2023

Are human-like robots trusted like humans? An investigation into the effect of anthropomorphism on trust in robots measured by expected value as reflected by feedback related negativity and P300

Wilson, L.O.

Wilson, L.O. (2023) 'Are human-like robots trusted like humans? An investigation into the effect of anthropomorphism on trust in robots measured by expected value as reflected by feedback related negativity and P300', The Plymouth Student Scientist, 16(2), pp. 347-376. https://pearl.plymouth.ac.uk/handle/10026.1/21834

The Plymouth Student Scientist
University of Plymouth

*The Plymouth Student Scientist, 2023, **16**, (2), 347-376*

Are human-like robots trusted like humans? An investigation into the effect of anthropomorphism on trust in robots measured by expected value as reflected by feedback related negativity and P300

Lucy Olivia Wilson

Project Advisor: Prof Jeremy Goslin, School of Psychology, University of Plymouth, Drake Circus, Plymouth, PL4 8AA

Abstract

Robots are becoming more prevalently used in industry and society. However, in order to ensure effective use of the trust, must be calibrated correctly. Anthropomorphism is one factors which is important in trust in robots (Hancock et al., 2011). Questionnaires and investment games have been used to investigate the impact of anthropomorphism on trust, however, these methods have led to disparate findings. Neurophysiological methods have also been used as an implicit measure of trust. Feedback related negativity (FRN) and P300 are event related potential (ERP) components which have been associated with processes involved in trust such as outcome evaluation. This study uses the trust game (Berg et al., 1995), along with questionnaires and ERP data to investigate trust and expectations towards three agents varying in anthropomorphism, a human, an anthropomorphic robot, and a computer. The behavioural and self-reported findings suggest that the human is perceived as the most trustworthy and there is no difference between the robot and the computer. The ERP data revealed a robot driven difference in FRN and P300 activation, which suggests that robots violated expectations more so than a human or a computer. The present findings are explained in terms of the perfect automation schema and trustworthiness and dominance perceptions. Future research into the impact of voice pitch on dominance and trustworthiness and the impact of trust violations is suggested in order to gain a more holistic picture of the impact of anthropomorphism on trust.

Keywords: Trust, Anthropomorphism, Robots, EEG, FRN, ERP, Trustworthiness, Likeability, Dominance, Intelligence, Investment Game, Trust Game, Trust Violation.

Ethical Statement

This study was granted full ethical approval by the Faculty of Health and Human Sciences at the University of Plymouth and adheres to the current BPS guidelines (British Psychological Society, 2022). Prior to taking part in the study, all participants were fully briefed and informed of their right to withdraw. All participants were required to sign a consent form (Appendix A) before being allowed to participate in the study and were given a debrief form following the experiment (Appendix B) Numerical values were assigned to participants to protect the anonymity of their data. All raw data was kept in a secure file only accessible to myself (Lucy Olivia Wilson) and Dr Jeremy Goslin. All data was handled in accordance with the Data Protection Act (1998).

The procedures and stimuli used in this study posed minimal risk to the participants. In application of the EEG cap saline gel was used. Participants were told to inform the researcher of any discomfort. All apparatus was cleaned and disinfected between uses.

Introduction

Robots are becoming more prevalently used in day to day life. Labour that used to be done exclusively by humans is shifting to machines or robots (Brynjolfsson & McAfee, 2014). While some machines are very simple and transparent in their operation, other machines have become more complex and the way in which they work has become more opaque. This lack of knowledge leaves a gap which needs to be filled with trust that the machine will work as the user expects it to. In current technology, probably one of the most complex machines are robots.

In 2021, 3 million industrial robots were operational with 384,000 units shipped in 2020, which is expected to rise to 500,000 by 2024 (World Robotics Report, 2021). These machines can be found in various sectors, included but not limited to, manufacturing (Edwards, 1984), healthcare (Schäfer et al., 2019; Guizzo & Goldstein, 2005; Heerink et al, 2010; Bemelmans et al., 2012; Pennisi et al., 2016; Chang & Kim, 2013), education (Belpaeme et al., 2018; Chang et al, 2010), the military (Schaefer et al., 2014) and hospitality (Gombolay et al., 2016, Salem et al., 2015, Tolmeijer et al., 2021). However, robots are ineffective if they are not used due to lack of trust or if they are over trusted and used in inappropriate conditions (Muir, 1994, Parasuaraman, & Riley, 1997; Hancock et al, 2011).

Trust, at its most basic level, is the willingness of a party to be vulnerable to the actions of another, despite the ability to control or monitor the other party (Mayer, Davis & Schoorman, 1995). Trust can be characterized by the trustors beliefs (Hoff & Bashir, 2015) and the behavioural manifestations of trust (Lee & See, 2004; Madhavan & Wiegmann, 2007). In addition, the outcomes of trusting behaviours can impact the trustors beliefs about the trustee (Mayer et al., 1995; King-Casas et al., 2005). Trust is a fundamental part of how humans interact with other humans and machines (Hoff & Bashir, 2015; Lee & See, 2004; Hancock et al, 2011; Lee & Seppelt, 2009). The mechanisms which underly human-human trust can be extended to automation. For example, the antecedents of trust in humans are ability, benevolence and integrity (Mayer et al., 1995), these have been replaced with similar concepts, performance, purpose and process for trust in robots (Lee & See, 2004; Lee & Moray 1992), while keeping the rest of the model the same. In addition,

a review revealed, humans tend to react socially to machines as they do with humans, however, there are subtle differences in the way that humans perceive other humans and machines (Madhaven & Wiegmann, 2007). There are two dominant paradigms used to explain how humans perceive machines, the unique agent hypothesis (UAH) and computers as social actors (CASA).

