (2013) as part of the "social" brain network (Pfeiffer, Vogeley & Schilbach, 2013). Although this study does manipulate intentional stance using a deceptive cover story in the social condition, the results do not inform us of its role in joint attention because the spatial cues in the non-social condition were visually different. As such, we do not know whether differential activation between these conditions is due to the adoption of an intentional stance or the evaluation of eye gaze to achieve joint attention in a social context. Furthermore, as with Redcay et al. (2012), this study did not manipulate the success of joint attention, and therefore does not inform us as to the neural mechanisms of evaluating joint attention.

In an event-related potential (ERP) study of joint attention, Caruana and colleagues developed a similar paradigm called the Prisoner Task (Caruana, de Lissa & McArthur, 2015). Like the Burglar Game (Caruana, Brock & Woolgar, 2015), this paradigm contextualises joint attention so that the interaction is goal-driven. Participants were assigned the role of a prison watchperson who had to monitor the exterior of a prison that was represented by four grey buildings in the four corners of a computer screen. The middle of the screen displayed an avatar which was programmed by a gaze-contingent computer algorithm to respond to the participant's gaze cues. Again, to ensure that an intentional stance was adopted towards the virtual character, participants were told that the avatar was controlled by another member of the research team located in another room (again, 'Alan'). Each trial started with a spotlight appearing on one of the four buildings, indicating the

location of the escaping prisoner. Participants initiated joint attention by first looking at the spotlight until the prisoner appeared, and then shifting their gaze back to their partner. 'Alan' would then either respond congruently to achieve joint attention by looking at the cued location or incongruently by looking at one of the other three locations. Successful (i.e., congruent) and unsuccessful (i.e., incongruent) joint attention episodes took place with equal probability (i.e., the avatar responded congruently on 50% of trials) in a randomised presentation order. The observation of incongruent gaze shifts elicited significantly larger centro-parietal P350 responses compared to congruent gaze shifts. The same effect of congruency was not observed in a control group who completed a non-social version of the task, in which the avatar's eyes were closed, and a computer-controlled arrow stimulus 'responded' congruently or incongruently in lieu of the avatar's eye movements. The authors interpreted the P350 effect in the social group as reflecting the evaluation of gaze cues which signal the avoidance of joint attention. The failure to replicate the same effect with arrow stimuli suggests that this is unlikely a domain-general effect of attention or spatial congruity. However, because both belief and the appearance of the directional cue (eyes versus arrow), it is not known whether the observed P350 effect is related to the appearance of the cue (social versus non-social) or the adoption of an intentional stance. More specific research is needed to specifically explore the effects of adopting an intentional stance on the neural encoding of joint attention.

The Intentional Stance

To summarise, we have insight into the neural responses associated with evaluating self-initiated joint attention bids using virtual interaction paradigms (Schilbach et al., 2010; Caruana et al., 2015). This methodology maximises experimental control as well as ecological validity by manipulating beliefs about the agency of the virtual partner to induce an "intentional stance" without requiring a genuine dyadic interaction. Whilst adopting an

intentional stance is a core feature of real-life joint attention interactions, few studies have directly manipulated intentional stance, in order to investigate its influence on the neural evaluation of joint attention experiences. This is also critical in confirming the extent to which mentalising mechanisms are engaged during joint attention (Williams et al., 2005).

The effects of adopting an intentional stance towards a social partner have, however, been explored in a number of non-joint attention contexts. Several studies have presented evidence suggesting that adopting an intentional stance modulates the neural processing of social information. For example, a series of ERP studies by Schindler and colleagues investigated how the brain responds to language-based personality feedback (i.e., adjectives presented on a screen, e.g., "weak"). This feedback was either believed to be generated by another human partner who endorsed or disagreed with the feedback, or believed to be randomly selected by a computer program (Schindler, Wegrzyn, Steppacher & Kissler, 2014, 2015; Schindler & Kissler, 2018). Another study compared neural processing when participants believed personality feedback to originate from a human evaluator and a "socially intelligent" (i.e., not random) computer software interface (Schindler & Kissler, 2016). Across all these studies, the belief that the feedback originated from a human source consistently enhanced late positive potentials (LPPs) measured at centro-parietal sites, across all four studies, particularly with positive (e.g., "happy") or negative (e.g., "weak") emotional adjectives. Central components in the P2 (150-200 ms) and P3 (300-450 ms) temporal ranges post-stimulus onset, and occipital early posterior negativity (EPN) components were also enhanced during the human-sender condition (Schindler et al. 2015; Schindler & Kissler, 2016, 2018), although this was not the case for every study (Schindler et al. 2014). A more recent fMRI study using a similar paradigm found more significant neural activity in the superior frontal, medial prefrontal and orbitofrontal cortices during the 'human-sender' condition, compared to computer feedback conditions (Schindler, Kruse,

Stark & Kissler, 2019). These are all brain regions which have been previously associated with the "social brain" or "mentalising network" (Frith & Frith, 1999; Van Overwalle & Baetens, 2009). It appears that even without visual social information, adopting the intentional stance towards a virtual interlocular still alters neural processing of social information and communication. These studies, however, do not inform us what these effects might look like when viewing and evaluating non-verbal social cues embedded in faces.

Wiese, Wykowska, Zwickel & Müller (2012) were among the first to study the effects of intentional stance on evaluating non-verbal cues, in the context of a gaze-cueing paradigm. The basic gaze-cueing task presents a centralised face, either a human face or a humanoid robot. After presentation, this face would look to the left or to the right, before a target letter (T or K) appeared to on either the left or right side of the face. Participants were asked to press the corresponding letter on the computer keyboard to detect the target's location as quickly as they could. Intentional stance was manipulated by inducing human agency beliefs using a deceptive cover story. Participants were instructed that the human face was either a human or a human-like mannequin, and that the robot was either pre-programmed or human-controlled. As is typical of gaze-cueing paradigms, the reaction time for participants to respond was longer when the gaze shift invalidly cued the target location than when it was valid (i.e., a validity effect; Posner, 1980). Interestingly, the validity effect was larger when participants believed the face was human-controlled. This effect was consistent for both human and robot faces. In other words, participants found it harder to ignore gaze cues when they were believed to signal the perspective of another, intentional and sentient human being. The authors argue that this is because greater significance is ascribed to the gaze cues when the observer adopts an intentional stance, even though the belief is not directly relevant to the task at hand.

In a follow-up study using the same gaze-cueing paradigm, Wykowska and colleagues performed two ERP experiments to investigate the neural correlates of intentional stance within this gaze-cueing context (Wykowska, Wiese, Prosser & Müller, 2014). The first experiment manipulated intentional stance indirectly using stimulus appearance, where participants interacted with either a human or a robot face. During the second experiment, participants interacted with a robot face only, but were instructed either that the robot was controlled by a human or by a computer. Thus, intentional stance was directly manipulated in the second experiment. This study discovered that in addition to enhancing behavioural validity effects, adopting an intentional stance resulted in significantly larger P1 responses to validly-cued targets (100-140ms, time-locked to the appearance of the target stimulus) at posterior parietal and occipital electrodes than invalidly-cued targets, and cued targets when an intentional stance was *not* adopted. These effects were consistent regardless of whether intentional stance was manipulated directly (i.e., by using a deceptive cover story) or indirectly, by manipulating the aesthetic humanness of the stimulus. As a reaction to these findings, the authors proposed the Intentional Stance Model (ISM) of social cognition which posits that neural mechanisms associated with mentalising (such as the medial prefrontal cortex and temporoparietal junction) are recruited when an intentional stance is adopted towards an entity. These cortical mechanisms then have the potential of modulating (e.g., prioritising) the early visual processing of social information. This account also aligns with Schindler and colleagues' interpretation of the enhancement of the ERPs associated with processing emotive personality feedback (i.e., P2, EPN, P3 and LPP) when it is believed to originate from another person (Schindler et al. 2015; Schindler & Kissler, 2016, 2018). However, what remains unclear from these studies is the extent to which the aesthetic realism of the face stimuli impacts on the adoption of an implicit or explicit intentional stance – and the corresponding neural consequences of this.

To elucidate this further, Abubshait and Wiese (2017) used a similar gaze-cueing paradigm to investigate the effects of aesthetic anthropomorphic realism and behaviour on the adoption of an intentional stance or "mind perception". Two stimuli were created by morphing a photographed human face and a photographed humanoid robot face (the Meka S2). Specifically, these faces were morphed to produce images that were "human-like" (80% human, 20% robot), or "robot-like" (20% human, 80% robot), whilst equating both stimulus categories for overall structural and low-level features. One group performed the gaze-cueing task with faces which looked at the correct location 80% of the time (i.e., the target was validly cued on 80% of trials), while for the other group, only 50% of the gaze shifts were valid cues. Participants were asked to rate the likelihood that the agent "had a mind" at the beginning and end of each block. The study found that while the aesthetically human-like stimulus (80% human) resulted in higher participant ratings of mind perception, it was the face that cued a higher percentage of trials validly (80% reliable gaze cues) that had the largest gaze-cueing effect. However, the mind perception ratings (i.e., the rating that the agent was likely to "have a mind") did not change significantly pre- and post-test, indicating that they were not significantly affected by the predictability of the agents' behaviour. The results from these studies suggest that the social behaviour of a virtual partner (e.g., an agent who provides more reliable and predictive social information) is more likely to promote a subjective intentional stance than increased aesthetic realism. However, the neural consequences of these intentional stance effects, and the role of aesthetic realism unclear.

Wiese, Buzzell, Abubshait & Beatty (2018) used a similar gaze-cueing paradigm to both Abubshait and Wiese (2017) and Wykowska et al. (2014) to investigate the neural correlates of "mind perception". Six stimuli were created by morphing a photograph of a human face with a photograph of the Meka S2 to produce six images that varied in their "human-likeness" (100% robot; 80% robot, 20% human; 60% robot, 40% human; 40%

robot, 60% human; 20% robot, 80% human; 100% human). During fMRI, participants were shown each of these faces repeatedly and each time were asked to rate the likelihood that the agent "had a mind". Participants then completed a non-predictive gaze-cueing task (outside the scanner) using each of the stimuli. The authors reported that the more human-like stimuli were associated with higher 'mind perception' ratings and these were associated with increased activation within the ventromedial prefrontal cortex. Outside the scanner, more human-like stimuli also predicted larger gaze-cueing validity effects. Activation in the left temporoparietal junction, right fusiform cortex and middle temporal gyrus were also correlated to performance in the gaze-cueing task, but this pattern of activation was not correlated with mind perception ratings. Whilst the neural consequences of adopting an intentional stance are not measured during the behavioural cueing task, these findings suggest that the extent to which we subjectively adopt an intentional stance may be influenced by aesthetic realism, but more critically, this influences the neural encoding of faces and predicts the social significance we ascribe to gaze cues. Given these findings, it stands to reason that adopting an intentional stance should also influence how we evaluate eye gaze cues during joint attention interactions.

To investigate this, Pfeiffer et al. (2014) utilised the virtual reality and eye-tracking joint attention paradigm developed by Schilbach et al. (2010) to investigate the neural correlates of the subjective experience of social interaction, and the adoption of an intentional stance using fMRI. In this paradigm, participants interacted with an animated human face, which was programmed by a gaze-dependent computer algorithm to create the experience of joint attention. Participants had to initiate joint attention over two grey blocks (left and right) with the avatar. They were told that on some blocks the avatar would be controlled by another human partner, and on others by a computer. However, they were not told when this would occur. On each block of five trials, the avatar responded congruently

either 20%, 40%, 60% or 80% of the time. After every five trials, participants were required to make a Turing-task decision, indicating whether they believed the avatar in that block was 'human' or 'computer'. Participants were more likely to indicate that the avatar was controlled by a person during blocks where the avatar responded more congruently (i.e., blocks where the majority of trials led to the achievement of joint attention). This study found that the experience of social interaction (i.e., during the blocks in which participants rated the avatar more likely to be controlled by another human) activated the mesolimbic reward system, including the medial orbitofrontal cortex and ventral striatum. This is consistent with the findings from Schilbach et al. (2010) described earlier. However, Pfeiffer et al. (2014) did not directly manipulate intentional stance using a deceptive cover story. Rather, here we can only use subjective and retrospective ratings from participants to gauge whether they adopted an intentional stance. Given that these ratings were made retrospective (i.e., after completing the associated block) it is also unclear as to whether they were explicitly adopting an intentional stance towards the virtual character during the analysis period. Finally, in this study the authors report that participants were more likely to adopt an intentional stance when their partner responded more congruently. As such, it is unclear whether the differential activation observed in social reward networks here reflect the adoption of an intentional stance or the evaluation of successful joint attention bids since the two experiences are conflated.

