Cognitive, Affective, & Behavioral Neuroscience
https://doi.org/10.3758/s13415-019-00734-y

The mind minds minds: The effect of intentional stance on the neural encoding of joint attention

Nathan Caruana [1,2] · Genevieve McArthur [1,2]

© The Psychonomic Society, Inc. 2019

Abstract
Recent neuroimaging studies have observed that the neural processing of social cues from a virtual reality character appears to be affected by "intentional stance" (i.e., attributing mental states, agency, and "humanness"). However, this effect could also be explained by individual differences or perceptual effects resulting from the design of these studies. The current study used a new design that measured centro-parietal P250, P350, and N170 event-related potentials (ERPs) in 20 healthy adults while they initiated gaze-related joint attention with a virtual character ("Alan") in two conditions. In one condition, they were told that Alan was controlled by a human; in the other, they were told that he was controlled by a computer. When participants believed Alan was human, his congruent gaze shifts, which resulted in joint attention, generated significantly larger P250 ERPs than his incongruent gaze shifts. In contrast, his incongruent gaze shifts triggered significantly larger increases in P350 ERPs than his congruent gaze shifts. These findings support previous studies suggesting that intentional stance affects the neural processing of social cues from a virtual character. The outcomes also suggest the use of the P250 and P350 ERPs as objective indices of social engagement during the design of socially approachable robots and virtual agents.

Keywords Joint attention · Eye gaze · Virtual reality · Social interaction · Agency

Introduction

Our ability to use non-verbal cues such as eye gaze to communicate and coordinate with others allows us to navigate the many social encounters that fill our daily lives. Eyes provide a special mode of social communication as they are the only sensory organ with the dual function of both signalling and perceiving communicative cues during coordinated face-to-face interactions (Gobel, Kim & Richardson, 2015). As such, eye gaze can intentionally or unintentionally signal information about a person's current mental state, perspective, and intentions. Humans have a unique sensitivity for detecting and using this information during social interactions (Grossmann, 2017).

A paradigmatic example of gaze-based social interaction is joint attention, which is the ability of two individuals to coordinate their attention with each other so that they are attending to the same thing (Bruinsma, Koegel, & Koegel, 2004). This core social skill is believed to be a precursor for language and social cognition development (e.g., Charman, 2003; Dawson et al., 2004; Mundy & Newell, 2007). Previous work has established that the successful achievement of joint attention recruits regions in the social brain network associated with both mentalizing processes (i.e., our ability to understand and infer the mental states of others; Williams, Waiter, Perra, Perrett, & Whiten, 2005) and reward processing (Gordon, Eilbott, Feldman, Pelphrey, & Vander Wyk, 2013; Pfeiffer, et al., 2014). Divergent joint attention development – often seen in autism – can make social interactions confusing and stressful, which can limit opportunities for social learning (Pelphrey, Shultz, Hudac, & Vander Wyk, 2011; Mundy & Newell, 2007). It has therefore become a priority for social cognition and neuroscience research to elucidate the neurocognitive mechanisms that support social interactions.

A challenge facing this research is the creation of experimental measures that offer both experimental control and

✉ Nathan Caruana
nathan.caruana@mq.edu.au

Genevieve McArthur
genevieve.mcarthur@mq.edu.au

[1] Department of Cognitive Science, Macquarie University, Level 3, 16 University Avenue, Sydney, NSW 2109, Australia

[2] ARC Centre of Excellence in Cognition and its Disorders, Sydney, NSW, Australia

Published online: 26 June 2019

ecological validity. This is particularly difficult for neurophysiological studies using equipment that can only test one person at a time (e.g., MRI scanners; see Caruana, McArthur, Woolgar, & Brock, 2017b; Schilbach et al., 2013, for relevant reviews). Some research studies have tackled this problem using displays of a virtual reality character who achieves or avoids joint attention with a participant whilst their brain is being scanned or measured (Caruana, de Lissa, & McArthur, 2015b, 2017a; Caruana, Brock, & Woolgar, 2015a; Schilbach et al., 2010; Pfeifer et al., 2014; Wilms et al., 2010; also see Georgescu, Kuzmanovic, Roth, Bente, & Vogeley, 2014, for a relevant review).

For the sakes of ecological validity, many of the above-mentioned studies went to considerable lengths to deceive participants into believing that their virtual partner was controlled by another human. This encouraged participants to adopt an "intentional stance" towards their virtual partner, believing that it had a mind of its own, and hence had intentions, desires, and motivation (i.e., "*My partner is capable and willing to achieve joint attention with me*"). Intentional stance is thought to be critical to the measurement of joint attention, which relies on (1) our ability to infer and evaluate others' behavior as a function of their beliefs, desires, and goals ("mentalizing"; Premack & Woodruff, 1978); and (2) our acceptance that a social partner is a sentient being capable of sharing attention and mental states (Emery et al., 2000). When we believe an entity has a mind (i.e., adopt an intentional stance), we automatically engage mentalizing mechanisms that have a top-down influence on the neurocognitive processing of social information (Wykowska et al., 2014).

Several studies have aimed to measure the influence of intentional stance on the neural mechanisms engaged during gaze-based social interactions. In a recent fMRI study by Wiese, Buzzell, Abubshait, and Beatty (2018), participants viewed morphed face stimuli that varied from appearing "more human" to "more robotic" along a 6-point continuum. Participants judged how likely a face was to have "internal states." Increased activation in the ventromedial prefrontal cortex – a region associated with mentalizing processes – was greater for faces judged to be more human than more robotic. This differential activation was associated with larger gaze-cueing effects in a behavioral task conducted outside the scanner using the same stimuli (i.e., faster detection of targets preceded by a valid than an invalid gaze cue). The authors interpreted the findings as evidence for greater relevance assigned to social cues from agents eliciting an intentional stance.

In a similar study using event-related potentials (ERPs), Schindler and colleagues (2017) found linear increases in the "late positive potential" (LPP) response to six face stimuli that increased in face-realism from photographs to stylized cartoons. Larger responses were observed for more realistic faces, with increased activity measured at visual and centroparietal sites. The LPP is believed to reflect higher-order stages of face evaluation, including the labelling of expressed emotions (see Schupp, Flaisch, Stockburger, & Junghofer, 2006; Hajcak, MacNamara, & Olvet, 2010, for reviews). Whilst it is possible that these ERP effects may reflect the attribution of internal states, and therefore the representation of emotional mental states, this study did not directly aim to measure or manipulate intentional stance.

Schindler and Kissler (2016) conducted another study where they did directly manipulate participants' intentional stance using a written verbal feedback paradigm. Participants were initially asked to describe themselves in a semi-structured interview whilst they were video-recorded. They were told that this video would be shown to another person whom they would be interacting with during the experiment. In one block of trials, participants were told that they would interact with a human partner, and in another, a "socially intelligent" computer interface. During each type of interaction, participants were shown a mixture of positive, neutral, and negative adjectives presented as text on a screen, each of which was followed by a color cue that indicated whether the human or computer partner thought the adjective fitted the description of the participant, based on their video. Participants' mean centroparietal P2, P3, and LPP ERPs were larger to feedback from a supposed human partner than a computer partner. The authors interpreted these findings in line with the "motivated attention" framework, in which stimuli that are intrinsically relevant to an individual (e.g., because they represent another person) are subjected to prioritized perceptual processing (Schupp et al., 2003, 2004).

Whilst the centroparietal P3 has been widely investigated in various contexts, Schildler and Kissler's (2016) findings are the first to associate the socially evoked P2 with intentional stance modulations. However, the P3 has been previously implicated in the allocation of attention to unexpected, novel, and rare events in non-social contexts (Polich, 2007), and is believed to reflect an observer's judgment of the probability of a stimulus occurring (Donchin & Coles, 1988). Similarly, the P3 has been thought to reflect the "motivational significance" of an eliciting stimulus, which in turn modulates attention allocation and subsequent perceptual processing (Nieuwenhuis et al., 2005). The motivational significance of a stimulus may relate to its personal relevance or utility (e.g., Yeung & Stanfey, 2004). Interpreted together, these findings suggest that the P3 may reflect the prioritized neural processing of social stimuli that are believed to be personally relevant to an observer. Thus, social information believed to convey the perspectives of other humans, who may judge, collaborate, or harm us, may be prioritized over social information from artificial agents, even if they are believed to be "intelligent" (Schindler et al., 2015; Schindler & Kissler, 2016).