UAH suggests humans view machines differently to humans based on the characteristics of the machine (de Visser et al, 2016), therefore, each agent and situation respectively is unique and may or may not evoke similar mental models to that of human-human interaction (Hoff & Bashir, 2015). However, CASA suggests humans attribute humans-like characteristics onto machines, applying social rules and expectations to automation (Nass & Moon, 2000; Rai and Diermeier, 2015), this is known as anthropomorphism (Duffy, 2003) and is an important factor in human-robot trust (Hancock et al, 2011; Hancock et al, 2021; Natarajan & Gombolay, 2020). However, the interaction between anthropomorphism and robots is complicated. Wang & Quadflieg (2015) suggests robots are viewed as less believable, intelligent and capable of experiencing emotions, as well as more eerie. There are an number of factors which influence whether humans anthropomorphise machines, including past experience, behaviour of the machine and how the machine looks. For example, human-like communication in intelligent virtual assistants such as Siri leads to increases in anthropomorphism (Sin & Munteanu, 2020; Guzman). Research shows humans have higher expectations of, and have more trust in, more anthropomorphic agents (Hinds et al, 2004; Klüber & Onnasch, 2022; Yang et al., 2022; Waytz et al., 2014; Chen & Park, 2021; Pak et al, 2012; de Visser et al, 2012). Human-human trust and human-computer trust have similarities and differences. Understanding the role of anthropomorphism in modulating human-machine trust is a key issue. This experiment will compare human-human trust, with trust in a computer and a humanlike robot to investigate the impact of anthropomorphism. Within the current literature this issue has primarily been explored through the use of questionnaires and behavioural data.

The trust in Automation Scale (TAS, Jian, Bisantz, & Drury, 2000) and the Trust Perception Scale-HRI (TPS-HRI, Schaefer, 2013) are two examples of questionnaires measuring trust. Some findings investigating trust in robots suggest that anthropomorphism increases self-reported trust (Kulms & Kopp, 2019; Jensen et al., 2021; Alarcon et al 2023; Tulk & Weise 2018), while other researchers suggest anthropomorphism has no effect on self-reported trust (Alarcon et al., 2023; Alarcon et al., 2021; Jain et al., 2022). The differences may be due to some of the issues with questionnaire data, for example, they are subject to biases. These impact the way participants behave or respond to stimuli and can lead to demand characteristics and discordant findings (Kessler et al., 2017). In addition, questionnaires alone do not allow trust to be measured within social contexts, instead they rely on participants making an subjective judgement of how they trust robots without having any experience within a social context. There are a number of different ways to measure trust implicitly, without the use of questionnaires. One of the most commonly used methods is through the use of economic games, for example the trust game (Berg et al., 1995).

The trust game (Berg et al, 1995) has been used to investigate social aspects of human-human trust such as gender (Buchan et al, 2008), race (Simpson et al, 2007)

and personality (Evans & Revelle, 2008). In the trust game participants are endowed points and can choose how much they wished to invest, then partner returns a percentage of this original investment. When the outcome is revealed it can be different to what was expected, the extent of the difference is known as the reward prediction error and it is used to evaluate whether or not to continue trusting someone (King-Casas et al., 2005). The trust game uses investments as an implicit measure of trust. The assumption is that player A will only invest more if they trust player B will reciprocate and return a portion of that money. Recently the trust game has also been used when investigating human-robot trust, finding that there is no impact of anthropomorphism on investment behaviour (Tulk & Wiese, 2018; Kulms & Kopp, 2019; Alarcon et al., 2021; Alarcon et al., 2023). Investments, in the trust game, suggest anthropomorphism has no effect on trust, however this contrasts with the split view in self-reported data. The use of neurophysiological methods has been suggested (Drnec et al., 2016; Parasuraman, 2003) allowing implicit measures of trust to be recorded alongside behavioural measures, such as the trust game.

Functional magnetic resonance imagery (fMRI) and electroencephalography (EEG) are two neurophysiological methods which have been used to investigate the spatial and temporal neurological origins of trust. fMRi data suggests activity in the caudate nucleus (King-Casas et al., 2005), anterior paracingulate cortex (McCabe et al., 2001) and orbitofrontal cortex (Krain et al., 2006) is associated with aspects of outcome and trust evaluation. fMRI data also more reliably distinguishes between trust and distrust compared to self-reported data (Dimoka, 2010). EEG results reveal alpha and beta band power have higher power in situations of trust, whereas gamma band power is stronger in situations of mistrust (Oh et al, 2017; Blais et al, 2018). Within event related potential (ERP) research two main ERP components have been associated with trust, feedback related negativity (FRN) (Long et al., 2012; Wang et al., 2016) and P300 (Bell et al, 2016; Wang et al, 2015).

In seminal work, Holroyd and Coles (2002) reported FRN, this a differential wave recorded at mid-central sites and peaks between 200ms to 300ms after feedback onset, sensitive to positive or negative feedback (Nieuwenhuis et al., 2004). FRN is thought to be associated with reward prediction errors and reinforcement learning (Holroyd & Coles, 2002). In addition mid-brain regions associated with FRN have been liked to reinforcement learning, reward predictions, outcome evaluation (Barto, 1995; Montague, Dayan & Sejnowski, 1996) and adapting future behaviour (Schultz, 2002), which are features of calibrating trust (Mayer et al., 1995). In a study involving a coin-flip task, FRN amplitude increased following trust decisions, which was significantly related to outcome evaluation (Long et al., 2012). The research suggests FRN is larger when participants decide to trust (Hu et al., 2018), has a greater amplitude in response to loss feedback (Wang et al., 2015; Hu et al., 2018), when an outcome is worse than expected (Holroyd & Coles 2002; Nieuwenhuis et al., 2004; Schultz et al., 1997) and is driven by the lack of an expected reward (Wu & Zhou, 2009; Holroyd & Coles 2002; Wang et al., 2015; Sambrook & Goslin, 2015). Therefore, a larger FRN amplitude will be expected following a trust decision which lead to unexpected negative feedback compared to expected positive feedback. In the context of this experiment, a larger FRN amplitude would be expected when a human returns less than was invested (loss trial) as this is worse than expected outcome. The human is the most anthropomorphic agent and therefore they will be expected to return the most and when the they fail to meet these expectations a

larger FRN will be seen. The amplitude of the FRN on loss trials will decrease as the agent becomes less anthropomorphic, such that a robot will have a slightly smaller FRN amplitude compared to the human and the computer will have a smaller FRN amplitude compared to the robot as expectations of these agents decreases.