In order to separate the neural consequences of evaluating the achievement of gaze-cued joint attention and adopting an intentional stance towards a virtual partner, Caruana, de Lissa & McArthur (2017) set out to directly manipulate both of these factors in a follow-up ERP study using their Prisoner Task paradigm (Caruana, de Lissa & McArthur, 2015). As described above, participants in this paradigm initiate joint attention bids with a virtual character during a goal-oriented game. The virtual partner responded congruently 50% of

the time to achieve joint attention, or incongruently to avoid joint attention. Intentional stance was manipulated between-subjects so that one group of participants were led to believe that the avatar was controlled by a person, and the other were truthfully told that it was controlled by a gaze-contingent computer program. The social stimulus was identical for both groups. ERPs time-locked to the virtual character's gaze shift elicited P250 (170-300ms) and P350 (310-440ms) peaks at centro-parietal electrode sites (CZ and PZ). Critically, P350 responses were significantly larger in response to incongruent gaze shifts (which avoided joint attention) than congruent gaze shifts (signalling the achievement of joint attention) – but only in the group of individuals who believed they were interacting with another human. Given that there was a high degree of variably in responses within the computer-belief condition, it is possible that the failure to identify the same effects in this group are due to individual differences between groups, or a failure to replicate the original effects of evaluating the achievement of joint attention.

To address this, Caruana and McArthur (2019) conducted a study using the same paradigm, in which they manipulated intentional stance beliefs within subjects. Each participant completed two blocks of the task, the order of which was counterbalanced between participants. They were told at the beginning of the experiment that during one block the virtual character would be controlled by a person, and that during the other block, it would be controlled by a computer. This study found a belief by congruency interaction at centro-parietal electrodes (CZ and PZ). That is, when participants adopted an intentional stance and observed congruent gaze shifts, P250 waveforms were significantly larger than those time-locked to gaze shifts that were either incongruent, or when the participant did not adopt an intentional stance. In contrast, P350 waveforms were significantly larger across both human and computer belief conditions when the virtual partner responded incongruently than congruently. However, it was evident that the absence of an international

stance modulation on the P350 was obscured by the fact that P350 responses were building upon the earlier P250 peaks. Thus, in order to control for the influence of P250 responses on the measurement of the P350 effect, Caruana and McArthur also analysed the difference between the P350 and P250 mean amplitudes (i.e., P350–P250). The greatest difference between the P350 and P250 waveforms was observed when the avatar responded incongruently, but only when it was believed to be controlled by a human – which was consistent with the findings of Caruana et al., (2017). Again, this study found that neural responses to gaze shifts were more variable across individuals when an intentional stance was not adopted. The authors suggest that this might be because the induction of an intentional stance standardises the extent to which individuals anthropomorphise their virtual partner, thus minimising the potential for individual differences in dispositional anthropomorphism to influence neural responses. However, this study was unable to provide any empirical support for this claim. Nevertheless, this study did establish that explicit intentional stance beliefs can be manipulated within subjects, and that this does influence – indeed, possibly underscores – the neural encoding of joint attention experiences.

To summarise, evaluating the success or failure of a self-initiated joint attention bid involves representing another's perspective and comparing it with one's own, which is reliant upon adopting an intentional stance towards the other. This perspective is known to affect neural processing in a number of contexts, but few studies have directly investigated the effects of intentional stance on neural processing in joint attention. Recent work by Caruana and McArthur (2019) presents a potential ERP neural marker for the evaluation of joint attention which also appears to be dependent on the explicit adoption of intentional stance. In this way, these ERPs may also present potential neural markers for adopting an intentional stance – which may be invaluable in human-robot interaction (HRI) settings to objectively evaluate whether humans engage with and evaluate socially-responsive robots

and artificial agents as they do other humans (Cross, Hortensius & Wykowska, 2019). However, in order for this to be used in such a way, the reliability of this neural marker, and the extent to which its measurement is influenced by (1) the aesthetic properties of virtual agent and (2) individual differences in the dispositional tendency to anthropomorphise non-human entities, needs further investigation.

Current Study

The current study aimed to investigate the reliability of the centro-parietal P250 and P350-P250 ERP effects reported by Caruana and McArthur (2019). Specifically, this study had three aims. First, we wanted to establish whether these ERP effects of joint attention and explicit intentional stance beliefs could be replicated, using the same stimuli as Caruana and McArthur, but a different experimenter, and a different electroencephalography (EEG) acquisition system in order to determine the reliably of these effects across time and research contexts. Second, we wanted to determine whether these effects could also be replicated using robot face stimuli (the Meka S2 robot; see Abubshait & Wiese, 2017) to determine whether these neural markers are also reliable across stimuli varying in aesthetic anthropomorphic realism. Finally, we wanted to explore whether individual differences in dispositional anthropomorphism explain individual differences in the ERP effects of joint attention when participants do not adopt an explicit intentional stance.

To this end, the current study adopted the same protocol as that used by Caruana and McArthur (2019). However, participants were randomly allocated into two stimulus groups – in which they either completed the experiment using the same animated human face or a robot face. We also used the Anthropomorphism Quotient (AQ; Neave et al. 2015) to measure each individual's dispositional tendency to anthropomorphise non-human entities.

For the group interacting with the human-looking avatar, we expected to find that when adopting an intentional stance (i.e., human belief) observing congruent gaze shifts

would elicit significantly larger P250 waveforms than incongruent gaze shifts, or either congruent and incongruent gaze shifts when the participant does not adopt an intentional stance (i.e., computer belief). We also expected to see larger differences between P350 and P250 amplitudes (i.e. P350-P250) following incongruent gaze shifts during the intentional stance condition than the computer belief condition. These expectations were based on the results obtained by Caruana and McArthur (2019). Additionally, and consistent with the findings from Abubshait and Weise (2017), we expected that these effects of intentional stance would be observed in both the human and robot face stimulus groups. Finally, we expected variability in the ERP effects of congruency to be associated with individual differences in dispositional anthropomorphism when participants did not adopt an intentional stance. However, these analyses were exploratory and directional hypotheses were not established prior to conducting the experiment.

Methods

Participants

All participants for this study were recruited from undergraduate Psychology cohorts at Macquarie University via an online recruitment platform. Participants gave written consent before testing commenced and received course credit for their time. All participants were right-handed (assessed using the Edinburgh Handedness Inventory; Oldfield, 1971) and had normal or corrected-to-normal vision (clear contact lenses were permitted as this did not interfere with eye-tracking calibration). Potential participants were also screened for any history of brain injury, neurological conditions (e.g. epilepsy) and psychiatric diagnoses (e.g. schizophrenia).

Twenty-nine participants (16 female) were recruited for the group who interacted with the human-like avatar. Three participants were excluded prior to data processing due to technical issues during data collection, and five participants were excluded before data

analysis as they did not believe the deceptive cover story. The final human-face group consisted of 21 participants (11 female; M_{age} = 19.3, SD = 1.80; M_{belief} = 8.5).

Thirty-one participants (22 female) were recruited for the group who interacted with the humanoid robot face. Two participants were excluded prior to data processing due to technical issues, nine were excluded prior to analysis when participants did not believe the cover story, and one was excluded due to an excessive number of trials lost during processing due to inaccuracy or artefacts (> 3 SD of the group mean). The final robot group consisted of 19 participants (16 females; M_{age} = 19.6, SD = 2.24; M_{belief} = 9.2).

Participants were randomly allocated to the two stimulus groups and the order of belief conditions within the human and robot groups was counterbalanced by order of recruitment. That is, the first participant was allocated to the human-like face stimulus group and first completed the task whilst believing their partner was human-controlled and then computer-controlled. The second participant completed the reverse order of belief conditions, again with the human-like face. The third participant was allocated to the robot face group and first believed the robot was human-controlled, and the fourth participant completed was allocated to the same robot stimulus group but completed the belief conditions in the reverse order. This pattern was repeated for the entire sample of 60 participants tested.

Stimuli

Participants in the *human-face* group interacted with an anthropomorphic avatar created in *FaceGen*, originally used in the first study utilizing the Prisoner Task (Caruana, de Lissa & McArthur, 2015). This animated human face subtended 10.1 x 6.5 degrees of visual angle, while the eye-well area (the area of interest defined around the eyes) subtended 1.5 x 4.9 visual degrees. Participants in the *robot-face* group interacted with a humanoid robot face (the Meka S2; Abubshait & Wiese, 2012) in place of the computer-generated

human avatar. The robot face subtended 5.9 x 6.3°, and the eye-well 2.3 x 5.3°. Both the human face and the robot face stimuli were edited in GIMP 2.0 to produced five images: looking to the four corners of the screen (i.e. to the top right, bottom right, top left and bottom left), as well as directly ahead to simulate eye-contact with the participant.

The experimental paradigm was created and run in SR Research Experiment Builder and presented on an AOC computer monitor (60 cm x 34 cm), situated 75 cm from each participant's eyes, with a refresh rate of 144 Hz. The faces were presented in the centre of the computer screen surrounded by four prison buildings located in each corner (each building subtended 9.0 x 10.2 degrees of visual angle).

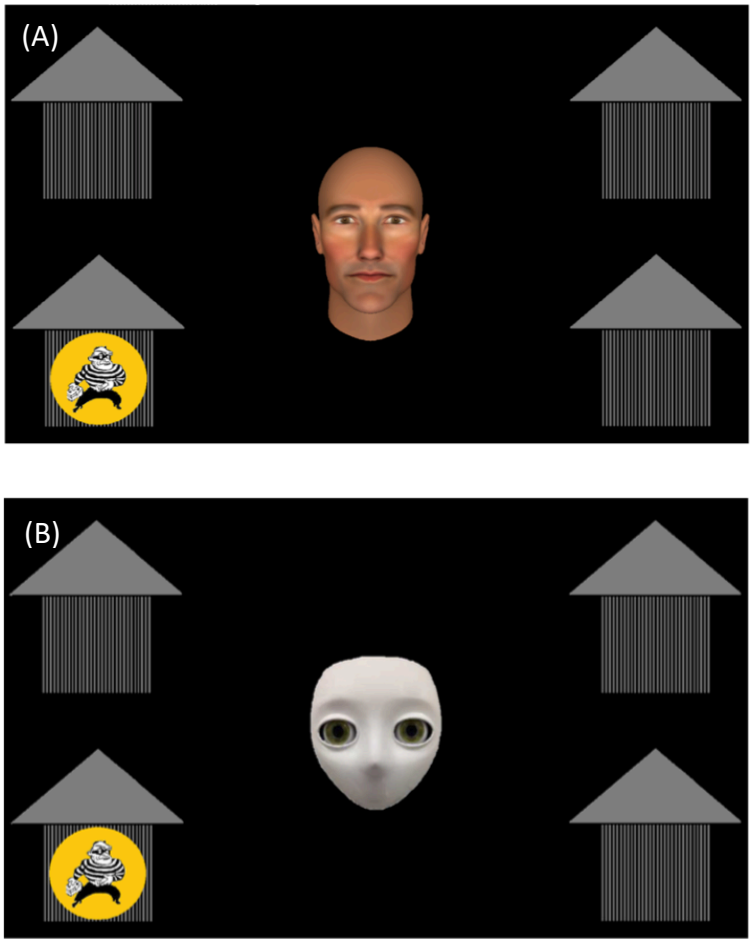


Figure 1. Face and task stimuli presented to the two groups of participants. (A) the human avatar face created in *FaceGen*. (B) The Meka S2 robot face taken from Abubshait & Weise, 2017.

Task

The task in this study was identical to the one described by Caruana & McArthur (2019). Participants were informed that a prison break was in progress in the compound that they could see on the computer monitor. As 'prison watchman' it was the participant's task to prevent the inmate's escape by informing their virtual partner – the 'prison guard' – of the location that the prisoner was attempting to escape from on each trial. Participants were required to do this by initiating joint attention towards the correct prison building. During one block of trials, the prison guard – represented by the human or robot face, shown in the centre of the screen – was purportedly controlled by a member of the research team, located in a nearby eye-tracking laboratory. Importantly, participants were told that just as they could only see the exterior of the prison, their partner could only see the interior. Thus, teamwork was required to complete the task. When the prisoner appeared at one of the prison blocks, participants had to guide their partner by initiating joint attention towards the correct location, with the understanding that the guard would then try to follow their gaze and lock down the breached exit.