However, the studies reviewed thus far have not investigated the influence of intentional stance on the neural processing of social information during reciprocal interactions.

Pfeiffer et al. (2014) took an interactive approach in an fMRI study of gaze-based joint attention. Participants were told that over a series of five trial blocks, they would initiate joint-attention bids with a virtual character. This character would be controlled by a human in some blocks and by a computer in others. The proportion of trials in which joint attention was achieved was systematically manipulated across blocks (i.e., how often the virtual character looked at the location indicated by the participant). At the end of each block, participants decided whether they were interacting with a human or a computer. The results showed that participants were most likely to believe that they had interacted with a human in blocks where most trials resulted in joint attention. Such blocks were associated with greater neural activation in the ventral striatum – a region associated with social-reward processing – than blocks in which the character was thought to be controlled by a computer.

A limitation of the aforementioned studies is that they manipulated intentional stance indirectly by varying the aesthetics of the stimuli (e.g., more human vs. more robot/caricatured; Schindler et al., 2017; Weise et al., 2018) or the behavior of the stimuli (e.g., achieving or avoiding joint attention; Pfeiffer et al., 2014). Such approaches cannot rule out the possibility that differences in neural responses between conditions arose from differences in the features or behavior of the stimuli, rather than differences in intentional stance. To our knowledge, only three studies of gaze-based communication have directly manipulated intentional stance between conditions without manipulating the features or behavior of the agent (Caruana, de Lissa, & McArthur, 2015b, 2017a; Wykowska et al. 2014).

Caruana, de Lissa, and McArthur (2015b) used a similar design to Pfeiffer et al. (2014). Participants initiated a joint attention bid with a virtual character to one of four on-screen locations. The virtual partner responded congruently or incongruently, with equal probability, to achieve or avoid joint attention, respectively. ERPs time-locked to the virtual partner's gaze shift were measured at occipitotemporal and centroparietal sites, consistent with previous gaze-processing ERP studies (see Carrick et al., 2007). One group of participants completed the task under the instruction that the virtual partner's gaze was controlled by another human in a nearby eye-tracking laboratory. This group had larger and later centroparietal P350 ERPs to incongruent gaze shifts compared to congruent gaze shifts. This effect was not observed in a control group of participants who completed a non-social version of the same task in which arrow cues – believed to be controlled by a computer – were superimposed over the avatar's face with closed eyes. Instead of a P350, this group showed a clear centro-parietal P250 ERP to arrow cues that did not vary with congruency. A retrospective analysis of the P250 data (not reported in the original paper, but using the same analysis protocol reported below) revealed a group-by-condition interaction, in which larger P250 ERPs were observed in response to congruent gaze shifts than incongruent gaze shifts when participants adopted an intentional stance. A reduced effect of congruency of the P250 response was observed for individuals who observed computer-controlled arrows. Together, these data suggested that the centro-parietal P250 and P350 may be sensitive to the role of intentional stance during joint attention.

In a second study, Caruana, de Lissa, and McArthur (2017a) examined a third group of participants who completed the social version of the task above, but were told that their virtual partner's gaze was controlled by a computer. In this group, the previously observed P350 effect of gaze congruency was less reliable at the individual level and was not statistically significant at the group level. Combined with the outcomes of Caruana et al. (2015b), these data suggest that the centroparietal P350 signals the achievement or avoidance of joint attention under conditions where an individual adopts an intentional stance with a virtual social partner.

A key limitation of the studies by Caruana and colleagues was the use of between-subjects designs to directly manipulate if participants adopted an intentional stance. Such designs cannot discount the possibility that condition effects were due to group differences in neuroanatomical structure, social information processing, or the tendency to anthropomorphize (see Waytz, Cacioppo, & Epley, 2010), rather than differences in intentional stance per se. What is needed, therefore, is a within-subjects experiment that measures the ERPs of the same subjects in two conditions that directly manipulate intentional stance. Such an experiment is difficult as it needs to provide a deceptive cover story that can convince participants that their virtual partner is sometimes controlled by a human and sometimes by a computer, despite no change whatsoever in the virtual partner's appearance or behavior.

To our knowledge, no study has successfully manipulated intentional stance during dynamic gaze-based social interactions using a within-subjects design. However, Wykowska et al. (2014) managed such a manipulation using a non-dynamic gaze-cueing paradigm. Participants completed a gaze-cueing task using a humanoid robot face stimulus. Participants were asked to report a target stimulus (T or F) that appeared on the right or left side of the robot. Just before a letter was presented, the robot gazed at the correct location of the letter (valid cue) or the opposite location (invalid cue). Two belief conditions were counter-balanced between participants: (1) the robot's gaze was controlled by a pre-programmed computer algorithm, or (2) by a human. Participants identified targets faster on valid than on invalid trials when they believed that the robot was controlled by a

human than by a computer. The same advantage could be seen in occopitotemporal P1 ERP responses that were time-locked to the presentation of the target.

The study by Wykowska et al. (2014) is the first to use a within-subjects design to measure the effect of intentional stance on the neural encoding of gaze-cued information during a non-interactive task. The current study takes the next step by measuring the effect of intentional stance on the encoding of gaze itself, during an interactive task that involves joint attention. Based on our previous work (Caruana et al., 2017a), we predicted that participants would have larger P250 ERPs and smaller P350 ERPs in trials where they achieved joint attention with a virtual character than in trials where they did not. Critically, these effects would be reduced or absent when participants did not adopt an intentional stance towards the virtual agent.

Method

The methods used in this study were approved by the Macquarie University Human Research Ethics Committee (Ref# 5201200021).

Participants

This study used a within-subjects design in which participants responded to two conditions of stimuli ("congruent" vs. "incongruent") and two conditions of belief ("human" vs. "computer"). Twenty-three participants (18 female) volunteered for the study. Three participants were excluded from the analyses due to difficulties obtaining a reliable eye-tracking calibration across all experimental blocks. The final sample comprised 20 adults (15 females, M_{age} = 24.70 years, SD = 9.05).

Stimuli

We used the same stimuli, experimental set-up, and task as that reported in Caruana, de Lissa, and McArthur (2017a). Specifically, a virtual-reality character ("Alan"; see Fig. 1) was presented in the center of a computer screen (60 × 34 cm) at a distance of 65 cm from the participant. The screen had a refresh rate of 120 Hz, and Alan subtended 8 × 12° of visual angle. There were five versions of Alan that differed only in the direction of his gaze: looking straight at the participant or looking towards one of the four corners of the screen (i.e., top-left, top-right, bottom-left, or bottom-right corner). Each corner of the screen displayed a cartoon prison building that subtended 11° of visual angle and was positioned 15° of visual angle away from Alan's eyes.

Task

Participants interacted with Alan in a cooperative game that comprised two blocks of trials. Participants were told that Alan would be controlled by a human partner in another room in one block of trials, and by a computer program in another block. In reality, Alan was always controlled by a gaze-contingent computer algorithm (see Caruana, de Lissa, et al., 2015b, 2017a). In each trial of both blocks, a prisoner on the screen attempted to escape from one of the four prison buildings in each corner of the screen. The participant was told that they were the "watch person" who was responsible for alerting Alan (the "guard") about any attempted escapes. They could alert Alan by initiating joint attention to the breached building. Alan's job was to respond to the joint attention bid to catch the prisoner. Critically, participants were told that whilst they (the watch person) and Alan (the guard) could see each other, only the watch person could see what was happening outside the prison, and only the guard could see what was happening inside the prison. Participants were also told that the guard was additionally responsible for stopping fights between inmates within the prison, which may distract him from responding to joint attention bids on some trials.