P300 is a positive component, which peaks around 200ms-600ms after feedback presentation. The scalp distribution of P300 is noted as the amplitude change across midbrain electrodes, typically P300 increases in magnitude from frontal to parietal sites (Johnson, 1993). P300 is linked to outcome expectation (Hajcak et al., 2005, Wu and Zhou, 2009), outcome evaluation (Yeung and Sanfey, 2004), decision making (Nieuwenhuis et al., 2005; Yeung & Sanfey, 2004) and updating mental representations (Donchin, 1981; Heslenfeld, 2003) which are key components of trust formation (Mayer et al., 1995). Research investigating trust has noted that the P300 component's amplitude is more positive following trust decisions (Long et al., 2012), gain feedback (Yeung & Sanfey, 2004; Hajcak et al., 2005; Holroyd et al., 2016; Wu & Zhou, 2009) and unexpected feedback (Hajcak et al., 2005; Hajcak et al., 2007; Wu & Zhou, 2009). Therefore, a larger P300 amplitude is expected following a trust decision that lead to unexpected positive feedback, compared to expected negative feedback. In this experiment a more positive P300 would be expected when a computer returns more than the initial investment as the computer is the least anthropomorphic and therefore the participant expects the least from it, therefore when it violates this expectation a more positive P300 amplitude will be seen. The P300 amplitude on gain trials will decrease as the partners become more anthropomorphic such that the P300 amplitude will be slightly smaller for robots compared to computers, and smaller for humans compared to robots as the participant will be expecting more from these agents.

The aim of this study is to use electrophysiological correlates, related to trust, to examine the neurophysiological correlates of trust and how they may be modulated by anthropomorphism. This study will use questionnaire data, implicit measures of trust through investments in the trust game, and electrophysiological correlates to understand the mechanisms of trust and its relationship to anthropomorphism. Based on the previous research the following hypotheses are made.

Hypothesis 1: There will be no difference in investment behaviour between the human, the pepper robot and the computer. (Kulms & Kopp, 2019; Alarcon et al., 2021; Alarcon et al., 2023; Tulk & Weise, 2018)

Hypothesis 2: Humans will be perceived as most trustworthy, followed by the pepper robot and lastly the computer (Kulms & Kopp, 2019; Jensen et al., 2021; Tulk & Wiese, 2018)

Hypothesis 3: There will be a more negative going FRN for loss trials with the humans, followed by a less negative FRN on lost trials with the pepper robot and the least negative FRN on lost trials for the computer. (Hinds et al., 2004; Hüber & Onnasch, 2022; Yang et al., 2022; Wang et al., 2015; Hu et al., 2018; Wu & Zhou, 2009; Holdroyd & Coles, 2002; Sambrook & Goslin, 2015)

Hypothesis 4: P300 will have a larger amplitude on gain trials for the computer, followed by a smaller P300 amplitude on gain trials for the pepper robot, and the

least positive P300 component on gain trials for the human (Hinds et al., 2004; Hüber & Onnasch, 2022; Yang et al., 2022; Sato et al., 2005; Yeung & Sanfey, 2004; Hajcak et al., 2005, 2007; Yeung et al., 2005; Wu & Zhou, 2009)

Methodology

In this experiment, each participant was invited to play a game with a human, a computer and a robot. The order of presentation of these partners was counterbalanced. In the human condition participants played with the researcher, in the computer condition they played with the computer, and in the robot condition they played with a pepper robot (SoftBank, 2014). The game used was an investment game in which participants were endowed points and could choose how much they wished to invest, this was used as an implicit measure of trust. The partner then returned a percentage of this original investment. The responses in this game, in each case were scripted, and the scripted payback was also counterbalanced between conditions. Following each game, the participants were asked to fill out a questionnaire regarding their perceptions of the partner they had played with. Throughout this procedure the participants EEG was recorded.

Participants:

45 participants were recruited for this study (F=31, M= 13, O= 1), however, due to technical issues, the data from 6 participants was not recorded (F=4, M=2). Therefore, in the final data set there were 39 participants (F= 27, M= 11, O= 1) for this experiment, all of which were aged 18 to 27 years old and students at the University of Plymouth. These participants were recruited via the Plymouth Psychology Participant Pool, using the SONA system (S, Systems, 2022). All participants had normal or correct-to-normal vision and were right handed.