Each trial was preceded by the presentation of a central crosshair subtending 0.8 degrees of visual angle. For the trial to start, participants were required to provide an ongoing check of the eye-tracking calibration. The crosshair disappeared and was replaced by the avatar (either the human or robot face) surrounded by the four prison buildings. After a jittered time delay of 200-1000ms, a "spotlight" (subtending 4.9 degrees of visual angle) appeared on one of the four prison buildings, indicating the prisoner's location. Participants fixated on the spotlight for at least 150ms to initiate joint attention towards the breached exit. The prisoner then appeared in the centre of the spotlight, after which participants were then required to return their gaze to the avatar's eyes and monitor their response. After another jittered time delay (350-650 ms), the avatar either responded congruently by

following the participant's gaze cue to establish joint attention (50% of trials), or incongruently by looking at another of the three buildings (50% of trials). The direction of the incongruent gaze shift was counterbalanced across trials. Trial order was also randomized throughout each block, so that participants could not predict the avatar's response. Participants were cued to blink during the inter-trial interval, to minimise their need to blink during the trial.

In order to implicitly explain the high proportion of incongruent gaze shifts made by the virtual partner, participants were told that their partner was sometimes distracted by 'fights' occurring within the prison compound (not visible to the participant) which they were required to detect and stop. A mock display of this task was shown to participants before the experiment so that they understood the task from their partner's perspective. This ensured that participants would not reject the belief that their partner was human, even though there was a low joint attention rate.

Each participant completed two blocks of the task, one in which the stimulus was purportedly controlled by another person, and one for which participants knew the avatar was controlled by a computer. Each block contained 120 trials, and participants were given a short break every 30 trials. During these breaks, they were asked to estimate the proportion of trials that they had achieved joint attention with their partner. This encouraged participants to engage with the task, whilst ensuring attention to their partner's response was maintained throughout the experiment.

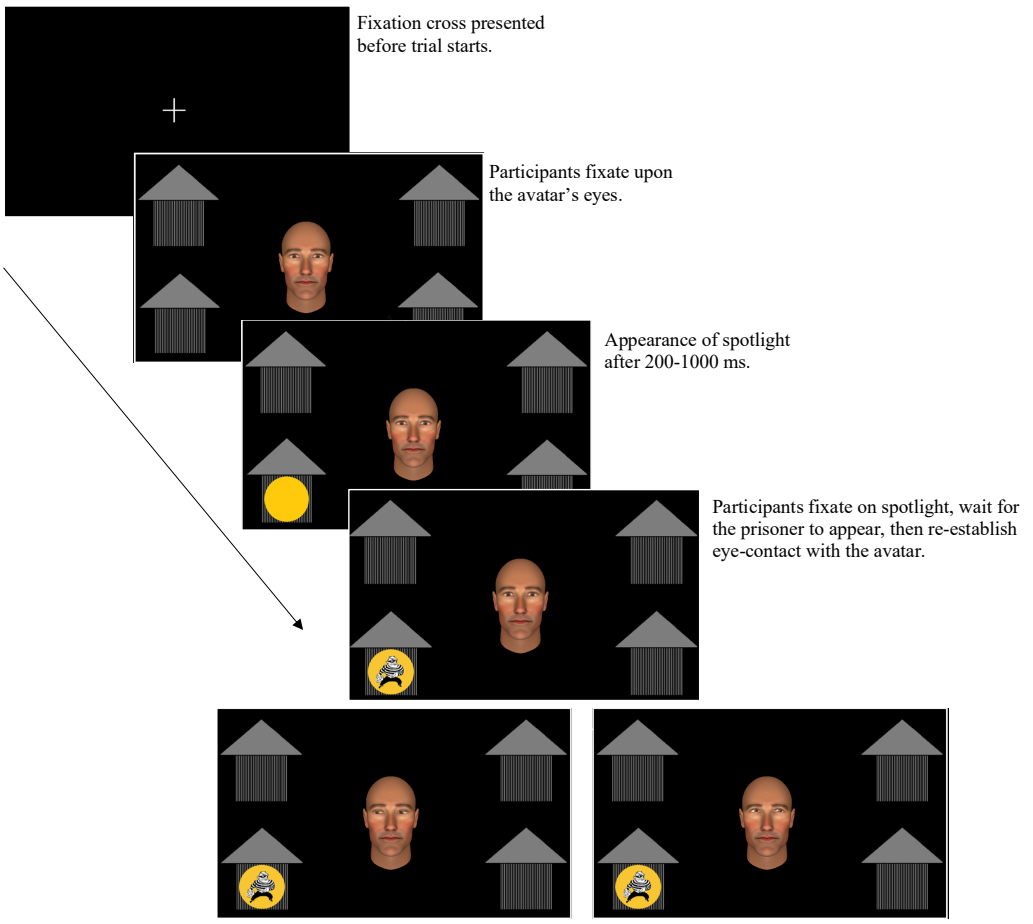


Figure 2. The trial sequence for participants interacting with the human face. Note that the location of the escaping prisoner, as well as the location that the avatar looked during the incongruent trials were counterbalanced.

Eye movements

Participant eye movements were recorded during the task using an EyeLink 1000 tower-mounted eye-tracker (SR Research), monitoring the right eye only, with a refresh rate of 500Hz. A chinrest was used to stabilise head movements and standardize viewing distance during the task. The eye-tracker was calibrated for each participant using a 9-Point calibration sequence, and recalibrated if participants moved their head during the experiment. The validity of the eye-tracking calibration was checked before each trial using a gaze-contingent crosshair fixation.

ERPs

EEG data was collected from a montage of 29 electrodes positioned according to the International 10-20 system, using an EasyCap32 (Synamps). The ground electrode was located between Fp1, Fp2 and Fz, and reference electrodes were placed on the outer earlobes of both ears. Ocular movements were recorded by electrodes positioned on the temples (HEOG) and above and below the left eye (VEOG). All electrode impedances were maintained below 15kΩ.

Raw EEG data was processed offline in MATLAB (version 2017a) using EEGLAB (version 14.1.2; Delorme & Makeig, 2004). A 50 Hz notch filter was applied to remove any electrical noise, then the data was re-referenced to the M2 reference electrode (positioned on the right earlobe). An independent-components analysis (ICA) was used to then identify blinks, which – where identified – were rejected. It is noteworthy that we minimized the occurrence of blinks experimentally by encouraging participants to blink between trials. Furthermore, the visual event of interest (i.e., the avatar's gaze shift) was gaze contingent, occurring, on average half a second after fixation. As such, the occurrence blinks during the analyzed event epochs were unlikely, and indeed rare.

Following the ICA, the continuous data was bandpass filtered (0.1-30 Hz) with a 12dB octave roll-off. The continuous data was then epoched starting 100 ms before the onset of the avatar's responsive gaze shift and ending 800 ms later (i.e., 0 ms to 700 ms). Epochs containing extreme voltages (+/- 100 mV) were automatically rejected. We also rejected error trials. These included trials where the participant failed to correctly fixate the spotlight 3000 ms after appearing on the screen, or trials in which the participant, upon initiating joint attention and fixating back on the avatar's face, looked away from the face before the end of the trial. This ensured that we only retained epochs in which the participant maintained fixation on the avatar's face. Finally, data from accepted epochs were averaged to create

four ERPs for each participant in each group (human-face, robot-face) for each condition (human-belief congruent, human-belief incongruent, robot-belief congruent, robot-belief incongruent.

We followed the same protocol for measuring ERPs at CZ and PZ as those established in previous work using this paradigm (see Caruana & McArthur, 2019). Specifically, we calculated mean amplitudes for the P250 (170-300ms) and the P350 (310-440ms). The former was subtracted from the latter to calculate the P350-P250 metric. We report the effects separately for CZ and PZ to be consistent with previous findings. However, the same pattern of results is found at CPZ and when we use a cluster of CZ, CPZ and PZ electrodes (see supplementary material).

Post-Experimental Interview and Additional Measures

Before testing commenced, in addition to the Edinburgh Handedness Inventory, participants also completed an Anthropomorphism Questionnaire (AnthQ; Neave et al. 2015), a series of 20 questions to measure individual anthropomorphic tendencies.

At the end of the testing session, participants completed a verbal subjective experience interview that asked participants to rate on a scale of 1 (‘not at all’) to 10 (‘extremely’), how pleasant, natural, and difficult they found each block of the task, and how human-like the human or robot avatar ‘behaved’, ‘felt’ and ‘appeared’. Participants also rated how cooperative they felt their partner was during the human-belief condition. Participants were also asked to indicate whether they preferred the block in which they interacted with a human or with the computer, and whether they would choose to complete such an interaction face-to-face with a stranger or using a virtual interface such as the one in the experiment.

Once the questionnaire had been completed, the researcher debriefed the participants, outlining the nature of and justification for the deception used in the

experiment. All participants provided their written consent once more to be involved in the study. Finally, to check the belief manipulation, participants were asked to rate how convinced they had been that they were interacting with a real person during the human-belief block, using the same 10-point scale. All participants who responded with less than seven to this question were excluded from further analysis ($n_{human\text{-}face} = 5$; $n_{robot\text{-}face} = 9$).

Statistical Analyses

The between-subjects effect of stimulus group (human-face versus robot-face) and within-subjects effects of belief (human-belief versus computer-belief) and congruency (congruent versus incongruent) were analysed in a 2 x 2 x 2 mixed ANOVA for the effects on the P250 and P350-P250 mean amplitudes at CZ and PZ.

Consistent with previous work, we also ran planned follow-up analyses to test for any potential effects of block order (i.e., human-belief first versus computer-belief first) using 2 x 2 repeated measures ANOVAs for P250 and P350-P250 mean amplitudes at CZ and PZ. This was to determine whether the strength of the intentional stance manipulation was influenced by belief order (i.e., whether you first engage with the stimulus believing it human-controlled).

Exploratory analysis. We also performed an exploratory analysis to determine whether individual differences in dispositional tendencies to anthropomorphise non-human entities was associated with variation in our joint attention ERP effects when an intentional stance was not explicitly adopted (i.e., during the computer-belief condition). To this end, we conducted Pearson correlational analyses between participants' AnthQ scores and the difference between the ERPs following congruent and incongruent responses during the computer-belief condition. This correlation analysis was conducted for both the P250 and P350-P250 measures.

Results

ERPs

P250. There was a significant main effect of congruency for the P250 response at CZ ($F(1, 39) = 19.89$; $p < .001$; $\eta^2 = .02$) and PZ ($F(1, 39) = 16.21$; $p < .001$; $\eta^2 = .02$). The mean P250 waveform was larger when the avatar responded congruently to participants' gaze cues, regardless of the type of stimulus or intentional stance belief. There was no main effect of belief [CZ ($F(1, 39) = 2.45$; $p = .126$; $\eta^2 = .004$); PZ ($F(1, 39) = 3.45$; $p = .071$; $\eta^2 = .006$)], and no interaction between belief and congruency [CZ ($F(1, 39) = 0.18$; $p = .677$; $\eta^2 < .001$); PZ ($F(1, 39) = 0.77$; $p = .385$; $\eta^2 = .001$)].

The main effect of stimulus was not significant at CZ ($F(1, 39) = 0.94$; $p = .338$; $\eta^2 = .02$) or at PZ ($F(1, 39) = 1.21$; $p = .278$; $\eta^2 = .03$). There was also no evidence of a significant interaction between stimulus group and congruency [CZ ($F(1, 39) = 1.01$; $p = .323$; $\eta^2 = .001$); PZ ($F(1, 39) = 1.86$; $p = .181$; $\eta^2 = .003$)], belief and stimulus [CZ ($F(1, 39) = 0.43$; $p = .517$; $\eta^2 = .001$); PZ ($F(1, 39) = 0.12$; $p = .726$; $\eta^2 < .001$)] or stimulus, congruency and belief [CZ ($F(1, 39) = 0.18$; $p = .675$; $\eta^2 < .001$); PZ ($F(1, 39) = 1.15$; $p = .291$; $\eta^2 = .001$)]. The group average ERPs for both stimulus groups at PZ and CZ are seen in Figure 3.