At the start of each trial, a crosshair (subtending 1.5° of visual angle) was presented in the center of the screen. Once the participant had maintained fixation upon the crosshair for at least 150 ms, it was replaced by Alan's face, with his nasion in the same location as the crosshair. After a jittered delay of 200–1,000 ms, a "spotlight" appeared over one of the four cartoon buildings surrounding Alan's face to indicate the breached location (see Fig. 1). Participants were required to initiate joint attention towards the breached location by fixating on the spotlight for at least 150 ms, which triggered a cartoon prisoner to appear within the spotlight stimulus. At this point, participants were required to look back at Alan to evaluate his response. Once the participant fixated back on Alan's eyes, for a minimum of 150 ms, Alan "responded" after a 350- to 650-ms delay (jittered randomly between trials). Alan either shifted his gaze congruently towards the correct building to achieve joint attention and catch the prisoner (50% of trials; congruent condition), or he looked towards an incorrect building (50% of trials; incongruent condition). ERPs were time-locked to Alan's responsive gaze shift. Participants were also required to maintain fixation of Alan's eye gaze until the end of the trial (i.e., 1,000 ms after his responsive gaze shift) to ensure they attended to the gaze shift and that their eyes did not move during the epoch period.

The two blocks of trials comprised 120 trials each. A short break was provided after every 30 trials. At each break, participants were asked to estimate the percentage of trials in which Alan responded congruently. This encouraged participants to maintain attention to Alan's response over all trials. To mitigate any additional load on working memory,

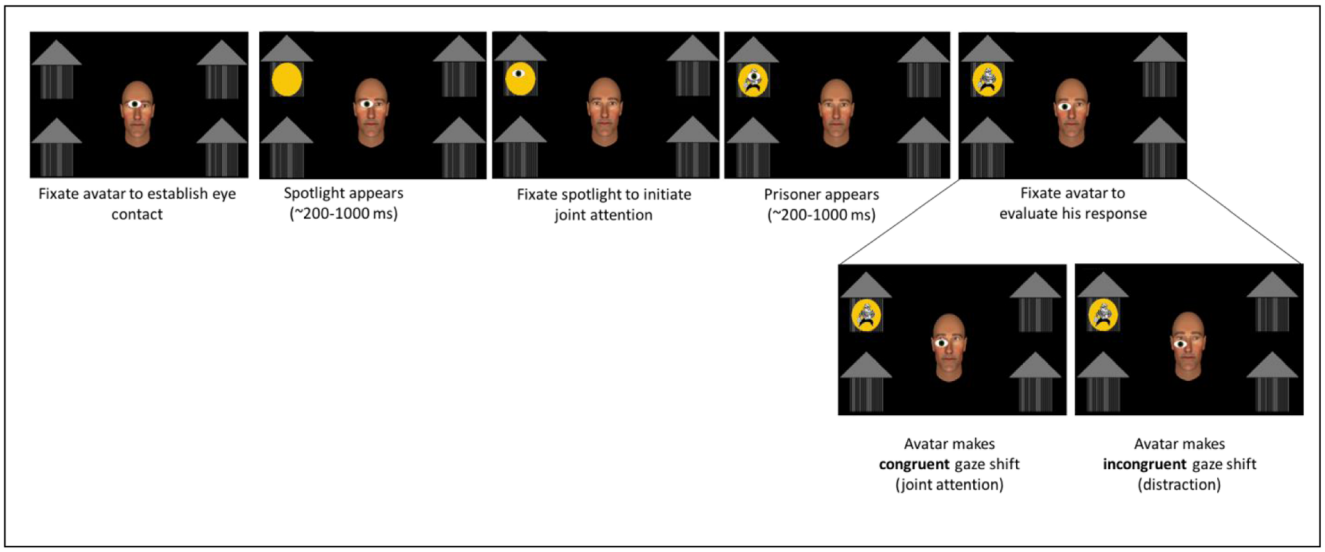

Fig. 1 Schematic representation of trial sequence. Schematic eye-balls represent the location of the participant's gaze and was not part of the stimuli visible to the participant

participants were told to avoid any explicit strategy (e.g., counting congruent trials) and were asked to simply pay attention and provide their subjective estimate. Overall participants' congruency estimations were highly accurate during both human and computer belief conditions, with mean estimate discrepancies of 8.31% (SD = 8.04) and 6.41% (SD = 9.25), respectively, and no significant differences between conditions ($t(19) = .68$, $p = .519$, $BF_{10} = .282$).

The order of the blocks was counterbalanced across participants (human-controlled first or computer-controlled first), and trial order was randomized within blocks. The location of breached exits and the direction of congruent and incongruent gaze shifts were also fully counterbalanced within each block.

Participants received negative feedback (i.e., text reading "Bad Fix" presented in the center of the screen) if they (1) did not fixate the spotlight, (2) prematurely fixated away from the spotlight before the prisoner appeared, (3) failed to fixate the avatar's face within 3,000 ms of finding the prisoner, or (4) fixated the avatar's face for less than 1,000 ms after finding the prisoner. This ensured that participants engaged with the task and were fixated on the avatar's face when gaze-related ERPs were being measured.

Eye movement and electroencephalogram (EEG) recording

Eye movements were recorded at 1,000 Hz using an EyeLink 1000 monocular (right eye) tower-mounted tracker with a head-stabilizing chin-rest. Online EEGs were recorded with a sampling rate of 1,000 Hz and an online band pass (.05–100 Hz) and notch filter (50 Hz) using a Synamps II amplifier. Recordings were measured from 29 electrodes positioned according to the 10–20 system (EasyCap; FP1, FP2, F7, F3, FZ, F4, F8, FT7, FC3, FC4, FT8, T7, C3, CZ, CPZ, C4, T8, TP7, CP3, CP4, TP8, P7, P3, PZ, P4, P8, O1, OZ, O2). Earlobes were used as sites for the online (left ear) and offline (right ear) reference electrodes. Horizontal and vertical electro-ocular activity (HEOG) was measured using bipolar electrodes positioned at the outer canthi (HEOG) and above and below the left eye (VEOG), respectively. A ground electrode was positioned between FP1, FP2, and FZ. Impedances were maintained below 5 kΩ for all electrodes.

ERP analysis

The raw EEG data was processed offline using Neuroscan 4.5 (El Paso, TX, USA). Electrical activity measured from VEOG was removed from the continuous data using the Scan 4.5 ocular reduction algorithm. These corrected data were then band-pass filtered (0.1–30 Hz) with a 12-dB octave roll-off and then segmented into epochs that were time-locked to the onset of Alan's gaze shifts. Epochs comprised a pre-stimulus baseline (-100 – 0 ms) and event period (0–700 ms). The Scan 4.5 artefact rejection algorithm was used to remove epochs containing voltages exceeding ±100 mV, and retained epochs were baseline-corrected. No electrodes were interpolated, and on average 89% of the 60 trials per condition were retained in the final analysis (Human-congruent: M = 53.2, SD = 7.05; Human-incongruent: M = 54.1, SD = 4.61; Computer-congruent: M = 53.15, SD = 5.37; Computer-incongruent: M = 52.75, SD = 5.52). Accepted epochs were averaged separately for

each of the four conditions to create the gaze-related ERPs (i.e., congruent-human, incongruent-human, congruent-computer, incongruent-computer).

We took eight indices of the ERPs. In line with our previous studies, we measured the mean amplitude of the P250 (170–300) and P350 (310–440 ms) at CZ and PZ. Data from these electrodes are reported separately, rather than as a centroparietal cluster, to allow direct comparison of results with our previous findings (Caruana, de Lissa, & McArthur, 2015b, 2017a). As noted above, whilst the P250 interval was previously not analyzed, a retrospective analysis of previous data (from Caruana, de Lissa, & McArthur, 2015b, 2017a) revealed that reliable peaks were observed in this interval. Specifically, in our original study where participants believed the avatar was controlled by a human, all 19 participants had clear P250 peaks in at least one task condition, with the majority (n=14) exhibiting clear P250 peaks in all task conditions. This was also the case for all participants excluded from the final analyses (n=5; Caruana, de Lissa, & McArthur, 2015b). In our following study, a second sample of 19 participants completed the task, believing that the avatar was computer-controlled. Fifteen participants exhibited clear P250 peaks in all task conditions. The remaining four had P250 responses, which emerged during the P250 interval, but peaks were obscured by a subsequent P350 response (Caruana, de Lissa, & McArthur, 2017a). In the current study, 18/20 participants exhibited a clear P250 peak in at least two conditions (11 of these had clear P250 peaks in all task conditions). Only two participants had P250 responses that did not exhibit a clear peak in the P250 interval, as they were obscured by subsequent P350 responses. Taken together, across all three samples of individuals who have completed this task (n=63), 90.5% have exhibited a clear P250 peak in at least one task condition. Thus, our previous and current data reveals that there is a reliable P250 response elicited during this task, which appears to be independent of, although occasionally obscured by, the subsequent P350 response.