Procedure:

Participants were asked if they had any allergies then following this were asked to read and sign a consent form. After this the EEG cap was applied to their head using the 10-20 system. Once the cap had been applied, the instructions for the trust game were presented on the computer screen, which the participants could read through at their own pace, and participants were asked if they had any questions. Following this the participants would start playing the game with the partner, the human, the computer or the pepper robot. The robot and human were situated to the right side of the monitor when playing the game, within the visual field of the participant. In this trust game, modified from Berg et al (1995), participants are endowed with 2 points each trial. They have the option of investing 0, 1 or 2 of these points into the partner, giving their responses on a key pad kept in their hands for the whole of the experiment. If the participant invested 0 points the screen turns red, 1 point the screen turns blue and 2 points the screen turns green. There was then a blank screen for 4200 to 4600ms. This was followed by a screen which states how many points the player banked and invested and that they were waiting for the partner to return the investment for 2000 to 6000ms. At this point the robot would make a random happy or thinking animation and say ‘Thank you for investing X in me this round, let me see what I can return’, the computer would present the same statement on the screen and the human would make some similar thinking movement and then press a button on their keypad. The counterpart’s reciprocation

strategy was manipulated by the experimenter (50% reinforcement rate). All return percentages were preprogramed with the same probability of returning a round percentage between 50%-150%, there were three orders for these percentages and the order of returns was counterbalanced along with the order of the investors. Following this there was a fixation cross which was on the screen for between 2000-2500ms before the percentage returned was on the screen for 1500ms. After the percentage was presented, there was another fixation point on the screen for 1000ms followed by the actual number of points returned for 1500ms. To complete the trial the screen returned to black and there was a summary of the total points the participant had in the bank, the participant the pressed enter to continue to the next trail. There were 45 trials per game. After the completion of the trust game participants were asked to fill out a generic trust questionnaire about their perception of the partner they were playing with.

EEG recording, processing and analysis:

EEGs were recorded from 64 scalp sites using tin electrodes mounted in an elastic cap (actiCHamp Plus, Brain Products GmbH, Gilching, Germany) following the international 10 –20 system. Vertical electro-oculograms (EOGs) were recorded from the underneath the right eye which was cleaned using alcohol before application. Electrode impedance was kept below 10 k Ω for all channels. All electrodes were referenced using the left mastoid electrode, we then re-referenced offline to an average of the left and right mastoid electrodes.

The FPz electrode was used to ground. ERPs were time-locked on the visual onset of the presentation partner's investment return with 1000msec time window spanning from 200 to 800msec before and after the time-lock. A fully-automatic trial rejection procedure was run on these ERPs to exclude segments violating the following parameters: maximum or minimum voltages of +200μV and 150μV respectively, a maximal voltage difference of 200μV over a 100 msec interval window, and minimum amplitude of 0.5μV within 100 msec intervals. To maximize the signal/noise ratio, these parameters were slightly manually adapted for each participant and leading to 15% of segments rejected. Any individual electrodes which had more than ~8% of rejected segments were substituted with topographically interpolated replacements (Perrin et al., 1989), over all of the participants 1.4% of the electrodes replaced.

A pairwise comparison, based on a cluster randomisation technique (Maris & Oostenveld, 2007), was used to conduct a the statistical analysis of ERP components. For the whole time-window, two-tailed t-tests were performed, each electrode-time and electrode-signal sample pair were compared for the partners, separately for positive or negative returns. Samples with a t statistic above the significance threshold of $p < .05$ were clustered together with regard to their spatial and temporal features. Each cluster included a minimum of eight samples and was subsequently used for the cluster analysis. The cluster-level t statistic was calculated as the sum of the t statistic of all electrode-time samples of a given cluster. The cluster with the largest t statistic was selected for a Monte-Carlo simulation. During this, the original pairs of t-tests sample that made up the cluster were repeated 1000 times, with arrangements of each paired samples assigned randomly to the human, robot or computer with either a positive or negative return. This created a Monte-Carlo distribution of summed t statistic corresponding to the null hypothesis. A Monte-Carlo p-value was calculated as the ratio of the 1000 summed t statistics in the random distribution that was above the cluster-level t statistic. A p-value above p

< .025 was considered significant. Averaged ERP were re-plotted as t-values in the time domain, derived from t-tests against baselines of zero. Topographic maps were created using the t-values of the significant cluster in Brain Vision Analyzer (Brain Products, Munich, Germany, v. 2.1), using spherical spline interpolation with an order of splines of 5 and a maximum degree of Legendre polynomials of 10 in order to smooth the map.

Results

Within this study three forms of data were collected. Behavioural data, this consisted of investments within the trust games, self-reported data, a questionnaire regarding perceptions of each partner after playing the trust game, and finally EEG data, ERP's following feedback presentation. The behavioural and self-reported data was analysed using Jamovi (The jamovi project, 2022; R Core Team, 2021; Singmann, 2018; Lenth, 2020) and the EEG data was analysed using Brain Vision Analyzer (BrainVision Analyzer, Version 2.2.2, Brain Products GmbH, Gilching, Germany).

Behavioural results:

The mean investments were 1.36 for Human, 1.18 for Robot, and 1.23 for Computer (see figure 1). To analyse any differences between the three conditions a repeated measures ANOVA was used. This analysis found there was a significant difference in investments between the three conditions ($F=76, 2) = 7.95$, $p<.001$). The post hoc test, Ptukey, revealed main effects were found for comparison between human-robot ($p=0.002$) and for human-computer ($p=0.026$), however, not for robot-computer ($p=0.509$). This goes against the first hypothesis, that there would be no differences in behavioural trust between the three conditions, as participants invested statistically more in the human partner compared to both the human and the robot. However, there was no statistically significant difference between investments for the robot and computer partners.