P350-P250. The main effect of congruency was also significant for the P350-P250 waveform at CZ ($F(1, 39) = 35.02$; $p < .001$; $\eta^2 = .09$) and at PZ ($F(1, 39) = 14.48$; $p < .001$; $\eta^2 = .04$). This effect was the opposite of that seen in the P250 range. That is, incongruent gaze shifts resulted in larger increases in mean amplitudes regardless of participants' belief. There was no main effect of belief [CZ ($F(1, 39) = 1.30$; $p = .261$; $\eta^2 = .003$); PZ ($F(1, 39) = 0.67$; $p = .419$; $\eta^2 = .001$)] or stimulus [CZ ($F(1, 39) = 1.49$; $p = .23$; $\eta^2 = .038$); PZ ($F(1, 39) = 2.21$; $p = .146$; $\eta^2 = .055$)]. There was no interaction between belief and congruency [CZ ($F(1, 39) = 2.59$; $p = .116$; $\eta^2 = .002$); PZ ($F(1, 39) = 0.80$; $p = .376$; $\eta^2 = .001$)], belief

and stimulus [CZ ($F(1, 39) = 1.24$; $p = .272$; $\eta^2 < .001$); PZ ($F(1, 39) = 0.64$; $p = .428$; $\eta^2 = .001$)] or congruency and stimulus [CZ ($F(1, 39) = 0.31$; $p = .582$; $\eta^2 = .001$); PZ ($F(1, 39) = 0.02$; $p = .888$; $\eta^2 < .001$)]. Likewise, there was no three-way interaction between belief, congruency and stimulus at CZ ($F(1, 39) = 0.26$; $p = .613$; $\eta^2 < .001$) or PZ ($F(1, 39) = 0.34$; $p = .565$; $\eta^2 < .001$). These effects are summarised in Figure 4.

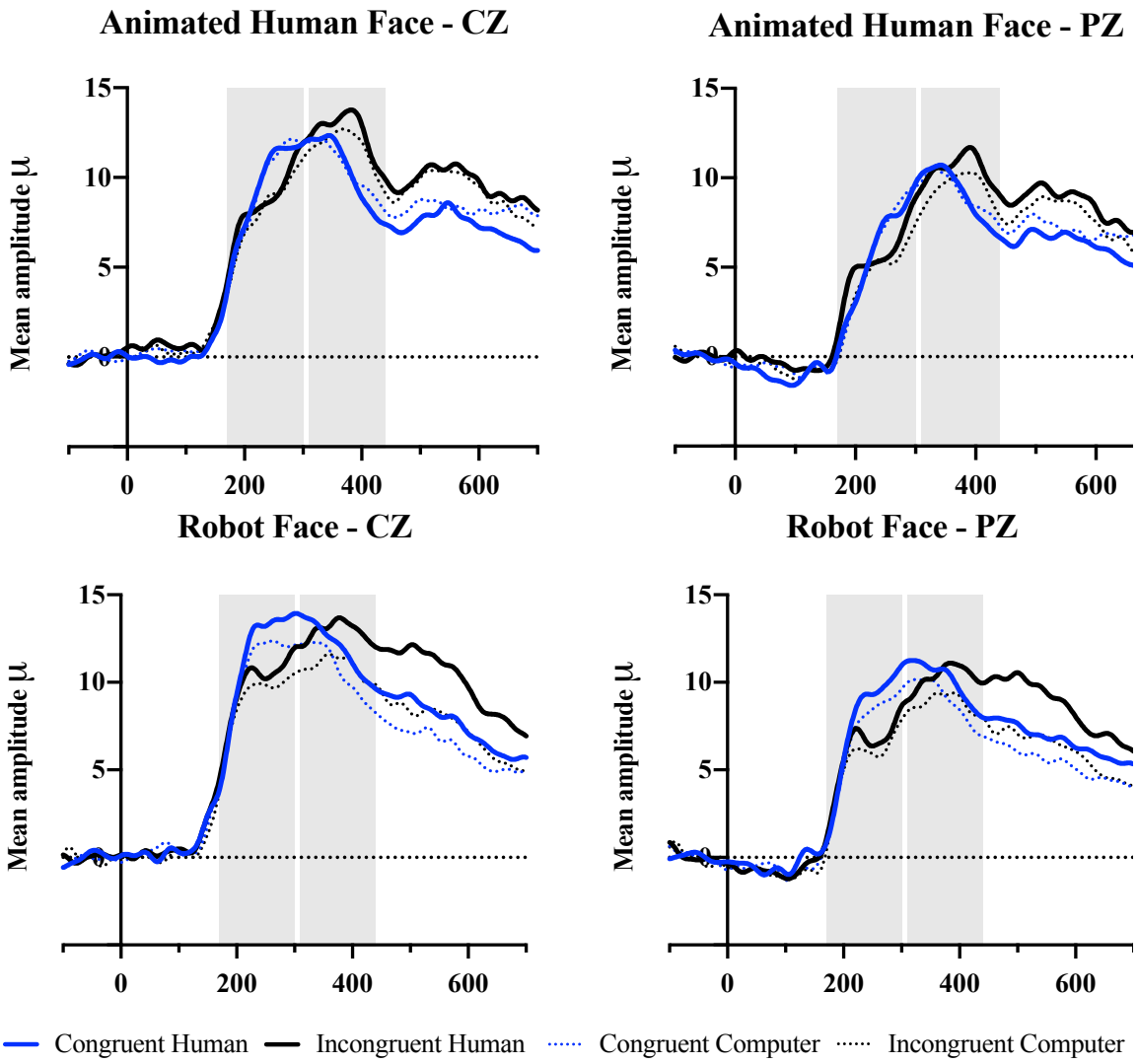


Figure 3. Grand average waveforms for the P250 and P350 at CZ and PZ for both stimulus groups.

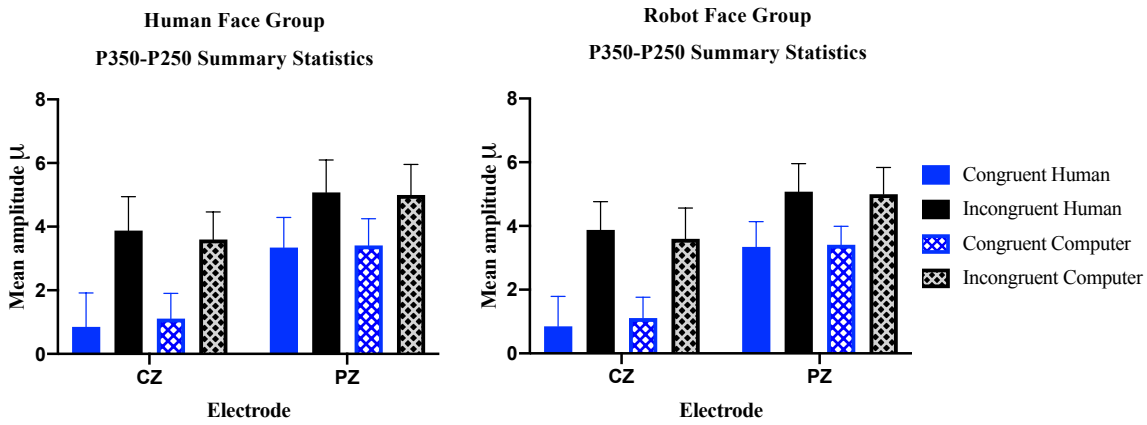


Figure 4. Bar graphs summarizing the difference between P250 and P350 mean amplitudes for each condition at CZ and PZ, for both stimulus groups.

Belief Order

In order to test whether any of the effects of congruency and belief were dependent on the order in which participants adopted an intentional stance, we re-ran 2 (congruency) x 2 (intentional stance belief) ANOVAs for each subgroup (i.e., for those who completed the human-belief task first versus those who completed the computer-belief condition first), for each stimulus group separately.

Human face. For those who first believed they were interacting with a human (n=10) we found no evidence for a congruency effect for the P250 at either electrode [CZ ($F(9)$ = 0.64, p = .444, η^2 = .003); PZ ($F(9)$ = 0.10, p = .758, η^2 < .001)]. However we did find evidence for a congruency by belief interaction for the P250 at PZ ($F(9)$ = 7.04, p = .026, η^2 = .02) but not at CZ ($F(9)$ = 4.79, p = .056, η^2 = .008). Furthermore, there was a a significant main effect of congruency for P350-P250 at CZ ($F(9)$ = 14.54, p = .004, η^2 = .09) but not at PZ ($F(9)$ = 4.74, p = .057, η^2 = .03).

The subgroup who engaged in the computer-belief task first (n=12) exhibited a main effect of congruency on P250 responses at both electrode sites [CZ ($F(11)$ = 12.26, p = .005, η^2 = .02); PZ ($F(11)$ = 6.36, p = .028, η^2 = .03)] and for P350-P250 at CZ ($F(11)$ = 8.08, p = .016, η^2 = .09) but not at PZ ($F(11)$ = 3.71, p = .08, η^2 = .04). To summarise, a belief-

congruency interaction appeared only for the human-avatar subgroup who performed the human-belief condition first. A visual inspection of the waveforms (see Appendix 1) shows that mean P250 amplitudes were largest following congruent gaze shifts during the computer-belief condition.

There was no significant main effect of belief for either the human-belief first [P250: CZ ($F(9) = 0.19$, $p = .678$, $\eta^2 = .001$); PZ ($F(9) = 0.05$, $p = .824$, $\eta^2 = .001$); P350-P250: CZ ($F(9) = 3.55$, $p = .092$, $\eta^2 = .04$); PZ ($F(9) = 1.72$, $p = .223$, $\eta^2 = .02$)] or the computer-belief first subgroups [P250: CZ ($F(11) = 0.77$, $p = .398$, $\eta^2 = .005$); PZ ($F(11) = 2.60$, $p = .135$, $\eta^2 = .02$); P350-P250: CZ ($F(11) = 2.19$, $p = .167$, $\eta^2 = .02$); PZ ($F(11) = 1.78$, $p = .209$, $\eta^2 = .01$)].

Robot face. For those who completed the human-belief condition first (n=10), a significant congruency effect was noted on the P250 at CZ ($F(9) = 9.09$, $p = .012$, $\eta^2 = .09$) and PZ ($F(10) = 13.52$, $p = .005$, $\eta^2 = .12$), and for P350-P250 at CZ ($F(9) = 10.54$, $p = .01$, $\eta^2 = .10$) but not at PZ ($F(9) = 3.62$, $p = .089$, $\eta^2 = .04$). In addition, for this subgroup, a main effect of belief on the P350-P250 was noted at CZ ($F(9) = 8.84$, $p = .016$, $\eta^2 = .06$) but not at PZ ($F(9) = 3.83$, $p = .082$, $\eta^2 = .02$).

For the subgroup who completed the computer-belief task first (n=9), the main effect of belief appeared in the P250 time window at PZ ($F(8) = 11.34$, $p = .01$, $\eta^2 = .03$) but not at CZ ($F(8) = 4.98$, $p = .056$, $\eta^2 = .02$). There was also a significant main effect of congruency on P350-P250 responses at CZ ($F(8) = 6.07$, $p = .039$, $\eta^2 = .09$), but not at PZ ($F(8) = 3.07$, $p = .118$, $\eta^2 = .07$). However, there was no significant congruency effect for the P250 at CZ ($F(8) = 2.94$, $p = .125$, $\eta^2 = .02$) or PZ ($F(8) = 1.94$, $p = .201$, $\eta^2 = .01$).

There was no significant belief-congruency interaction for either the human-belief first subgroup [P250: CZ ($F(9) = 1.02$, $p = .34$, $\eta^2 = .006$); PZ ($F(9) = 1.87$, $p = .205$, $\eta^2 = .008$); P350-P250: CZ ($F(9) = 0.32$, $p = .588$, $\eta^2 = .002$); PZ ($F(9) = 0.14$, $p = .72$, $\eta^2 = .001$)]

or computer-belief first subgroup [P250: CZ ($F(8) = 1.02$, $p = .342$, $\eta^2 = .003$); PZ ($F(8) = 0.97$, $p = .355$, $\eta^2 = .003$); P350-P250: CZ ($F(8) = 1.64$, $p = .236$, $\eta^2 = .01$); PZ ($F(8) = 2.65$, $p = .142$, $\eta^2 = .02$)].

Subjective Interview Scores

To examine whether there were any differences in subjective experience between the human-belief and computer-belief conditions, we compared task ratings collected during the post-experimental interview. Participants provided similar ratings across the human-belief and computer-belief conditions.

Participants interacting with the anthropomorphic avatar rated both conditions as relatively easy [human-belief ($M = 3.24$, $SD = 2.05$); computer-belief ($M = 3.05$, $SD = 1.91$)], but neither natural nor unnatural [human-belief ($M = 5.57$, $SD = 2.27$); computer-belief ($M = 5.38$, $SD = 2.31$)], and neither pleasant nor unpleasant [human-belief ($M = 5.19$, $SD = 1.91$); computer-belief ($M = 5.05$, $SD = 1.86$)]. Subjective ratings did not significantly differ across intentional stance belief conditions (all ps > .239).