We also took two new measures of the P350, calculated by subtracting each individual's P250 mean amplitude from their P350 mean amplitude in each condition at CZ and PZ (see P350–P250 in Table 2 for summary statistics). We added these measures because it has become apparent across our studies that the P350 ERP "builds upon" the P250 ERP response. P250 effects may therefore have a carry-over influence on the P350 peaks, which can be controlled by subtracting P250 responses from the P350 interval.

Our final two ERP measures were the peak amplitudes of the N170 (107–237ms) at P7 and P8. We measured the N170 in all our studies to date because the N170 is sensitive to the perceptual encoding of faces and biological cues within them – including eye gaze (see Itier & Batty, 2009, for review). If the N170 reflects the same condition effects as the P250 or P350, this would suggest that these effects were due to condition differences in stimulus features rather than intentional stance.

Subjective experience interview

At the end of the testing session, participants were asked to rate their experience of the task across the two belief conditions using a five-point Likert scale (1 = not at all to 5 = extremely). Participants were asked to rate how difficult, natural, and pleasant the task was, and how natural and pleasant the social interaction was. Participants were also asked to rate how human-like the avatar felt, appeared, and behaved. Separate ratings were provided for the Human and Computer conditions. At the end of the experiment, participants were debriefed and asked to rate how convinced they were that a real person controlled the avatar during the human block.

Statistical analysis

The effects of congruency (congruent vs. incongruent) and belief (human vs. computer) on each ERP measure were assessed in two-way ANOVAs using Jamovi (Jamovi Project, 2018). Significant interactions were evaluated using *post hoc* paired comparisons with Bonferroni corrections for multiple comparisons. To test if any effects of belief were driven by block order, we re-calculated the ANOVAs for two separate subgroups of participants: those who were told that Alan was controlled by a human in the first block versus those who were told he was controlled by a computer. Within-subjects comparisons of subjective ratings across the two belief conditions were evaluated using Wilcoxon signed-rank tests. An alpha level of $p < 0.05$ was used for all planned analyses.

Results

Subjective experience interview

Participant ratings indicated that they found the task easy, natural, and pleasant. Participants rated the task as being significantly less difficult and more natural when they believed Alan was controlled by a human than by a computer. Similarly, they rated the interaction itself as being significantly more natural during the Human than the Computer condition. Finally, participants indicated that the Alan "felt" and "behaved" significantly more human-like when they believed it was being controlled by a human. There were no significant differences between belief conditions for any other rating. The descriptive and test statistics for these subjective ratings analyses are summarized in Table 1.

Table 1. Subjective experience ratings

	Human *M(SD)*	Computer *M(SD)*	*z*	*p*
Task				
Difficult	1.65 (0.81)	2.15 (1.09)	-2.49	.01*
Natural	3.95 (1.15)	3.30 (1.30)	-2.14	.03*
Pleasant	3.95 (0.89)	3.50 (1.15)	-1.90	.06
Interaction				
Natural	3.6 (1.9)	3.05 (1.15)	-2.07	.04*
Pleasant	3.6 (1.19)	3.15 (1.14)	-1.73	.08
Anthropomorphism				
Felt human	3.50 (1.15)	2.50 (1.10)	-2.84	<.01*
Appeared human	3.70 (0.86)	3.70 (0.98)	0.00	1.00
Behaved like a human	3.85 (0.81)	3.40 (1.05)	-2.53	.01*

Note. Ratings provided on a 5-point scale (1=not at all, 5=extremely)

*Denotes a significant difference between belief conditions (human vs. computer)

ERPs

Summary statistics for the mean amplitude of P250, P350, and P350-P250 at CZ and PZ, and the N170 at P7 and P8 are shown in Table 2. Group average waveforms and topographic maps for the P250 and P350 are shown in Fig. 2, and for the N170 in Fig. 3. The mean amplitude differences between the P350 and P250 analysis intervals are summarized by condition and electrode in Fig. 4.

P250. There was a significant main effect of congruency measured at PZ ($F(19) = 5.26$, $p = .033$, $\eta^2 = .217$), characterized by larger responses overall to congruent compared to incongruent gaze shifts. This was not significant at CZ ($F(19) = 3.47$, $p = .078$, $\eta^2 = .154$). There was no significant main effect of belief at either CZ ($F(19) = 3.19$, $p = .090$, $\eta^2 = .144$) or PZ ($F(19) = 1.97$, $p = .177$, $\eta^2 = .090$). However, there was a significant belief by congruency interaction at both CZ ($F(19) = 13.44$, $p = .002$, $\eta^2 = .414$) and PZ ($F(19) = 16.91$, $p = .001$, $\eta^2 = .471$). *Post hoc* tests ($\alpha = .017$, Bonferroni corrected for three comparisons) revealed that the P250 generated in the human-congruent condition was significantly larger than those observed in the human-incongruent [CZ ($t(28.6) = 3.34$, $p = .002$); PZ ($t(27.9) = 3.89$, $p < .001$)], computer-congruent [CZ ($t(30.3) = 3.38$, $p = .002$); PZ ($t(27.8) = 3.08$, $p = .005$)], and computer-incongruent [CZ ($t(37.6) = 2.58$, $p = .014$); PZ ($t(38.0) = 2.61$, $p = .013$)] conditions.

To determine if these effects were influenced by belief order, we separately analyzed the data from participants who were told that Alan was controlled by a human in the first block (n=9) and those who believed that the agent was computer-controlled in the first block (n=11). Despite modest sample sizes, the same effects were observed in these analyses. Specifically, we found a significant belief-by-condition interaction for the P250 at both CZ [human first ($F(8) = 6.39$, $p = .035$, $\eta^2 = .444$); computer first ($F(10) = 6.51$, $p = .029$, $\eta^2 = .394$)] and PZ [human first ($F(8) = 12.58$, $p = .008$, $\eta^2 = .611$); computer first ($F(10) = 5.65$, $p = .039$, $\eta^2 = .361$)] in both groups of participants.

P350. There was a significant main effect of belief measured at CZ ($F(19) = 5.09$, $p = .036$, $\eta^2 = .211$) and PZ ($F(19) = 5.02$, $p = .037$, $\eta^2 = .209$), characterized by larger responses overall when participants believed they were interacting with a human than a computer. There was no significant main effect of condition at CZ ($F(19) = 0.25$, $p = .625$, $\eta^2 = .013$) or PZ ($F(19) = 0.24$, $p =$

Table 2. Summary statistics for amplitude measures by electrode

Mean amplitude	**Congruent**	**Incongruent**	**Congruent**	**Incongruent**
	CZ		**PZ**	
P250				
Human	9.30 (5.07)	7.26 (3.77)	6.49 (4.31)	4.23 (3.44)
Computer	7.39 (4.86)	7.42 (4.73)	4.69 (5.08)	4.56 (4.27)
P350				
Human	11.36 (5.68)	11.24 (6.46)	11.17 (6.01)	10.28 (5.86)
Computer	9.05 (4.78)	9.71 (5.09)	8.39 (5.35)	9.08 (4.91)
P350-P250				
Human	2.06 (4.95)	3.98 (5.60)	4.68 (5.25)	6.06 (5.17)
Computer	1.66 (4.64)	2.29 (4.55)	3.70 (4.39)	4.53 (3.66)
	P7		**P8**	
N170 Peak Amplitude				
Human	-4.37 (2.75)	-4.71 (3.07)	-7.03 (2.66)	-7.66 (3.35)
Computer	-4.35 (2.82)	-4.15 (2.29)	-7.35 (3.11)	-7.22 (3.12)

Note. Summary statistics are provided in the format of mean (standard deviation)