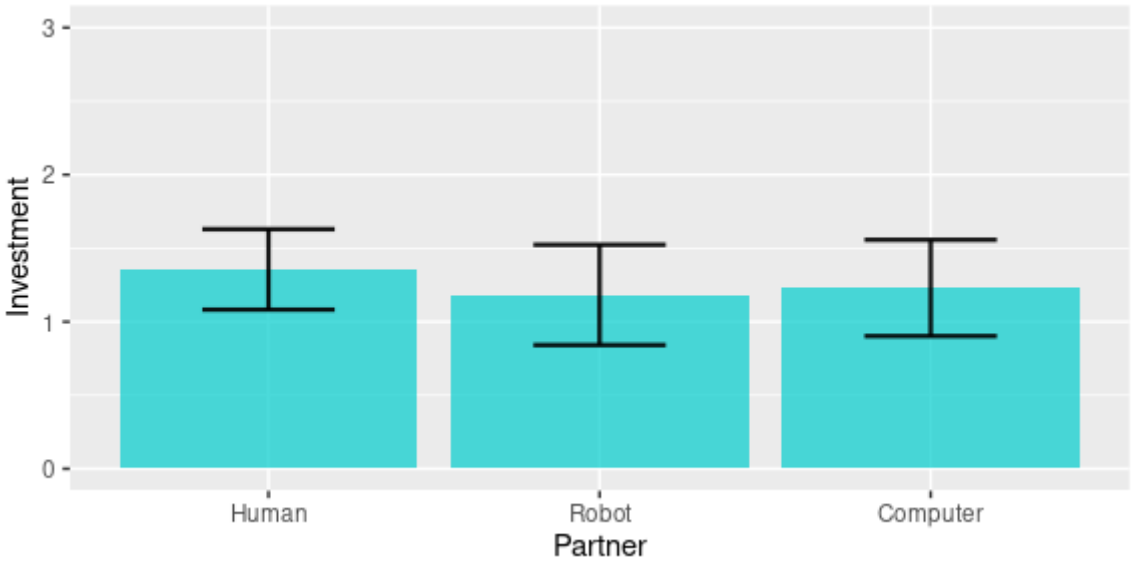

Figure 1: Mean investments with error bars showing standard deviation

Self-reported measures:
In this experiment, participants were asked to complete a questionnaire after each game. This questionnaire measured perceptions of trustworthiness, likeability, intelligence and dominance. Although this study primarily investigates perceived trust, in line with previous research the peripheral traits were also investigated (Calvo-Barajas et al., 2020; Kim et al., 2020; Kraus et al., 2018; McAleer et al., 2014).

Trustworthiness:
Mean self-reported measures of trustworthiness were 5.29 for human, 4.11 for robot and 3.97 for computer (See figure 2). To analyse any differences between the three conditions a repeated measures ANOVA was used. This analysis found there was a significant difference in perceived trustworthiness between the three conditions ($F(76, 2) = 17.8$, $p<.001$). The post hoc test, Ptukey, revealed main effects were found for comparison between human-robot ($p=<.001$) and for human-computer ($p=<.001$), however, not for robot-computer ($p=0.827$). This partially supports the second hypothesis, that anthropomorphism would increase self-reported trust, as the more anthropomorphic agent, the human, was perceived as more trustworthy. However, the robot was not statistically perceived as being more trustworthy than the computer, therefore not fully supporting hypotheses 2 as even though the robot was more anthropomorphic it was not perceived as any more trustworthy than the computer.

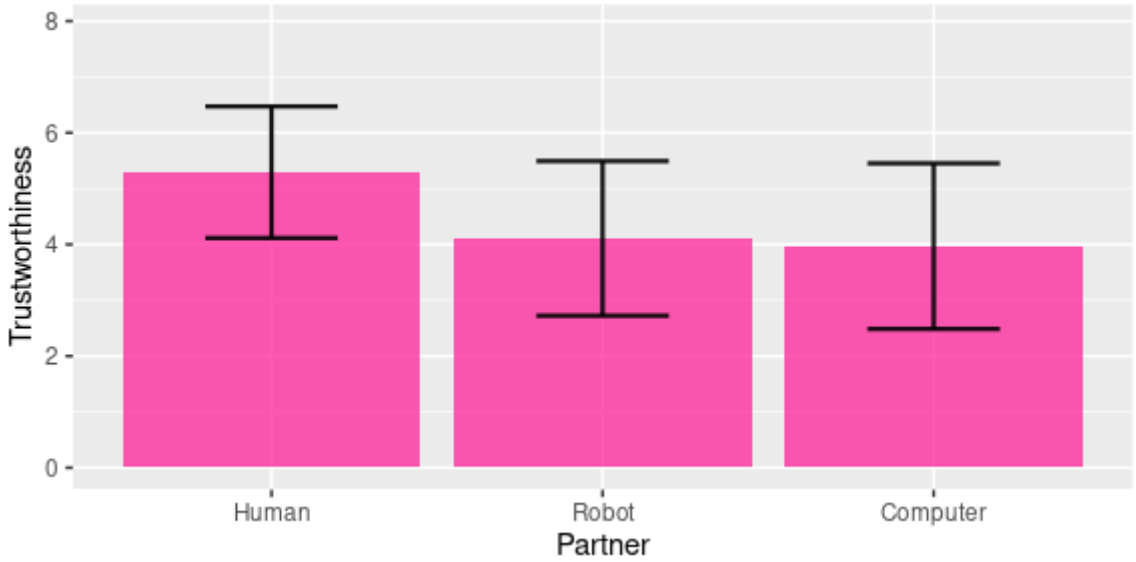

Figure 2: Mean self-reported trust scores with error bars showing standard deviation

Peripheral self-reported measures:

Likeability:
Mean scores for perceived likeability were 5.98 for human, 5.07 for robot and 3.82 for computer (see figure 3). To analyse any differences between the three conditions a repeated measures ANOVA was used. This analysis found there was a significant difference in perceived likeability between the three conditions ($F(76, 2) = 36.8$, $p<.001$). The post hoc test, Ptukey, revealed there were main effects for all three comparisons, human-robot ($p= 0.002$), human-computer ($p=<.001$) and robot-computer ($p=<.001$).