Participants interacting with the robot face rated their task as being less difficult than the human-avatar group [human-belief ($M = 2.84$, $SD = 1.77$); computer-belief ($M = 2.79$, $SD = 1.75$)]. This group also rated their task less natural [human-belief ($M = 4.84$, $SD = 1.89$); computer-belief ($M = 4.16$, $SD = 2.04$)] and less pleasant [human-belief ($M = 4.90$, $SD = 0.82$); computer-belief ($M = 4.11$, $SD = 1.63$)]. However, while ratings for the difficulty and naturalness of the task did not significantly differ across intentional stance belief conditions [difficulty ($W = 20$, $p = .832$); felt natural ($W = 37$, $p = .095$)], participants found the task significantly more pleasant during the human-belief condition than during the computer-belief condition ($W = 62$, $p = .009$).

For the human-belief condition, participants across both groups rated the virtual character as moderately cooperative [human avatar ($M = 4.91$, $SD = 2.19$); robot avatar (M

= 5.11, *SD* = 1.10)]. Participants interacting with the human-like avatar stated that the virtual character felt (*M* = 5.29, *SD* =1.88) and behaved (*M* = 5.57, *SD* = 2.11) more human-like than did the group interacting with the robot avatar [felt human (*M* = 4.32, *SD* = 2.34); behaved human (*M* = 4.63, *SD* = 2.31)]. However, none of these differences in ratings were statistically significant between groups (all *p*s > .08).

Unsurprisingly, the group interacting with the computer-generated human face considered the avatar to look more like a human (*M* = 6.24, *SD* = 1.70) than did the robot group (*M* = 3.68, *SD* = 1.73; $t(39) = 4.70$, $p < .001$).

Exploratory Analysis – Dispositional Anthropomorphism

We conducted exploratory analyses to determine whether there was any evidence for a correlation between dispositional anthropomorphism (assessed using the AnthQ; Neave et al., 2015) and the difference between our ERP measures during the computer-belief condition. Caruana and McArthur (2019) observed a difference in the P250 and P350-P250 ERP measures for successful and unsuccessful joint attention in the human-belief condition (i.e., when the avatar was believed to be controlled by a person), but during the computer-belief condition these ERP effects were reduced or absent. Most strikingly, the P250 response following congruent gaze shifts when an intentional stance was adopted was significantly larger than when the gaze shift was incongruent, or when an intentional stance was not adopted.

In the current study, we replicated previous work by Caruana and colleagues (Caruana, de Lissa & McArthur, 2015; 2017; Caruana & McArthur, 2019) with respect to the effect of congruency on ERPs at centro-parietal electrodes. Specifically, we found that both the P250 and P350-P250 were reliably sensitive to joint attention outcomes, with larger P250 responses to congruent than incongruent gaze shifts and the opposite pattern for P350-

P250. However, unlike previous studies, we did not find evidence for a reliable modulation of explicit intentional stance beliefs on these joint attention effects.

With respect to the P250, we found that individuals in the current study exhibited larger responses to congruent than incongruent gaze shifts both when they believed their partner was human- and computer-controlled, whereas this effect was only exclusively observed in the human-belief condition in previous work (Caruana & McArthur, 2019). Given that we noticed a high degree of variability in P250 responses across individuals, we wanted to test whether variability in the congruency effect in the computer-belief condition – when participants did not hold explicit intentional stance beliefs – were associated with individual differences in anthropomorphism tendencies. This follows in that one's disposition to attribute human characteristics to a non-human entity may be considered an *implicit* adoption of an intentional stance, which may have similar neural consequences during the evaluation of joint attention.

Our exploratory analyses revealed that stronger dispositional tendencies to attribute human characteristics to non-human entities, as measured using the AnthQ, was significantly and positively correlated with P250 difference scores for the effect of congruency (i.e., congruent-incongruent) at both CZ ($r(38) = .33$, $p = .041$) and PZ ($r(38) = .34$, $p = .033$). However, we probed this effect in both stimulus groups separately and found that this correlation was only significant in the human face group [CZ ($r(19) = .56$, $p = .008$); PZ ($r(19) = .52$, $p = .016$)], not for the robot face group [CZ ($r(17) = .10$, $p = .672$); PZ ($r(17) = .18$, $p = .462$)]. AnthQ scores were not significantly correlated with P350-P250 congruency effects within the whole sample, for the human face group or the robot face groups (all ps > .229; see Appendix 2).

Discussion

In this study, we examined whether the neural encoding of successful joint attention was modulated by the explicit adoption of an intentional stance towards an interactive virtual partner. We also wanted to test whether this modulation depended on the anthropomorphic realism of the virtual partner's appearance. Specifically, this study was designed to test the reliability of the P250 and P350-P250 ERP effects previously identified by Caruana and McArthur (2019). Caruana and McArthur found that the P250 and P350-P250 responses were sensitive to the observation of gaze shifts which signaled the achievement or avoidance of joint attention – but that this was modulated by explicit intentional stance beliefs. That is, these effects were larger and more reliably observed across individuals when they believed their virtual partner was human- and not computer-controlled. The current study attempted to replicate this finding in a new sample of participants to probe the reliability of these ERPs as neural markers of both joint attention achievement and of adopting an intentional stance towards a virtual partner. We also examined whether the same effects are observed when individuals interact with robot faces that are not prototypically anthropomorphic to examine the reliability of these effects across stimulus type.

The Neural Encoding of Joint Attention Achievement

Our study replicated the congruency effect reported by Caruana and McArthur (2019), so that at the group level, larger P250 mean amplitudes followed congruent than incongruent gaze shifts, and a larger difference between P350 and P250 responses was observed for incongruent than congruent gaze shifts. This is consistent with the effects reported in past studies for people who believed they were interacting with a person in this same paradigm (Caruana, de Lissa & McArthur, 2015, 2017; Caruana & McArthur, 2019) This demonstrates that the P250 and P350-250 ERP measures provide reliable neural markers of joint attention achievement. Moreover, these effects were replicated despite using a new EEG acquisition system, new data processing software (EEGLAB vs SCAN –

different software environments may have different built-in functions or pre-processing steps which could affect how data is processed), and a different experimenter collecting the data and administering the task instructions. Most strikingly, however, the congruency effect was also present regardless of the type of stimulus that participants interacted with (i.e., human vs robot face). This indicates that anthropomorphic appearance does not appear to significantly influence the neural encoding of joint attention success or failure. Overall, we therefore have strong evidence to suggest that the centro-parietal P250 and P350 ERPs (measured by the difference between P250 and P350) provide reliable markers of joint attention evaluation.

The Influence of Explicit Intentional Stance

The current study, however, was particularly interested in the extent to which an explicit intentional stance modulated subjective experiences and the neural encoding of joint attention. Unlike Caruana and McArthur (2019), we did not find significant differences in subjective experience across the intentional stance belief conditions. More critically, and contrary to our hypotheses, the ERP effects of joint attention evaluation (i.e., congruency) were not significantly modulated by explicit intentional stance beliefs, thus failing to replicate the congruency by belief interaction reported by Caruana and McArthur. Unlike Caruana and McArthur's study, which found that the congruency effect on the P250 and P350-P250 ERP measures were largely exclusive to the human-belief condition, we found similar congruency effects under *both* the human- and computer-belief conditions. As such, these data suggest that P250 and P350-P250 measures may not be reliable indicators of adopting an explicit, deception-induced intentional stance. One possible explanation for the reduced reliability of this intentional stance effect in the current study is that some participants may have *implicitly* adopted an intentional stance during the task, even when they *explicitly* believed they were interacting with a computer. A second possible

explanation is that the order in which participants adopted an intentional stance (i.e., human-belief first vs. computer-belief first) impacted the psychological perception of the virtual partner's behavior across belief conditions, adding measurement noise to our ERP analyses. We have conducted two follow-up analyses to explore these possible explanations. We discuss these in turn below.

Implicit intentional stance and dispositional anthropomorphism. We postulated that the observation of the congruency effect under the computer-belief condition (for the group interacting with the human face only) in the current study might be due to some individuals treating their virtual partner as if it had a sentient and intentional mind, despite holding the explicit belief that it is not a human mind. To explore this possibility further, we retrospectively tested whether dispositional anthropomorphism was associated with the extent to which participants exhibited a congruency effect on P250 and P350-P250 measures when they explicitly believed their partner was computer-controlled (i.e., computer-belief condition). We examined individual differences in anthropomorphism because it is a dispositional trait that is highly variable across individuals and is also associated with individual differences in the size of brain structures implicated in mentalising processes, such as the temporoparietal junction (Cullen, Kanai, Bahrami & Rees, 2013). Furthermore, dispositional anthropomorphism should index the likelihood of an individual implicitly adopting an intentional stance towards a non-human entity.

Our exploratory analyses revealed a significant and positive correlation between participants' AnthQ test scores, and the congruency difference scores (i.e., congruent minus incongruent mean amplitudes) during the computer-belief condition (i.e., when participants did *not* adopt an explicit intentional stance belief). That is, individuals with a greater tendency to attribute human characteristics to non-human entities exhibited larger P250 responses to congruent than incongruent gaze shifts in the computer-belief condition. If we

assume that anthropomorphism is a proxy measure for implicitly adopting an intentional stance towards non-human entities, then these data tentatively suggest that variability in the congruency effect in the computer-belief condition is driven, in part, by variability in implicit intentional stance adoption. These findings not only help us understand the factors that impact on the reliability of neural markers of intentional stance and joint attention but also inform on the possible individual differences that should be considered in human-robot interaction research. Furthermore, these findings highlight that there is a need for future work on interactions with artificial agents (e.g., robots) to better distinguish between *implicit* and *explicit* intentional stance – both theoretically and in its operationalisation in research. This will be critical for the intersecting fields of social neuroscience and human-robot interaction (discussed further below). However, given the exploratory nature of these findings, prospective investigations are needed to confirm the role of dispositional anthropomorphism on the neural encoding of joint attention and the validity of this as a proxy measure for implicit intentional stance.

Belief order. A second possible explanation for the observed variability in the influence of intentional stance on the neural encoding of joint attention could be differences across participants in the order in which they adopted an intentional stance (i.e., human-belief first vs computer-belief first). Belief order was counter-balanced in this study to mitigate the influence of order effects observed at the group level. However, it is possible that those who completed the human-belief condition first treated the avatar in the subsequent computer-belief condition in a similar way, such that the effects of intentional stance on neural processing persist into the computer-belief condition. If this were so, we might expect to see larger and more reliable congruency effects on the P250 and P350-P250 ERP measures for individuals who completed the computer-belief condition first than those who completed the human-belief condition first.

As detailed in the methods section above, for each stimulus group (human vs robot face), we performed separate 2 (congruency) x 2 (belief) ANOVAs for each subgroup of individuals who completed the human-belief condition first and those who completed the computer-belief condition first. Contrary to our expectation, our analyses revealed that belief order in most subgroups did not change the pattern of results. The only exception was that those in the human face group who completed the human-belief condition first exhibited a significant belief by congruency interaction on the P250 response. However, this was not in the expected direction. That is, when participants completed the computer-belief condition second, the effect of congruency was larger during the second condition than it was in the first (see Appendix 1 for separate ERP plots by belief order and stimulus group). Together, these follow-up results do not suggest that belief order is reliably contributing to the larger congruency effect observed during computer-belief trials at the group level. Furthermore, since splitting the groups into subgroups by belief order markedly reduces the power of our statistical analyses, future work is needed to prospectively test, with larger samples, whether belief order influences the influence of explicit intentional stance beliefs on the neural encoding of joint attention.

Aesthetic Realism

A novel aim of the current study was to investigate the influence of aesthetic anthropomorphism realism on the neural encoding of joint attention and the modulatory effect of explicitly adopting an intentional stance. We compared neural encoding of joint attention – and the possible modulatory effect of intentional stance – across two samples of individuals who either interacted with an animated human face (from Caruana, de Lissa & McArthur, 2015, 2017; Caruana, Brock & Woolgar, 2015; Caruana & McArthur, 2019) or a humanoid robot face (the Meka S2, from Abubshait & Wiese, 2017).