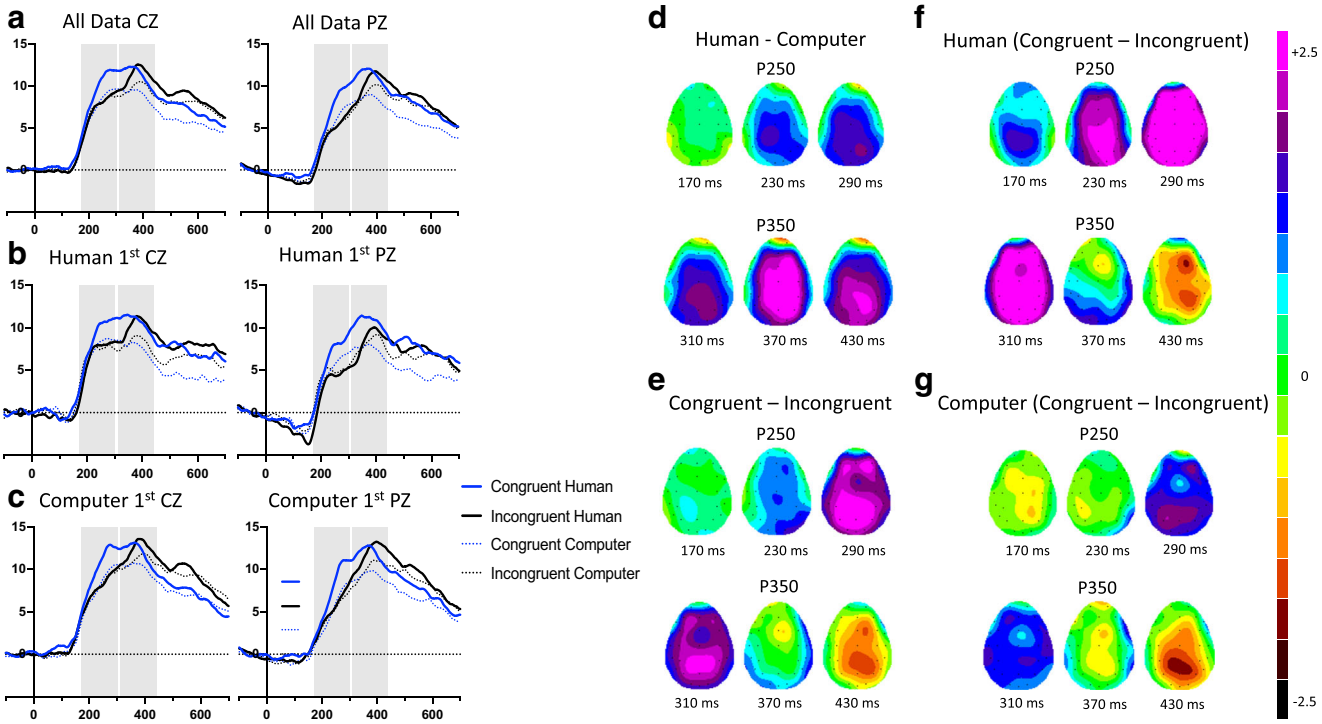

Fig. 2 Group average waveforms comprising the P250 and P350 intervals measures at Cz (left column) and Pz electrodes (right column) for (**A**) all participants irrespective of belief order (n=20), (**B**) participants who believed partner was human-controlled in first block (n=9), and (**C**) participants who believed partner was computer-controlled in first block (n=11). Mean amplitude (μV) and latency (ms) are plotted on the X and Y axes, respectively. Analysis intervals indicated in greyshade. Voltage change topographies (i.e., scalp differences) are visualized for the whole sample, for the main effect of belief in both intervals (**D**), the main effect of congruency (**E**), and the condition effect (Congruent – Incongruent) for the human-belief (**F**) and computer-belief (**G**) conditions

.879, η^2 = .001). There was a significant belief by condition interaction at PZ ($F(19)$ = 5.39, p = .031, η^2 = .221) but not CZ ($F(19)$ = 1.04, p = .321, η^2 = .052). *Post hoc* tests (α = .025, Bonferroni corrected for two comparisons) revealed that the difference between the P350 in the human- and computer-belief blocks was significant for the congruent condition [CZ (t (26.3) = 2.46, p = .020); PZ (t(24.5) = 2.92, p = .007)], but not for the incongruent condition [CZ (t (26.3) = 1.64, p = .327); PZ (t(24.5) = 1.26, p = .219)].

P350-P250. The main effects of belief ($F(19)$ = 4.16, p = .056, η^2 = .011) and condition ($F(19)$ = 2.96, p = .102, η^2 = .017) were not statistically significant at CZ. However, there was a significant belief-by-condition interaction ($F(19)$ = 4.56, p = .046, η^2 = .004) at CZ. *Post hoc* tests (α = .025, Bonferroni corrected for two comparisons) revealed that there was a significantly larger increase in mean amplitude in the human-incongruent compared to the computer-incongruent condition (t(30.6) = 2.84, p = .008; also see Fig. 4), but there was no significant difference between the human-congruent and computer-congruent conditions (t(30.6) = 0.68, p = .500). At PZ we found a main effect of belief, with larger increases in mean amplitude in the human belief conditions ($F(19)$ = 5.53, p = .030, η^2 = .018). There was no significant main effect of condition ($F(19)$ = 2.46, p = .134, η^2 = .014) nor a belief by condition interaction measured at this site ($F(19)$ = .98, p = .335, η^2 = .001).

In line with the P250 analyses, we separately analyzed the P350–P250 data from participants who believed Alan was human- or computer-controlled in the first block they completed. We found a significant main effect of belief at CZ ($F(8)$ = 11.46, p = .010, η^2 = .033) and PZ ($F(8)$ = 7.77, p = .024, η^2 = .039) for the human-first group (n=9) but not the computer-first group (n=11) at CZ ($F(10)$ = .343, p = .571, η^2 = .002) or PZ ($F(10)$ = 1.24, p = .292, η^2 = .009). We found no evidence for a main effect of condition, or a belief by condition interaction at either electrode, when each subgroup's P350–P250 data was analyzed separately (all ps > .09).

N170 peak amplitude. There was no significant main effect of belief [(P7: ($F(19)$ = .843, p = .370, η^2 = .042); P8: ($F(19)$ = .016, p = .900, η^2 = .001] or condition [(P7: ($F(19)$ = .110, p = .743, η^2 = .006); P8: ($F(19)$ = .965, p = .338, η^2 = .048], nor a belief-by-condition interaction [(P7: ($F(19)$ = 1.23, p = .061, η^2 = .061); P8: ($F(19)$ = 3.02, p = .099, η^2 = .137] for the N170 at either P7 or P8.