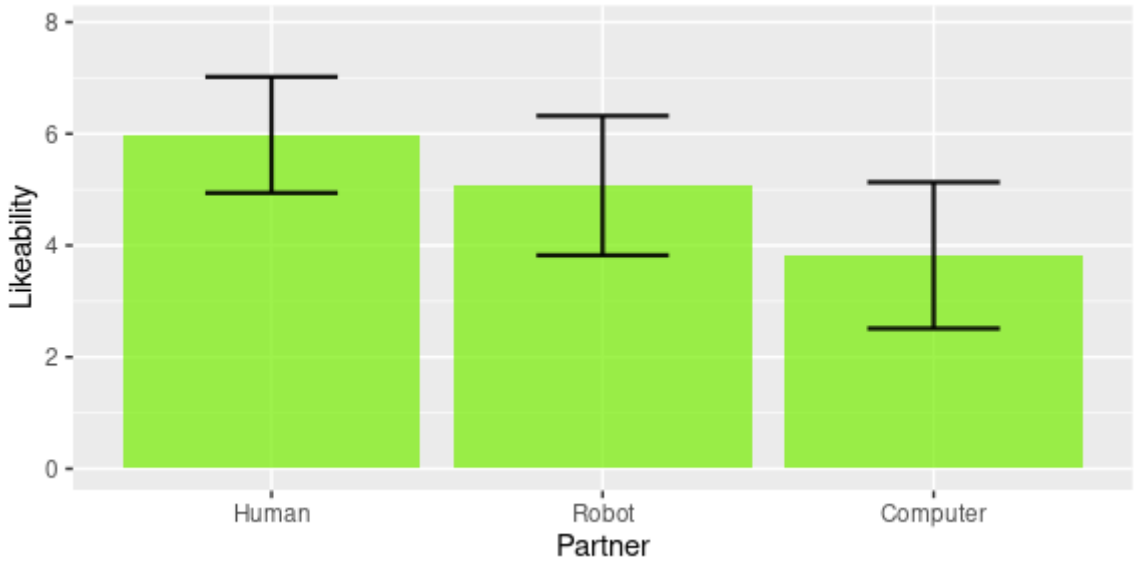

Figure 3: Mean self-reported likeability scores with error bars showing standard deviation

Intelligence:
Mean self-reported measures of intelligence were 5.42 for human, 4.98 for robot and 4.72 for computer (See figure 4). To analyse any differences between the three conditions a repeated measures ANOVA was used. This analysis found there was a significant difference in perceived intelligence between the three conditions ($F(76, 2) = 12.3$, $p<.001$). The post hoc test, Ptukey, revealed main effects were found for comparison between human-robot ($p=0.004$) and for human-computer ($p=<.001$), however, not for robot-computer ($p=0.195$).

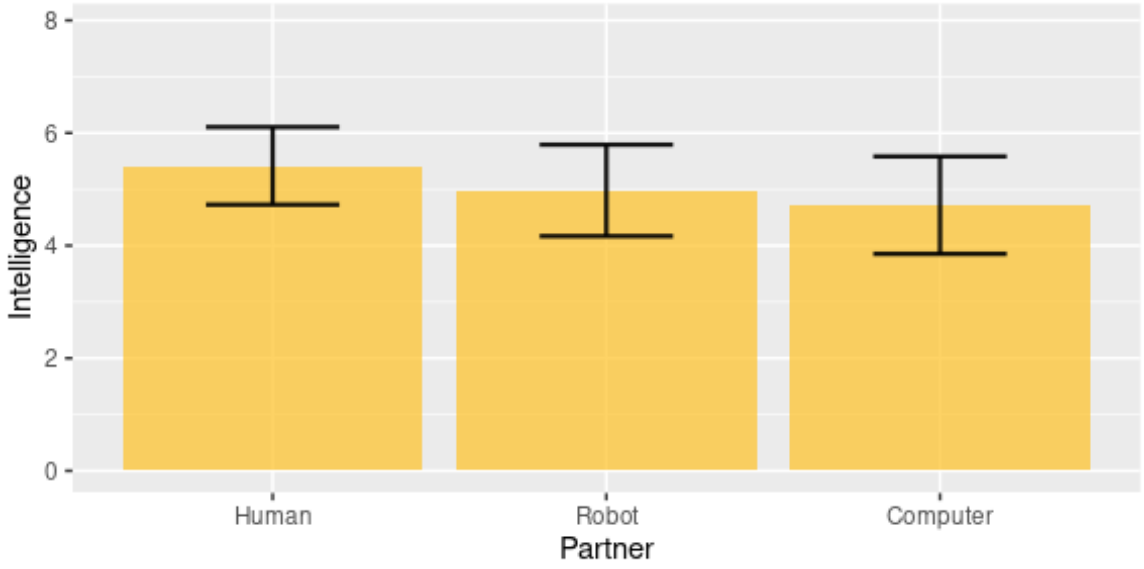

Figure 4: Mean self-reported intelligence scores with error bars showing standard deviation

Dominance:
Mean self-reported measures of dominance were 3.00 for human, 3.28 for robot and 3.80 for computer (See figure 5). To analyse any differences between the three conditions a repeated measures ANOVA was used. This analysis found there was a significant difference in perceived dominance between the three conditions (F(76, 2) = 4.30, p= 0.017). The post hoc test, Ptukey, revealed main effects were found for comparison between human-computer (p=0.005), however, not for human-robot (p=0.650) or robot-computer (p=0.150) .