We found no significant differences in our P250 and P350-P250 measures between stimulus groups for the neural encoding of joint attention. We observed the same congruency effects (i.e., larger P250 mean amplitudes after congruent responses from the avatar, and a larger mean amplitude gain from P250 to P350 intervals after incongruent responses), regardless of whether participants believed they were interacting with a human or a computer. These results suggest that humans flexibly encode eye gaze information from faces, even if a face is not prototypically human. The stability of these joint attention (i.e., gaze congruency) effects across stimulus groups is also striking given that these human and robot face stimuli differed with respect to (1) the physical size of the eyes, (2) the proportional size of the eyes relative to the rest of the face, and (3) the visual contrast generated by sclera visibility when the eyes shifted from direct to averted gaze. To fully account for the influence of these non-social stimulus differences on corresponding neural encoding, it would be ideal for future studies to use human and robot face stimuli that are better matched on low-level visual properties.

The consistent patterns of neural encoding of gaze-signalled joint attention across stimulus groups are comparable with the findings of Wykowska et al. (2014) who reported similar ERP effects on validly gaze-cued targets when participants either believed that a robot face or a human face stimulus was controlled by a person. The authors did however find differences in the neural encoding of gaze-cued targets between human and robot faces when participants were not given any instructions about the agency of the face stimulus. This suggests that the explicit instruction about whether the stimulus is human- or computer-controlled may be critical in standardising, to some extent, whether individuals adopt an intentional stance towards the face. This may override any influence the stimuli's anthropomorphic features may have on implicit intentional stance effects. That is, the aesthetic humanness of a face may only impact our intentional stance towards the face when

we have no other information about the stimulus available to us. Nevertheless, the current study demonstrates that people are willing to engage in social interactions and to use non-verbal cues to communicate with entities that are aesthetically, unambiguously non-human (i.e., robotic), even with the obvious absence of prototypical facial structures (e.g., mouth).

Implications for Human-Robot Interaction Research

Our findings have implications for the field of human-robot interaction, which looks to investigate the robotic features – regarding both function, responsivity and appearance – which impact how humans perceive, feel about, respond to, and perform during interactions with them. One goal of this field is to develop socially-responsive robots which can deliver useful applications for society. Some of these applications may benefit from robots that are perceived as more human-like and intentional (Hameed, Tan, Thomsen & Duan, 2016), such as in education (Fernandes, Fermé & Oliviera, 2006; Saerbeck et al. 2010), for companionship (Dautenhahn, 2007; Breazeal et al. 2004) and for therapeutic interventions (Ferrari, Robins & Dautenhahn, 2009; Robins, Dautenhahn & Dickerson, 2009). However, adopting an intentional stance towards artificial agents can also have negative consequences under certain conditions (e.g., distribution of responsibility; see Hortensius & Cross, 2018 for a review). Furthermore, the aesthetic realism of an agent is also known to influence human-robot relationships, such that more human-like robots tend to generate more empathetic and positive emotional responses from human observers than less aesthetically human-like robots (Bartneck, Kulić, Croft & Zoghbi, 2008). Research into human-robot interaction has determined that whilst robot appearance and behaviour can influence the extent to which people attribute mental states to robots, explicit instruction can override this (Abubshait & Wiese, 2017; Wykowska et al. 2014; for review, see Hortensius & Cross, 2018). The robot used in our study (the Meka S2) resembles a human face, in particular, its salient, human-like eyes. As a result, it is possible that other, less human-like robot faces

may influence the neural encoding of social information, even when observers hold explicit belief about their intentionality. To comprehensively determine whether aesthetic realism impacts the neural encoding of social cues in robots, a broader range of robot stimuli must be examined.

The current findings suggest that the ERPs identified by past studies may not provide reliable neural markers of *explicit* intentional stance adoption (Caruana & McArthur, 2019; Caruana, de Lissa & McArthur, 2015). However, our exploratory analyses suggest that these ERPs may be sensitive to individual differences in the *implicit* adoption of an intentional stance towards an agent, even when there is an explicit belief that it does not have a human-like mind. We attempted to indirectly measure the tendency to implicitly adopt an intentional stance using a self-report measure of dispositional anthropomorphism. We found tentative evidence to suggest that this was related to the degree to which the P250 was modulated by the perception of gaze-signalled joint attention. As such, the P250 may provide a neural marker of implicit intentional stance during social engagement with artificial agents. Prospective studies, with larger sample sizes are needed to (1) confirm the use of dispositional anthropomorphism measures such as the AnthQ as a proxy measure for implicit intentional stance, and (2) to confirm the association between implicit intentional stance adoption and the neural modulation of joint attention encoding as indexed by the centro-parietal P250 response.

If future studies confirm the P250 as a reliable neural marker of implicit intentional stance adoption, this will have significant implications for the field of human-robot interaction. Currently, human-robot interaction research largely relies on subjective measures of intentional stance to assess how robot design impacts human-to-robot interactions. For instance, participants may be asked to rate the likelihood that a robot (of varying aesthetic human-likeness) “has a mind” (Abubshait & Wiese, 2017; Wiese et al.

2018). A less direct measure adopted by de Graaf and Malle (2019) involved examining whether participants made inferences about the minds of a robot when describing its behaviour (i.e., phrases which explained or described the behaviour of a robot with references to projected wants, needs or intentions, such as "the robot ***wanted*** to be polite"). However, subjective responses are inherently biased and unreliable, since they require conscious deliberation. Furthermore, these subjective ratings are often retrospectively obtained from participants (i.e., after the interaction has occurred) and therefore do not directly index intentional stance at the time of observing or interacting with the robot. To this end, a neural marker, such as the P250, may be valuable in providing a direct, quantifiable and objective measure of intentional stance. Further, the potential for the P250 to index *implicit*, rather than *explicit* intentional stance is arguably of greater relevance to human-robot interaction research, given that in this context the aim is to determine whether participants implicitly perceive a mind in a robotic agent that they explicitly believe is not human, and does not have a human mind.

Additional advantages of validating an EEG-based neural marker of intentional stance are that it (1) could be used to assess real-time changes in intentional stance, (2) is relatively low-cost compared to other neurophysiology methods (e.g., fMRI) and (3) is increasingly portable, lending itself to application in various real-world settings. One exciting direction for future research would be to validate the P250 effects further, using portable EEG systems such as the Emotiv EPOC. These EEG systems have been shown to provide reliable ERP measures, when validated against traditional research-grade systems for both auditory (Badcock et al. 2012; 2013; 2015; de Wit et al. 2017) and visually-evoked ERPs (de Lissa, Sörensen, Badcock, Thie & McArthur, 2015). The ability to reliably measure intentional stance effects using these portable systems would enable more ecologically-valid human-robot interaction research, with the capacity to examine how

humans perceive and interact with robots in their intended context for application (e.g., in the classroom or home).

Conclusion

Joint attention is dependent upon adopting an intentional stance towards a social partner. This study investigated ERPs associated with evaluating the success of joint attention with and without the explicit adoption of an intentional stance. The results show that the centro-parietal P250 and P350 ERPs identified by Caruana and McArthur (2019) are reliable markers of joint attention achievement, but are not, in themselves, reliable neural markers of explicit intentional stance. However, we have tentative evidence to suggest that these ERPs may be sensitive to implicit intentional stance adoption during joint attention encoding. Pending future validation of these exploratory findings, these ERPs may prove useful neural markers of implicit intentional stance for use in human-robot interaction research.

Reference List

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*, 1393.

Adamson, L. B., Bakeman, R., Suma, K., & Robins, D. L. (2019). An expanded view of joint attention: Skill, engagement, and language in typical development and autism. *Child Development*, *90*(1).

Badcock, N. A., Mousikou, B., Mahajan, Y., De Lissa, P., Thie, J., & McArthur, G. (2012). Emotiv versus Neuroscan: Validating a gaming EEG system for research quality ERP measurement. In *Front. Hum. Neurosci. Conference Abstract: ACNS-2012 Australasian Cognitive Neuroscience Conference. DOI: 10.3389/conf. fnhum* (Vol. 122).

Badcock, N. A., Mousikou, P., Mahajan, Y., De Lissa, P., Thie, J., & McArthur, G. (2013). Validation of the Emotiv EPOC® EEG gaming system for measuring research quality auditory ERPs. *PeerJ*, *1*, e38.

Badcock, N. A., Preece, K. A., de Wit, B., Glenn, K., Fieder, N., Thie, J., & McArthur, G. (2015). Validation of the Emotiv EPOC EEG system for research quality auditory event-related potentials in children. *PeerJ*, *3*, e907.

Banakou, D., Hanumanthu, P. D., & Slater, M. (2016). Virtual embodiment of white people in a black virtual body leads to a sustained reduction in their implicit racial bias. *Frontiers in Human Neuroscience*, *10*, 601.

Baron-Cohen, S. (1997). *Mindblindness: An Essay on Autism and Theory of Mind.* Cambridge, MA: MIT press.

Bartneck, C., Kulić, D., Croft, E., & Zoghbi, S. (2009). Measurement instruments for the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots. *International Journal of Social Robotics*, *1*(1), 71-81.

Bohil, C. J., Alicea, B., & Biocca, F. A. (2011). Virtual reality in neuroscience research and therapy. *Nature Reviews Neuroscience*, *12*(12), 752.

Booth, T., Murray, A. L., McKenzie, K., Kuenssberg, R., O'Donnell, M., & Burnett, H. (2013). Brief report: An evaluation of the AQ-10 as a brief screening instrument for ASD in adults. *Journal of Autism and Developmental Disorders*, *43*(12), 2997-3000.

Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems*, *42*(3-4), 167-175.

Breazeal, C., Brooks, A., Gray, J., Hoffman, G., Kidd, C., Lee, H., ... & Mulanda, D. (2004). Humanoid robots as cooperative partners for people. *Int. Journal of Humanoid Robots*, *1*(2), 1-34.

Bruinsma, Y., Koegel, R. L., & Koegel, L. K. (2004). Joint attention and children with autism: A review of the literature. *Mental Retardation and Developmental Disabilities Research Reviews*, *10*(3), 169-175.

Bruner, J. S. (1974). From communication to language—A psychological perspective. *Cognition*, 3(3), 255-287.

Cañigueral, R., & Hamilton, A. F. D. C. (2019). The Role of Eye Gaze During Natural Social Interactions in Typical and Autistic People. *Frontiers in Psychology*, *10*.

Carlin, J. D., & Calder, A. J. (2013). The neural basis of eye gaze processing. *Current Opinion in Neurobiology*, *23*(3), 450-455.

Carrick, O. K., Thompson, J. C., Epling, J. A., & Puce, A. (2007). It's all in the eyes: neural responses to socially significant gaze shifts. *Neuroreport*, *18*(8), 763.

Carter, E. J., & Pelphrey, K. A. (2008). Friend or foe? Brain systems involved in the perception of dynamic signals of menacing and friendly social approaches. *Social Neuroscience*, *3*(2), 151-163.

Caruana, N., Brock, J., & Woolgar, A. (2015). A frontotemporoparietal network common to initiating and responding to joint attention bids. *Neuroimage*, *108*, 34-46.

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.

Caruana, N., de Lissa, P., & McArthur, G. (2017). Beliefs about human agency influence the neural processing of gaze during joint attention. *Social Neuroscience*, *12*(2), 194-206.

Caruana, N., & McArthur, G. (2019). The mind minds minds: The effect of intentional stance on the neural encoding of joint attention. *Cognitive, Affective, & Behavioral Neuroscience*, 1-13.

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.

Charman, T., Baron-Cohen, S., Swettenham, J., Baird, G., Drew, A., & Cox, A. (2003). Predicting language outcome in infants with autism and pervasive developmental disorder. *International Journal of Language & Communication Disorders*, *38*(3), 265-285.

Cross, E. S., Hortensius, R., & Wykowska, A. (2019). From social brains to social robots: applying neurocognitive insights to human–robot interaction. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 374(1771), 20180024. https://doi.org/10.1098/rstb.2018.0024*

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.

Dautenhahn, K. (2007). Socially intelligent robots: dimensions of human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 679-704.

Dawson, G., Toth, K., Abbott, R., Osterling, J., Munson, J., Estes, A., & Liaw, J. (2004). Early social attention impairments in autism: social orienting, joint attention, and attention to distress. *Developmental Psychology*, *40*(2), 271.

de Graaf, M. M., & Malle, B. F. (2019). People's Explanations of Robot Behavior Subtly Reveal Mental State Inferences. In *2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (pp. 239-248). IEEE.

de Lissa, P., Sörensen, S., Badcock, N., Thie, J., & McArthur, G. (2015). Measuring the face-sensitive N170 with a gaming EEG system: a validation study. *Journal of Neuroscience Methods*, *253*, 47-54.