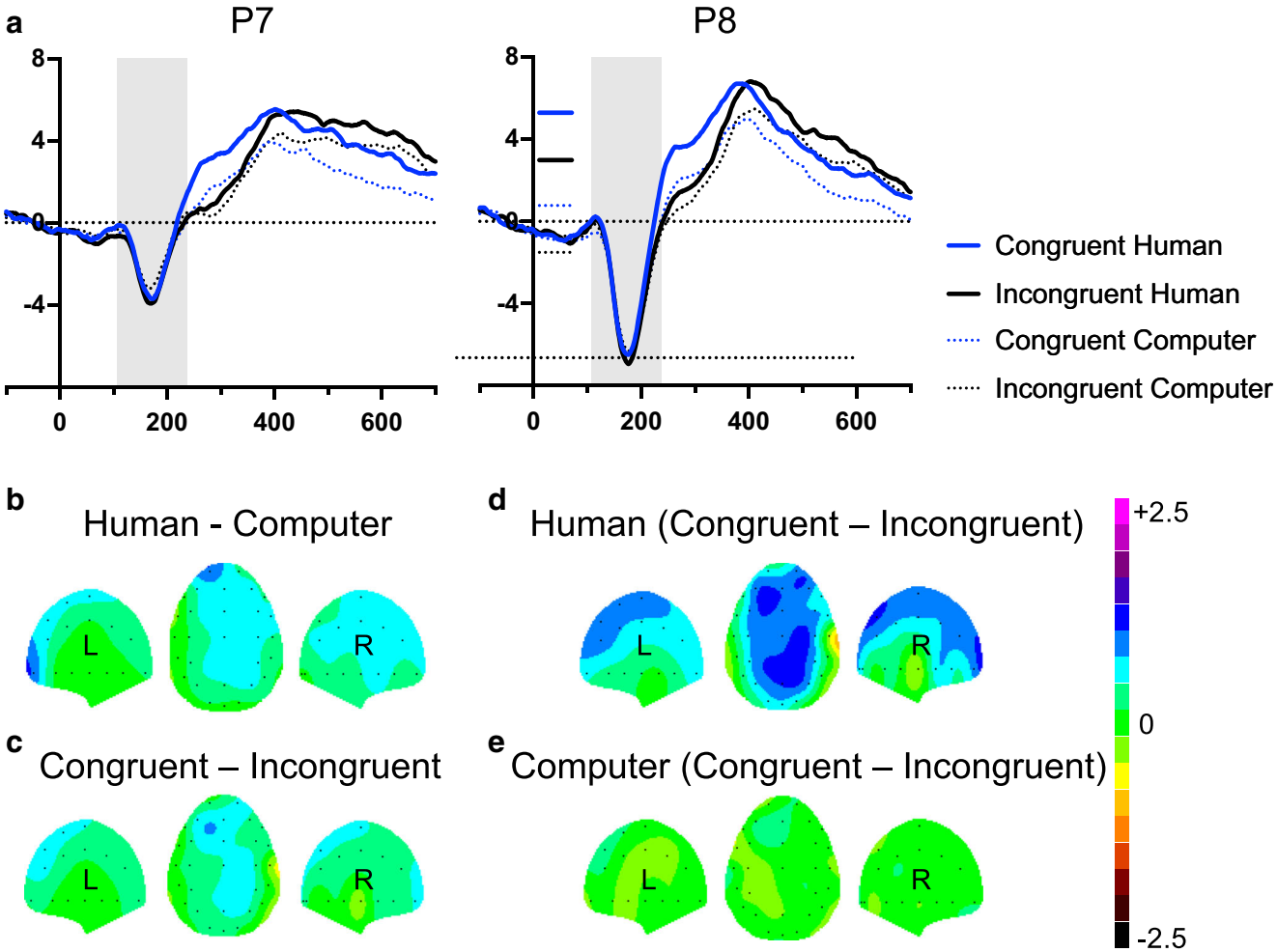

Fig. 3 (**A**) Group average waveforms comprising the N170 at P7 and P8 electrodes. Peak amplitude (μV) and latency (ms) are plotted on the X and Y axes, respectively. Analysis intervals indicated in greyshade. Voltage change topographies (i.e., scalp differences) are visualized for the main effect of belief in both intervals (**B**), the main effect of congruency (**C**) and the condition effect (Congruent – Incongruent) for the human-belief (**D**) and computer-belief (**E**) conditions. Left, top and right topography views are presented for each scalp difference

Discussion

There is growing evidence that how we react to gaze cues is influenced by the degree to which we believe them to originate from a human-like entity (Caruana, McArthur, Woolgar, & Brock, 2017b; Caruana, Spirou, & Brock, 2018; Weise et al., 2017). However, much of this evidence has been indirect, with studies manipulating the esthetic or behavioral features of stimuli to modulate intentional stance. Other studies have directly manipulated beliefs to test the role of intentional stance, but have relied on between-subjects designs that cannot rule out the impact of incidental individual differences between groups. The current study used a within-subjects design to directly test if adopting an intentional stance reliably modulates the neural encoding of gaze during joint attention using an interactive paradigm.

Consistent with our previous findings, and those of intentional stance manipulations in other social contexts (e.g., Schindler & Kissler, 2016), the centroparietal P250 and P350 ERPs were significantly larger when individuals believed that they were interacting with another human than when they believed they were interacting with a computer (Caruana, de Lissa et al., 2015b, 2017a). Of specific importance to our study was the P250 and P350–P250 belief-by-congruency interaction effects. The centroparietal P250 was significantly larger to congruent gaze shifts that signalled the achievement of joint attention than incongruent gaze shifts, but only when participants believed that they were interacting with another human. In contrast, the P250 ERP did not differ between gaze shifts that did not signal the achievement of joint attention with another human (i.e., incongruent-human, congruent-computer, incongruent-computer conditions). This

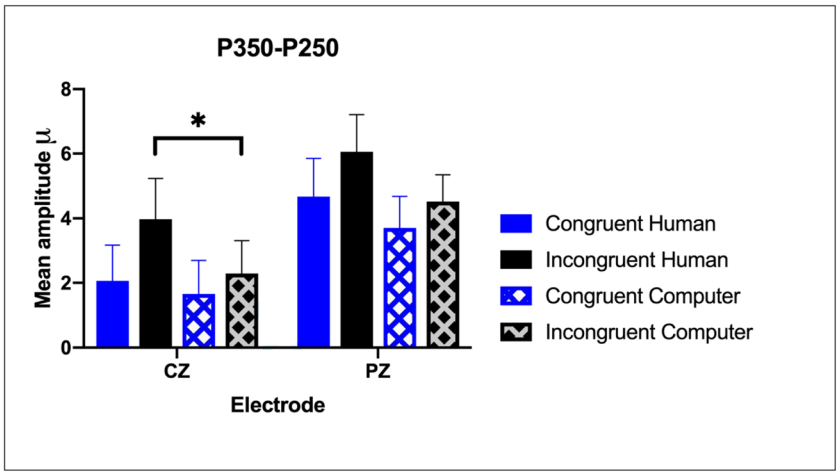

Fig. 4 Mean amplitude differences between the P350 and P250 analysis periods for each condition, measured at CZ and PZ. Error bars represent standard error

P250 interaction effect was observed irrespective of the order in which participants adopted an intentional stance during the experiment, suggesting that intentional stance has a powerful and consistent influence on the evaluation of interactive gaze. The belief order analysis also demonstrates that the observed interactions cannot be explained by habituation or fatigue across blocks, given that the direction of the effect is consistent across participants who completed the task under different belief orders.

We also found that the relative increase in ERP mean amplitude from the P250 to the P350 (i.e., the P350–P250) was larger to incongruent gaze shifts (i.e., failed joint attention) than to congruent gaze shifts (successful joint attention) when participants believed that they were interacting with a human. Unlike our previous findings, absolute mean amplitude for the P350 did not significantly vary as a function of joint attention outcome, although it was reliably larger when individuals adopted an intentional stance. One explanation for the lesser reliability of P350 mean amplitude as an index of joint attention is that it builds upon the P250, and hence its absolute amplitude reflects an interaction of the P250 and the P350. In this study, we attempted to remove the influence of the P250 from the P350 by subtracting the amplitude of the former from the latter for each participant (i.e., the P350–P250). This appeared to be successful since the P350–P250 was more sensitive to interaction effects than the P350. We will assess the reliability of these indices in future studies.

One potential interpretation of the observed P250 and P350–P250 effects is that they reflect the affective responses, both positive and negative, associated with achieving or failing to achieve joint attention with another person. Social neuroscience research continues to demonstrate the many factors that influence how we evaluate and respond to social information conveyed by social interlocutors – including their social influence, persuasiveness, and the value we place on our relationship with them (see Falk & Sholz, 2018, for a review). One critical factor is the human desire to be affiliated with, and approved of by, others (Baumeister & Leary, 1995). It therefore stands to reason that when we attribute humanness and intentionality to an entity, we initialize a suite of mentalizing computations that not only enable the representation of another's perspective, but also evaluate how they perceive, judge, and accept us. Thus, one might evaluate the social alignment experienced during joint attention as a sign of affiliation and acceptance, which has a positive affective consequence compared to the avoidance of joint attention.

Specific to the context of gaze-based interactions, fMRI studies implementing similar paradigms to the current study have found that evaluating the achievement of joint attention is associated with increased activation in mentalizing and social-reward substrates, including the medial prefrontal cortex (Williams et al., 2005; Wilms et al., 2010), amygdala (Gordon et al., 2013) and ventral striatum (Pfeiffer et al., 2014; Schilbach et al., 2010). Future work is needed to investigate whether the ERPs observed in this study reflect the social relevance of the spatial gaze cues or reflect a later stage of processing in which an affective evaluation is associated with this outcome. This could be done by manipulating whether the interactive context is cooperative or competitive. In the current task, achieving joint attention cooperatively with the virtual agent was associated with task success (i.e., catching the prisoner), and thus the P250 may reflect a hedonic response. It would be interesting to test whether a reversal of the observed ERP effect (i.e., larger P250 response for incongruent gaze shifts) is observed in a competitive context where task success is signalled by the independent capture of the prisoner, and the virtual partner's failure to achieve joint attention. This would provide evidence for the ERP effects reflecting affective responses to the achievement or avoidance of joint attention with another person.