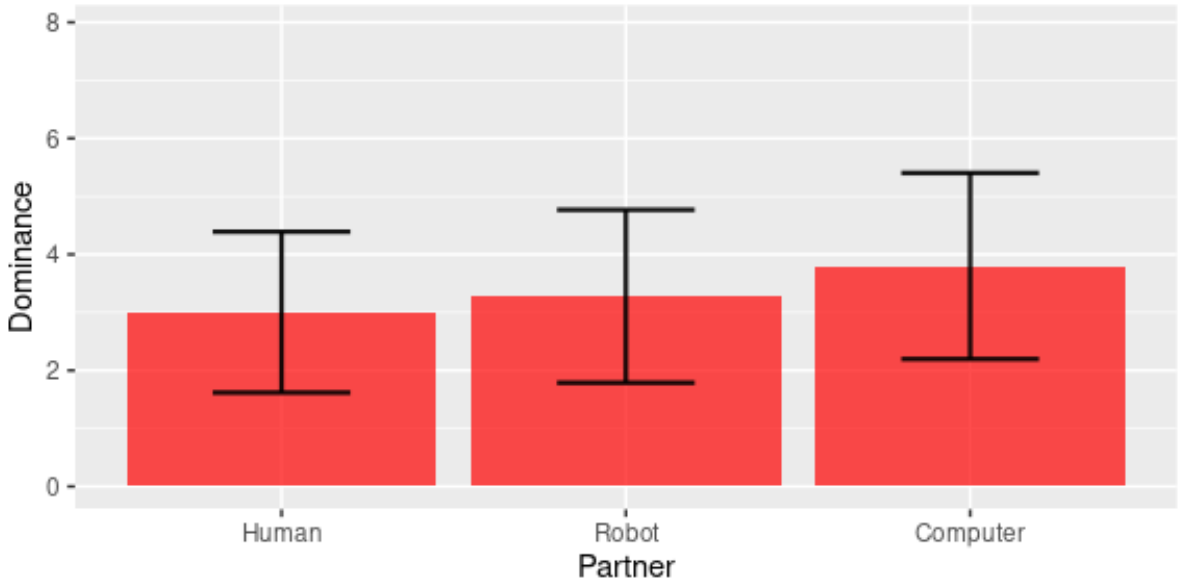

Figure 5: Mean self-reported dominance scores with error bars showing standard deviation

EEG Results:
ERP responses to feedback were sorted into six groups and entered into statistical analysis: human gain – robot gain, human gain – computer gain, computer gain – robot gain, human loss – robot loss, human loss – computer loss, computer loss – robot loss. The data was split like this to investigate any differences between the agents. By splitting the valence of the outcome into gain and loss, inferences can be made about the initial expectation of the agent. For example, if there is a larger FRN for gain trials for robots compared to humans, then it would suggest that they were expecting more from the robot and the actual outcome was worse than expected. If there is a larger FRN for humans on loss trials then it would suggest that they were expecting more from the human as the actual outcome was worse than expected. In terms of P300, if there was a larger P300 for computers on gain trials compared to a robot it would suggest that this is an unexpected outcome. These findings would support the hypotheses regarding FRN and P300 as the violation in expectation, as reflected by FRN and P300, would suggest differences modulated by anthropomorphism.

Analyses of these 6 comparisons were completed using the cluster randomization technique (Maris & Oostenveld, 2007). These analyses revealed a number of significant areas of differences between conditions, all of which were below the significance threshold at p=0.025.

Human gain – Robot gain:
Within this comparison, of human minus robot activation on gain trials, there was one cluster identified. This was a positive difference in activity between 196-544ms across central and right posterior sites as shown in figure 6. The distribution of the temporal and spatial significance suggests an early modulation of the FRN followed by a later modulation of P300.