De Wit, B., Badcock, N. A., Grootswagers, T., Hardwick, K., Teichmann, L., Wehrman, J., ... & Kaplan, D. M. (2017). Neurogaming technology meets neuroscience education: a cost-effective, scalable, and highly portable undergraduate teaching laboratory for neuroscience. *Journal of Undergraduate Neuroscience Education*, *15*(2), A104.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9-21.

Dennett, D. C. (1989). *The Intentional Stance*. Cambridge, MA: MIT press.

Fernandes, E., Fermé, E., & Oliveira, R. (2006). Using robots to learn functions in math class. *Technology Revisited*, 152.

Ferrari, E., Robins, B., & Dautenhahn, K. (2009). Therapeutic and educational objectives in robot assisted play for children with autism. In *RO-MAN 2009-The 18th IEEE International Symposium on Robot and Human Interactive Communication* (pp. 108-114). IEEE.

Frith, C. D., & Frith, U. (1999). Interacting minds--a biological basis. *Science*, *286*(5445), 1692-1695.

Frith, C., & Frith, U. (2000). The physiological basis of theory of mind: functional neuroimaging studies. *Understanding Other Minds: Perspectives from Developmental Cognitive Neuroscience*, *2*.

Gobel, M. S., Kim, H. S., & Richardson, D. C. (2015). The dual function of social gaze. *Cognition*, *136*, 359-364.

Georgescu, A. L., Kuzmanovic, B., Roth, D., Bente, G., & Vogeley, K. (2014). The use of virtual characters to assess and train non-verbal communication in high-functioning autism. *Frontiers in Human Neuroscience*, *8*, 807.

Gordon, I., Eilbott, J. A., Feldman, R., Pelphrey, K. A., & Vander Wyk, B. C. (2013). Social, reward, and attention brain networks are involved when online bids for joint attention are met with congruent versus incongruent responses. *Social Neuroscience*, *8*(6), 544-554.

Hameed, I. A., Tan, Z. H., Thomsen, N. B., & Duan, X. (2016). User acceptance of social robots. In *Proceedings of the Ninth International Conference on Advances in Computer-Human Interactions (ACHI 2016)*, Venice, Italy (pp. 274-279).

Heatherton, T. F., Macrae, C. N., & Kelley, W. M. (2004). What the social brain sciences can tell us about the self. *Current Directions in Psychological Science*, *13*(5), 190-193.

Heatherton, T. F., Wyland, C. L., Macrae, C. N., Demos, K. E., Denny, B. T., & Kelley, W. M. (2006). Medial prefrontal activity differentiates self from close others. *Social Cognitive and Affective Neuroscience*, *1*(1), 18-25.

Hortensius, R., & Cross, E. S. (2018). From automata to animate beings: the scope and limits of attributing socialness to artificial agents. *Annals of the New York Academy of Sciences*, *1426*(1), 93-110.

Izuma, K., Saito, D. N., & Sadato, N. (2008). Processing of social and monetary rewards in the human striatum. *Neuron*, *58*(2), 284-294.

Kaplan, F., & Hafner, V. V. (2006). The challenges of joint attention. *Interaction Studies*, *7*(2), 135-169.

Kassner, M. P., Wesselmann, E. D., Law, A. T., & Williams, K. D. (2012). Virtually ostracized: Studying ostracism in immersive virtual environments. *Cyberpsychology, Behavior, and Social Networking*, *15*(8), 399-403.

Lachat, F., Hugueville, L., Lemaréchal, J., Conty, L. & George, N. (2012). Oscillatory brain correlates of live joint attention: a dual-EEG study. *Frontiers in Human Neuroscience*, *6*, 156.

Liu, X., Powell, D. K., Wang, H., Gold, B. T., Corbly, C. R., & Joseph, J. E. (2007). Functional dissociation in frontal and striatal areas for processing of positive and negative reward information. *Journal of Neuroscience*, *27*(17), 4587-4597.

Mundy, P. (2018). A review of joint attention and social-cognitive brain systems in typical development and autism spectrum disorder. *European Journal of Neuroscience*, *47*(6), 497- 514.

Neave, N., Jackson, R., Saxton, T., & Hönekopp, J. (2015). The influence of anthropomorphic tendencies on human hoarding behaviours. *Personality and Individual Differences*, *72*, 214-219.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97-113.

Peck, T. C., Seinfeld, S., Aglioti, S. M., & Slater, M. (2013). Putting yourself in the skin of a black avatar reduces implicit racial bias. *Consciousness and Cognition*, *22*(3), 779-787.

Pelphrey, K. A., Morris, J. P., & McCarthy, G. (2005). Neural basis of eye gaze processing deficits in autism. *Brain*, *128*(5), 1038-1048.

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.

Pfeiffer, U. J., Vogeley, K., & Schilbach, L. (2013). From gaze cueing to dual eye-tracking: novel approaches to investigate the neural correlates of gaze in social interaction. *Neuroscience & Biobehavioral Reviews*, *37*(10), 2516-2528.

Pineda, J. A., & Hecht, E. (2009). Mirroring and mu rhythm involvement in social cognition: are there dissociable subcomponents of theory of mind? *Biological Psychology, 80*(3), 306-314.

Posner, M. I. (1980). Orienting of attention. *Quarterly Journal of Experimental Psychology*, *32*(1), 3-25.

Redcay, E., Dodell-Feder, D., Pearrow, M. J., Mavros, P. L., Kleiner, M., Gabrieli, J. D., & Saxe, R. (2010). Live face-to-face interaction during fMRI: a new tool for social cognitive neuroscience. *Neuroimage*, *50*(4), 1639-1647.

Redcay, E., Kleiner, M., & Saxe, R. (2012). Look at this: the neural correlates of initiating and responding to bids for joint attention. *Frontiers in Human Neuroscience*, *6*, 169.

Robins, B., Dautenhahn, K., & Dickerson, P. (2009, February). From isolation to communication: a case study evaluation of robot assisted play for children with autism with a minimally expressive humanoid robot. In *2009 Second International Conferences on Advances in Computer-Human Interactions* (pp. 205-211). IEEE.

Saerbeck, M., Schut, T., Bartneck, C., & Janse, M. D. (2010). Expressive robots in education: varying the degree of social supportive behavior of a robotic tutor. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1613-1622). ACM.

Saito, D. N., Tanabe, H. C., Izuma, K., Hayashi, M. J., Morito, Y., Komeda, H., ... & Sadato, N. (2010). "Stay tuned": inter-individual neural synchronization during mutual gaze and joint attention. *Frontiers in Integrative Neuroscience, 4*, 127.

Schilbach, L., Timmermans, B., Reddy, V., Costall, A., Bente, G., Schlicht, T., & Vogeley, K. (2013). Toward a second-person neuroscience 1. *Behavioral and Brain Sciences*, *36*(4), 393-414.

Schilbach, L., Wilms, M., Eickhoff, S. B., Romanzetti, S., Tepest, R., Bente, G., ... & Vogeley, K. (2010). Minds made for sharing: initiating joint attention recruits reward-related neurocircuitry. *Journal of Cognitive Neuroscience, 22*(12), 2702-2715.

Schindler, S., & Kissler, J. (2016). People matter: Perceived sender identity modulates cerebral processing of socio-emotional language feedback. *NeuroImage, 134,* 160–169

Schindler, S., & Kissler, J. (2018). Language-based social feedback processing with randomized "senders": an ERP study. *Social Neuroscience*, *13*(2), 202-213.

Schindler, S., Kruse, O., Stark, R., & Kissler, J. (2019). Attributed social context and emotional content recruit frontal and limbic brain regions during virtual feedback processing. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(2), 239-252.

Schindler, S., Wegrzyn, M., Steppacher, I., & Kissler, J. (2014). It's all in your head–how anticipating evaluation affects the processing of emotional trait adjectives. *Frontiers in Psychology*, *5*, 1292.

Schindler, S., Wegrzyn, M., Steppacher, I., & Kissler, J. (2015). Perceived communicative context and emotional content amplify visual word processing in the fusiform gyrus. *Journal of Neuroscience*, *35*(15), 6010-6019.

Severson, R. L., & Carlson, S. M. (2010). Behaving as or behaving as if? Children's conceptions of personified robots and the emergence of a new ontological category. *Neural Networks*, *23*(8-9), 1099-1103.

Slater, M., & Sanchez-Vives, M. V. (2014). Transcending the self in immersive virtual reality. *Computer*, *47*(7), 24-30.

Tomasello, M. (1995). Joint attention as social cognition. *Joint attention: Its Origins and Role in Development*, *103130*.

Van Overwalle, F., & Baetens, K. (2009). Understanding others' actions and goals by mirror and mentalizing systems: a meta-analysis. *Neuroimage*, *48*(3), 564-584.

Wesselmann, E. D., Wirth, J. H., Mroczek, D. K., & Williams, K. D. (2012). Dial a feeling: Detecting moderation of affect decline during ostracism. *Personality and Individual Differences*, *53*(5), 580-586.

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.

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.

Williams, K. D. (2007). Ostracism: The kiss of social death. *Social and Personality Psychology Compass*, *1*(1), 236-247.

Williams, J. H., Waiter, G. D., Perra, O., Perrett, D. I., & Whiten, A. (2005). An fMRI study of joint attention experience. *Neuroimage*, *25*(1), 133-140.

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

Wykowska, A., Chaminade, T., & Cheng, G. (2016). Embodied artificial agents for understanding human social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1693), 20150375.

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).

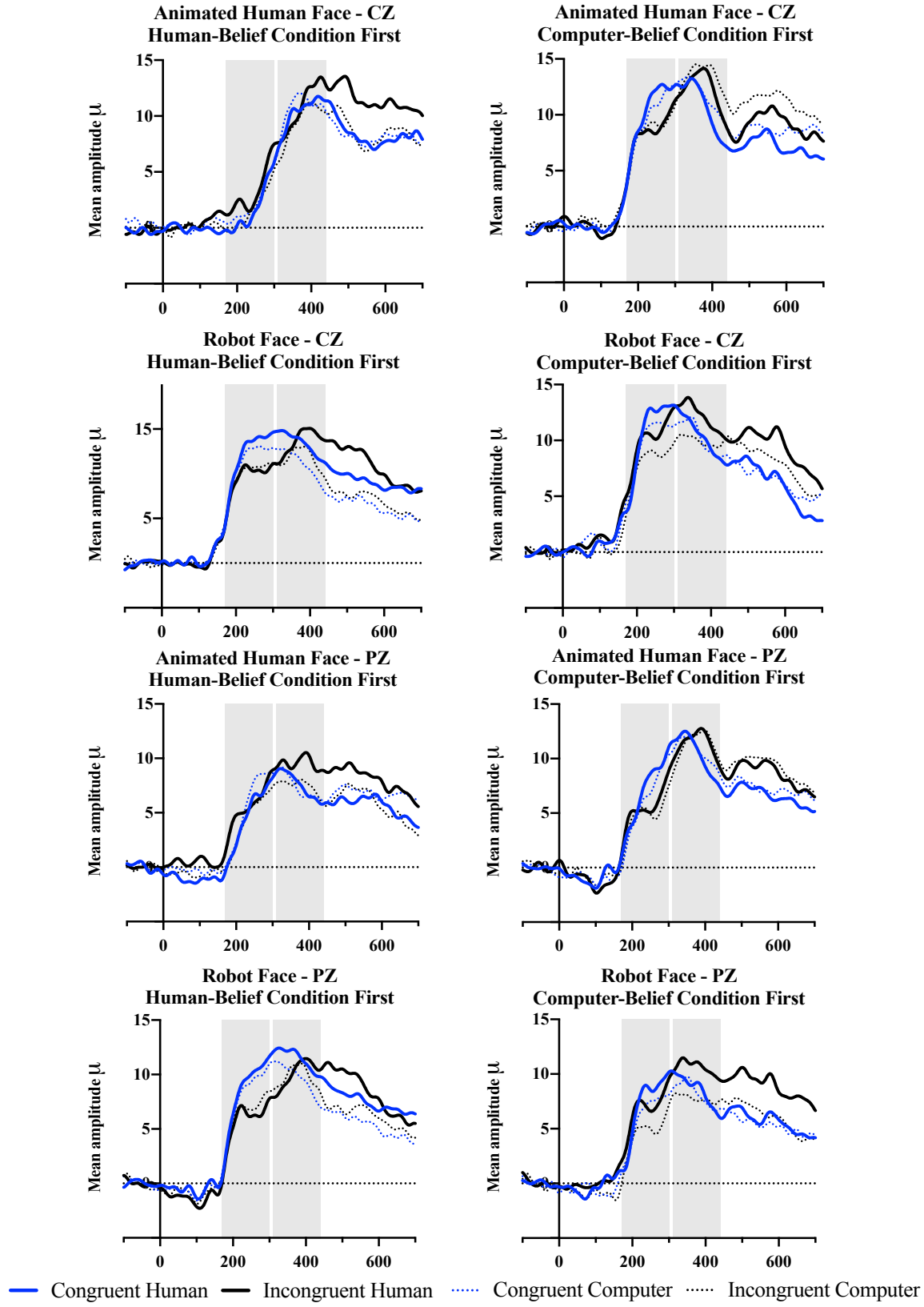
Yu, C., & Smith, L. B. (2013). Joint attention without gaze following: Human infants and their parents coordinate visual attention to objects through eye-hand coordination. *PloS One*, *8*(11), e79659.