Whilst the current study cannot determine whether the intentional stance effect on the P250 and P350 ERPs reflects processes of a cognitive or affective nature, the modulation of later ERPs (> 170 ms) suggests the intentional stance effect manifests at later, more evaluative, stages of social cortical processing. At the very least, our data suggest that social contexts that encourage self-representation or mentalizing processes (i.e., the human condition) – as reflected by enhanced P250 and 350 ERPs – are consistent with the motivated attention account (Schupp et al., 2003, 2004).

Furthermore, given that we did not find evidence of an occipitotemporal N170 modulation by belief in our study, our data also suggest that adopting an intentional stance is unlikely to impact on the perceptual encoding of gaze shifts. It is noteworthy that we did find tentative evidence of a left-lateralized belief effect on N170 peak amplitude in our previous work. However, this could have been an artifact of the between-subjects manipulation employed (i.e., an effect driven by inadvertent individual differences between groups; Caruana, de Lissa, & McArthur, 2017a). This emphasizes the importance of using within-subjects designs when investigating the role of social context, belief, and expectations, including whether one adopts an intentional stance.

Implications for social neuroscience research

The current study demonstrates that deception-induced intentional stance manipulations can be achieved in within-subjects designs, and that this results in differential neural processing. Therefore, adopting an intention stance is likely an important ingredient in obtaining an ecologically valid measure of the neurocognitive mechanisms of social interaction. This aligns with other findings using similar paradigms that have demonstrated that individuals adopt specific and distinct behavioral strategies when responding to and initiating joint attention bids with a virtual partner (Caruana, Spirou & Brock, 2018).

These methodological implications are also important for social neuroscience investigations of atypical development and psychiatric conditions (e.g., autism, schizophrenia, social anxiety), where there is a need for an optimal balance of ecological validity and experimental control (see Caruana et al., 2017b, for discussion and review). In these studies, there is a need to test cognitive explanations for social impairments whilst controlling for the impact of concurrent non-social cognitive difficulties. This requires the design of well-matched non-social control tasks that are equivalent in their visual and cognitive complexity. A longstanding challenge in gaze-processing research has been the design of non-social gaze stimuli that are not "social" but are visually matched. Deception-induced intentional stance manipulations may be one, albeit conservative, way to investigate the social communication difficulties that may be specific to contexts in which individuals truly believe they are interacting with another person. Together, our findings are important to the field of social neuroscience as they (1) provide unequivocal evidence for the neural time course for evaluating the achievement of gaze-based joint attention, which occurs approximately 250 ms after observing a social partner's responsive gaze shift, and (2) highlight the potential utility and influence of deception-induced intentional stance manipulations in elucidating the neurocognitive mechanisms of social interaction and their divergence in certain populations (cf. Schilbach et al., 2013; Schilbach, 2016).

Implications for virtual reality and HRI research

The outcomes of this study also have implications for virtual reality and human-robot interaction (HRI) research. Understanding how intentional stance influences the neurocognitive mechanisms of social interactions – including how we interpret and respond to non-verbal cues – is becoming a central focus in HRI research. Humanoid robots are already emerging as cohabitants in our world, joining our workforce and households, and fulfilling several social roles by providing companionship to the lonely, and increasing access to face-to-face healthcare delivery, education, social-cognitive training, and workplace collaboration (e.g., Birks et al., 2016; Hinds et al., 2004; Martin et al., 2016; Tapus et al., 2007; Warren et al., 2015). One fascinating goal of HRI research is to determine how self-controlled robots should be designed to induce the adoption of intentional stance in human end-users, which is thought to have a significant impact on human performance – both positive and negative – in various contexts (see Wiese, Metta, & Wykowska, 2017, for a comprehensive review).

One of the benefits of simulated agents – both virtual or robotic – is that they can engage in automated, yet compelling, social behavior (see Caruana et al., 2017b; Georgescu et al., 2014 for relevant reviews). However, current evidence suggests that the degree to which a user trusts, learns from, or enjoys interacting with a simulated agent may depend on the degree to which the agent can induce an intentional stance. This is supported by the current study, which found evidence for this in the subjective ratings analysis. Adopting an intentional stance made the task seem easier and the interaction with Alan more natural. This aligns with a growing body of research that has established that taking an intentional stance towards a non-human character can have a broad impact on social attitudes and behavior. Specifically, it has been found to strengthen empathic responses and prosocial moral decision-making (Gutsell & Inzlicht, 2012; Haley & Fessler, 2005; Haslam, 2006; Harris & Fiske, 2006) and increase the perceived relevance of social actions and cues (Özdem et al., 2017; Wiese et al., 2012).

As such, it has become necessary to assess the extent to which simulated agents (e.g., robots) induce the adoption of an intentional stance. To date, this has largely been achieved

using subjective measures, such as asking participants to rate the likelihood of an agent having a mind (e.g., Weise et al., 2017) or completing a Turing Test task where they must guess whether the agent is human- or computer-controlled (e.g., Pfeiffer et al., 2014). A limitation of this approach is the inherent unreliability of conscious and subjective judgments. The outcomes of the current study suggest an objective neural marker of intentional stance in the P250 ERP, which appeared reliable even when scrutinized in small subsets of the group data. Further, it was not affected by the order in which participants adopted or abandoned an intentional stance, suggesting that it may even be sensitive to transient shifts in intentional stance. This neural marker may provide a new objective tool for evaluating the esthetic and behavioral parameters that are most likely to induce a deception-free intentional stance towards virtual or robotic agents.

Conclusion

The signalling and perceiving of gaze cues are critical for navigating the constant stream of social interaction in our lives. Successfully interpreting these cues enables us to understand the behaviors and intentions of others so that we can respond appropriately. Our data reveal that adopting an intentional stance results in a shift in the neural encoding of gaze shifts that signal social engagement via the achievement of joint attention with a virtual character. This neural marker might prove useful for the future development and validation of simulated social agents capable of spontaneously inspiring an intentional stance. Social neuroscience researchers and technology developers interested in working with virtual or robotic agents should to be aware of the influence that adopting an intentional stance has on the way people (both research participants or consumers) evaluate, engage, and feel when they interact with these agents.

References

Baumeister, R. F., & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin, 117,* 497–529.

Birks, M., Bodak, M., Barlas, J., Harwood, J., and Pether, M. (2016). Robotic seals as therapeutic tools in an aged care facility: a qualitative study. *J. Aging Res.* 2016, 1–7. https://doi.org/10.1155/2016/8569602

Bruinsma Y, Koegel RL, Koegel LK. 2004. Joint attention and children with autism: a review of the literature. Mental Retardation & Developmental Disabilities Research Reviews 10(3):169–175. https://doi.org/10.1002/mrdd.20036.

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), 763-766.

Caruana, N., Brock, J., & Woolgar, A. (2015a). A frontotemporoparietal network common to initiating and responding to joint attention bids. *NeuroImage, 108*, 34-46. https://doi.org/10.1016/j.neuroimage.2014.12.041

Caruana, N., de Lissa, P., & McArthur, G. (2015b). The neural time course of evaluating self-initiated joint attention bids. *Brain and Cognition*. https://doi.org/10.1016/j.bandc.2015.06.001

Caruana, N., de Lissa, P., & McArthur, G. (2017a). Beliefs about human agency influence the neural processing of gaze during joint attention. Social. Neuroscience., https://doi.org/10.1080/17470919.2016.1160953.

Caruana, N., McArthur, G., Woolgar, A., and Brock, J. (2017b). Simulating social interactions for the experimental investigation of joint attention. *Neurosci. Biobehav. Rev.* 74, 115–125. https://doi.org/10.1016/j.neubiorev.2016.12.022

Caruana, N., Spirou, D., & Brock, J. (2018). Human agency beliefs influence behaviour during virtual social interactions*, PeerJ 5,* e3819

Charman, T. (2003). Why is joint attention a pivotal skill in autism? *Philosophical Transactions Royal Society London Biological Sciences, 358*, 315-324. https://doi.org/10.1098/rstb.2002.1199

Dawson, G., Toth, K., Abbott, R., Osterling, J., Munson, J. A., Estes, A., & Liaw, J. (2004). Early social attention impairments in autism: Social orienting, joint attention, and attention to distress. *Developmental Psychology, 40*(2), 271-283.