Yu, C., & Smith, L. B. (2017). Hand–eye coordination predicts joint attention. *Child Development*, *88*(6), 2060-2078.

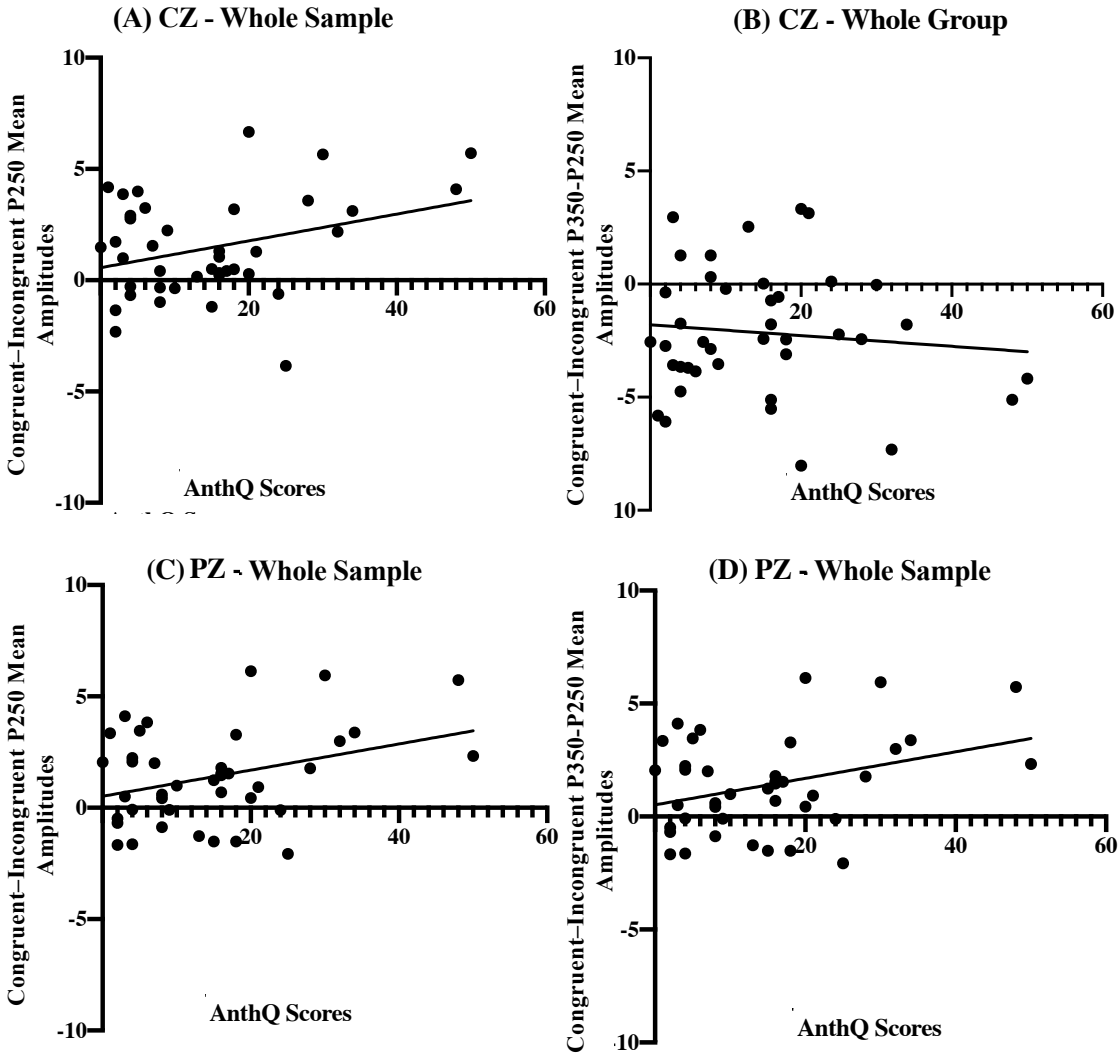
Yu, C., & Smith, L. B. (2017). Multiple sensory-motor pathways lead to coordinated visual attention. *Cognitive Science*, *41*, 5-31.

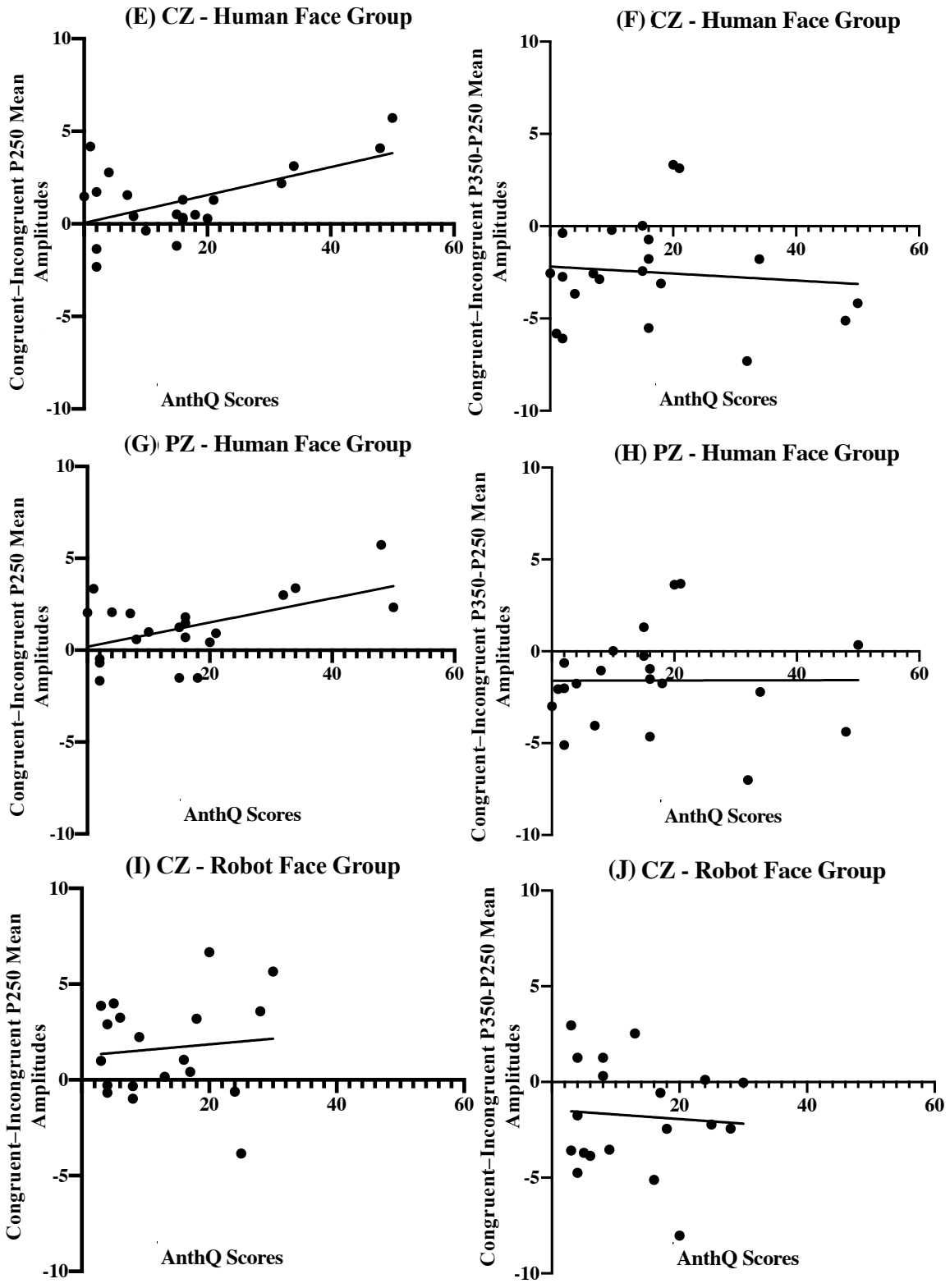
Appendix

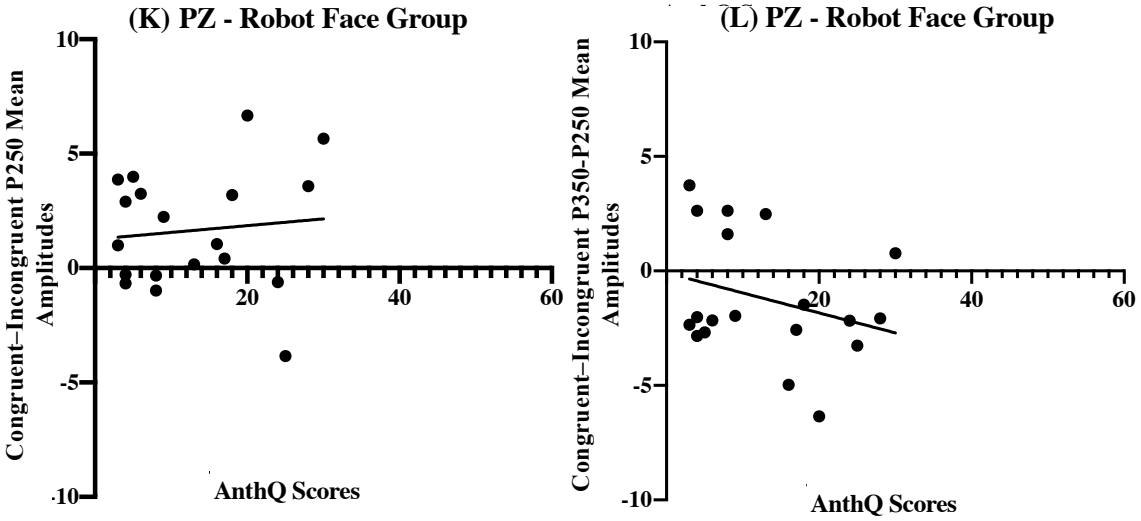
Appendix 1. Group average waveforms for the human-belief first subgroups (left column) and computer-belief first subgroups (right column).



Appendix 2. Correlation plots for correlational analyses between participants' AnthQ scores and the difference between the ERP measures following congruent and incongruent responses during the computer-belief condition. (A) – (D) Correlation plots for the whole sample. (E) – (H) Correlation plots for the human face group. (I) – (L) Correlation plots for the robot group. Note that graphs (A), (C), (E) and (G) represent significant, positive correlations between AnthQ scores and the difference between P250 ERPs following congruent and incongruent gaze shifts during the computer-belief condition.




(E) CZ - Human Face Group
Congruent–Incongruent P250 Mean Amplitudes
AnthQ Scores
(F) CZ - Human Face Group
Congruent–Incongruent P350-P250 Mean Amplitudes
AnthQ Scores
(G) PZ - Human Face Group
Congruent–Incongruent P250 Mean Amplitudes
AnthQ Scores
(H) PZ - Human Face Group
Congruent–Incongruent P350-P250 Mean Amplitudes
AnthQ Scores
(I) CZ - Robot Face Group
Congruent–Incongruent P250 Mean Amplitudes
AnthQ Scores
(J) CZ - Robot Face Group
Congruent–Incongruent P350-P250 Mean Amplitudes
AnthQ Scores


(K) PZ - Robot Face Group
Congruent–Incongruent P250 Mean Amplitudes
AnthQ Scores
10
5
0
-5
-10
20
40
60
(L) PZ - Robot Face Group
Congruent–Incongruent P350-P250 Mean Amplitudes
AnthQ Scores
10
5
0
-5
-10
20
40
60