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioural and Brain Sciences, 11,* 357-374.

Emery, N. J. (2000). The eyes have it: the neuroethology, function and evolution of social gaze. *Neuroscience and Biobehavioral Reviews, 24*(6), 581-604. doi:https://doi.org/10.1016/S0149-7634(00)00025-7

Falk, E., & Scholz, C. (2018). Persuasion, Influence, and Value: Perspectives from Communication and Social Neuroscience. *Annual Review of Psychology, 69*(1), 329–356. https://doi.org/10.1146/annurev-psych-122216-011821

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

Gobel MS, Kim HS, Richardson DC. 2015. The dual function of social gaze. Cognition 136:359–364. https://doi.org/10.1016/j.cognition.2014.11.040.

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*, 1-11. https://doi.org/10.1080/17470919.2013.832374

Grossmann, T. (2017). The eyes as a window into other minds: An integrative perspective. *Perspectives on Psychological Science, 12,* 107–121. https://doi.org/10.1177/1745691616654457.

Gutsell, J. N., and Inzlicht, M. (2012). Intergroup differences in the sharing of emotive states: neural evidence of an empathy gap. *Soc. Cogn. Affect. Neurosci.* 7, 596–603. https://doi.org/10.1093/scan/nsr035

Hajcak, G., MacNamara, A., & Olvet, D. M. (2010). Event-Related Potentials, Emotion, and Emotion Regulation: An Integrative Review. *Developmental Neuropsychology, 35*(2), 129–155. https://doi.org/10.1080/87565640903526504

Haley, K. J., and Fessler, D. M. T. (2005). Nobody's watching? Subtle cues affect generosity in an anonymous economic game. *Evolution and Human Behavior* 26, 245–256. https://doi.org/10.1016/j.evolhumbehav.2005.01.002

Harris, L. T., and Fiske, S. T. (2006). Dehumanizing the lowest of the low – neuroimaging responses to extreme out-groups. *Psychol. Sci.* 17, 847–853. https://doi.org/10.1111/j.1467-9280.2006.01793.x

Haslam, N. (2006). Dehumanization: an integrative review. *Pers. Soc. Psychol. Rev.* 10, 252–264. https://doi.org/10.1207/s15327957pspr1003-4

Hinds, P., Roberts, T., and Jones, H. (2004). Whose job is it anyway? A study of human–robot interaction in a collaborative task. *Hum. Comput. Interact.* 19, 151–181. https://doi.org/10.1207/s15327051hci1901&2_7

Itier, R. J., & Batty, M. (2009). Neural bases of eye and gaze processing: The core of social cognition. *Neuroscience & Biobehavioral Reviews, 33*(6), 843-863. https://doi.org/10.1016/j.neubiorev.2009.02.004

Jamovi Project (2018). jamovi (Version 0.9) [Computer Software]. Retrieved from https://www.jamovi.org

Martini, M. C., Gonzalez, C. A., and Wiese, E. (2016). Seeing minds in others–Can agents with robotic appearance have human-like preferences? *PLOS ONE* 11:e0146310. https://doi.org/10.1371/journal.pone.0146310

Mundy, P., & Newell, L. (2007). Attention, joint attention, and social cognition. *Current Directions in Psychological Science, 16*(5), 269–274. https://doi.org/10.1111/j.1467-8721.2007.00518.x

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus–norepinephrine system. *Psychological Bulletin, 131*(4), 510.

Özdem, C., Wiese, E., Wykowska, A., Müller, H., Brass, M., & Van Overwalle, F. (2017). Believing androids–fMRI activation in the right temporoparietal junction is modulated by ascribing intentions to non-human agents. *Social Neuroscience, 12*(5), 582–593.

Pelphrey, K. A., Shultz, S., Hudac, C. M., & Vander Wyk, B. C. (2011). Research Review: Constraining heterogeneity: the social brain and its development in autism spectrum disorder. *Journal of Child Psychology and Psychiatry, 52*(6), 631-644. https://doi.org/10.1111/j.1469-7610.2010.02349.x

Pfeiffer, U. J., Schilbach, L., Timmermans, B., Kuzmanovic, B., Georgescu, A. L., Bente, G., & Vogeley, K. (2014). Why we interact: on the functional role of the striatum in the subjective experience of social interaction. *NeuroImage, 101*, 124-137. https://doi.org/10.1016/j.neuroimage.2014.06.061

Polich, J. (2007). Updating P300: An Integrative Theory of P3a and P3b. *Clinical Neurophysiology, 118* (10), 2128–2148. https://doi.org/10.1016/j.clinph.2007.04.019

Premack D, Woodruff G. 1978. Does the chimpanzee have a theory of mind? Behavioral and Brain Sciences 1:515–526. https://doi.org/10.1017/S0140525X00076512.

Schilbach, L. (2016). Towards a second person neuropsychiatry. *Philosophical Transactions of the Royal Society, London, B. Biological Sciences, 19,* 371. https://doi.org/10.1098/rstb.2015.0081

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

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. https://doi.org/10.1016/j.neuroimage.2016.03.052

Schindler, S., Wegrzyn, M., Steppacher, I., & Kissler, J. (2015). Perceived Communicative Context and Emotional Content Amplify Visual Word Processing in the Fusiform Gyrus. *The Journal of Neuroscience, 35* (15), 6010–6019. https://doi.org/10.1523/JNEUROSCI.3346-14.2015

Schindler, S., Zell, E., Botsch, M., & Kissler, J. (2017). Differential effects of face-realism and emotion on event-related brain potentials and their implications for the uncanny valley theory. *Scientific Reports, 7*, 45003. https://doi.org/10.1038/srep45003

Schupp, H.T., Cuthbert, B., Bradley, M., Hillman, C., Hamm, A., Lang, P.J., 2004. Brain processes in emotional perception: motivated attention. *Cogn. Emot. 18* (5), 593–611.

Schupp, H. T., Flaisch, T., Stockburger, J., & Junghofer, M. (2006). Emotion and attention: event-related brain potential studies. Progress in Brain Research, 156, 31–51.

Schupp, H.T., Junghöfer, M., Weike, A.I., Hamm, A.O., 2003. Emotional facilitation of sensory processing in the visual cortex. *Psychol. Sci. 14*, 7–13.

Tapus, A., Mataric, M. J., and Scasselati, B. (2007). Socially assistive robotics [Grand challenges of robotics]. *IEEE Robot. Autom. Mag.* 14, 35–42. https://doi.org/10.1109/MRA.2007.339605

Warren, Z. E., Zheng, Z., Swanson, A. R., Bekele, E., Zhang, L., Crittendon, J. A., et al. (2015). Can robotic interaction improve joint attention skills? *J. Autism Dev. Dis.* 45, 1–9. https://doi.org/10.1007/s10803-013-1918-4

Waytz A, Cacioppo J, Epley N. 2010. Who sees human? The stability and importance of individual differences in anthropomorphism. Perspectives on Psychological Science 5(3):219–232 DOI https://doi.org/10.1177/1745691610369336.

Wiese, E., Buzzell, G., Abubshait, A., & Beatty, P. (2018). Seeing minds in others: Mind perception modulates social-cognitive performance and relates to ventromedial prefrontal structures, *CABN.* https://doi.org/10.31234/osf.io/ac47k

Wiese, E., Metta, G., & Wykowska, A. (2017). Robots As Intentional Agents: Using Neuroscientific Methods to Make Robots Appear More Social. Frontiers in Psychology, 8. https://doi.org/10.3389/fpsyg.2017.01663

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

Williams, J. H. G., Waiter, G. D., Perra, O., Perrett, D. I., & Whiten, A. (2005). An fMRI study of joint attention experience. *NeuroImage, 25*(1), 133–140. https://doi.org/10.1016/j.neuroimage.2004.10.047

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., Wiese, E., Prosser, A., & Müller, H. J. (2014). Beliefs about the Minds of Others Influence How We Process Sensory Information. *PLoS ONE, 9*(4), e94339. https://doi.org/10.1371/journal.pone.0094339

Yeung, N. & Stanfey, A. G. (2004). Indeopendent coding of reward magnitude and valence in the human brain. *Journal of Neuroscience, 24,* 6258–6264.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.