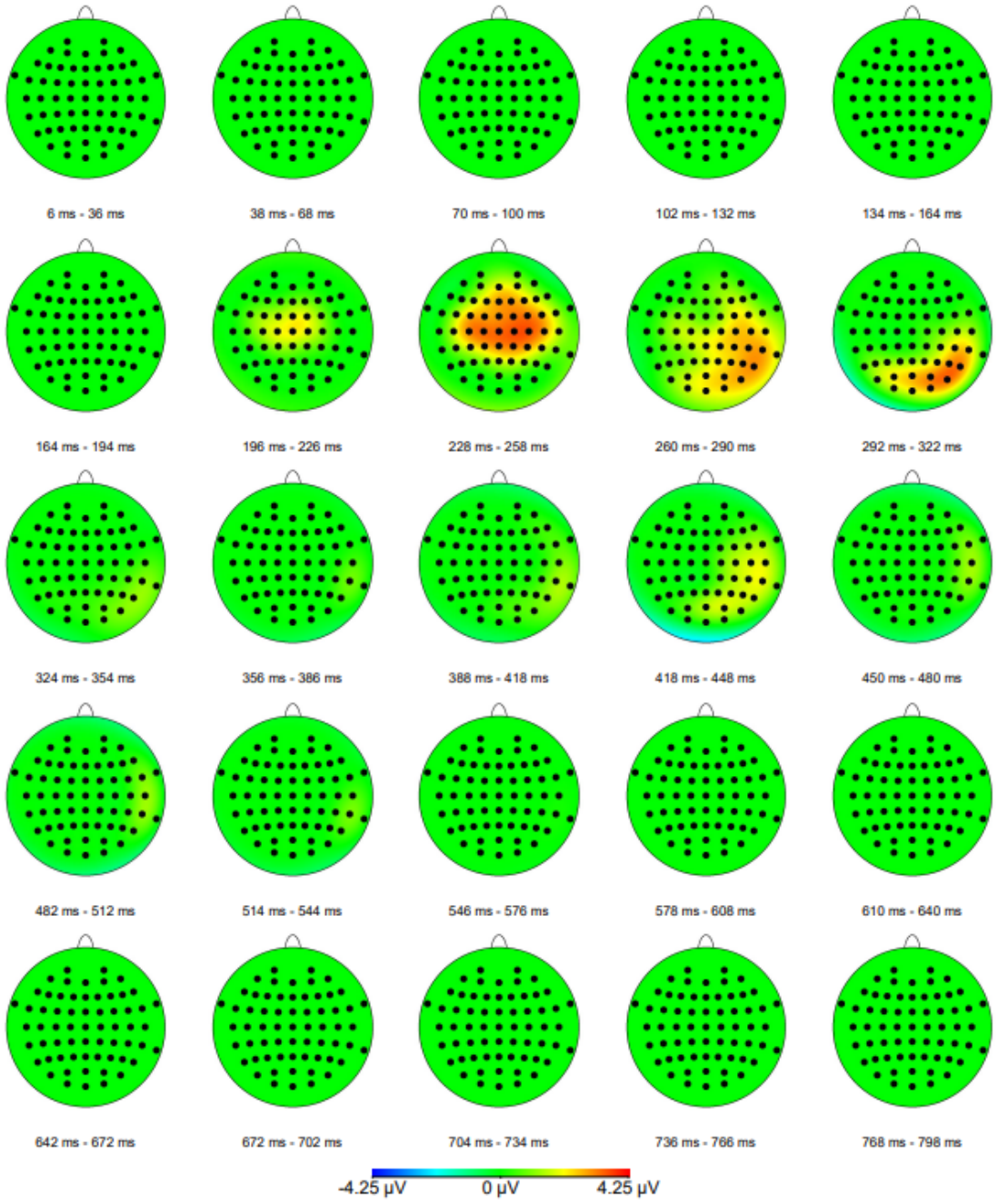

Figure 6: Human gain – robot gain topographical map.

Human gain – Computer Gain:
Within this comparison, of human minus computer activation on gain trials, there was one cluster identified. This was a positive difference in activity between 258-352ms across occipital sites as shown in figure 7. This does not have the spatial or temporal significance to be an FRN or P300.

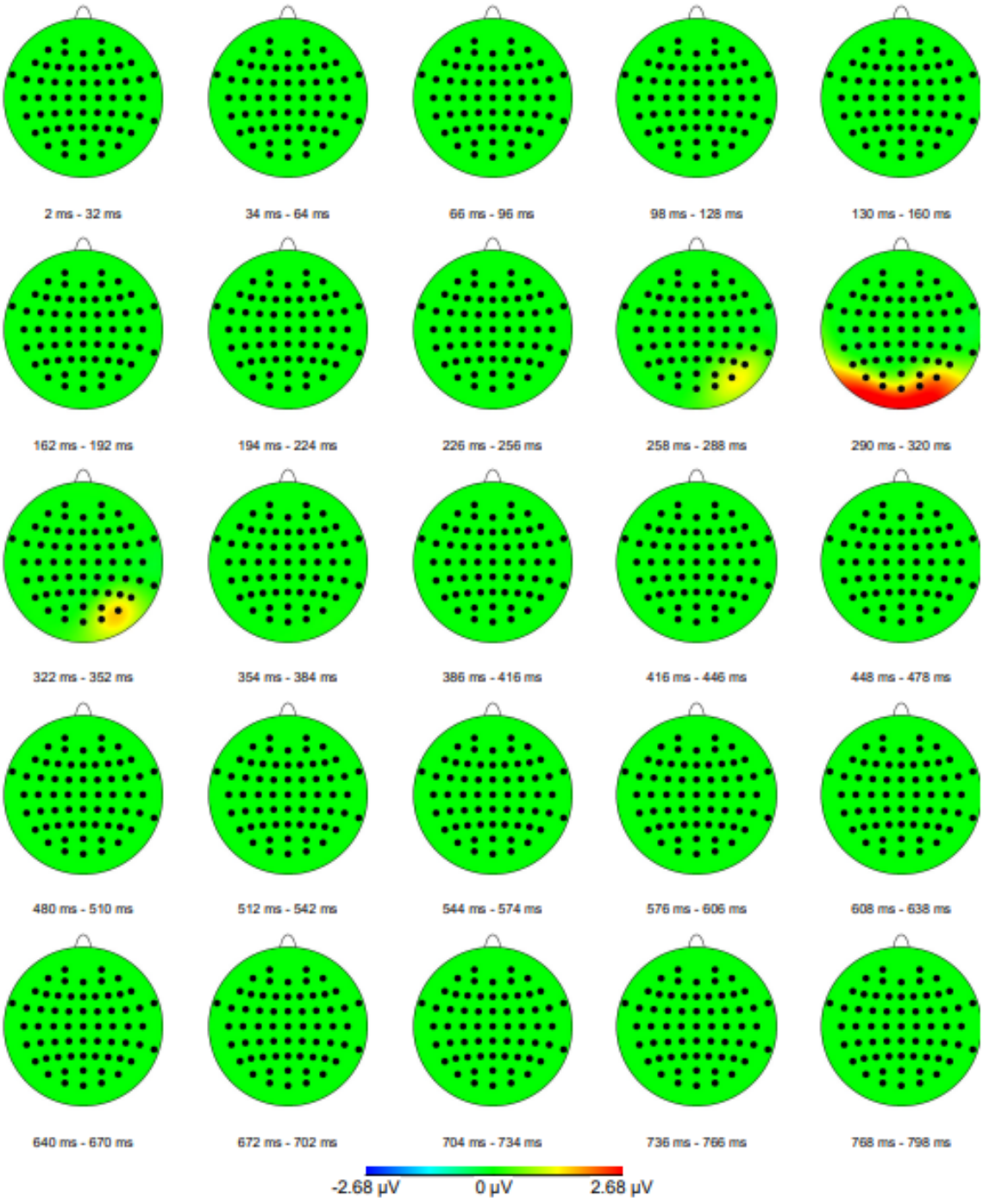

Figure 7: Human gain – computer gain topographical map.

Computer gain – robot gain:
Within this comparison, of computer minus robot activation on gain trials, there were two clusters identified. As shown in figure 8, the first is a positive difference in activity between 220-284ms across central sites, which has the temporal and spatial significance to suggest a modulation of the FRN. The second cluster was a negative difference in activity between 284-510ms across left anterior sites, however this does not have the temporal or spatial significance to suggest an FRN or P300.

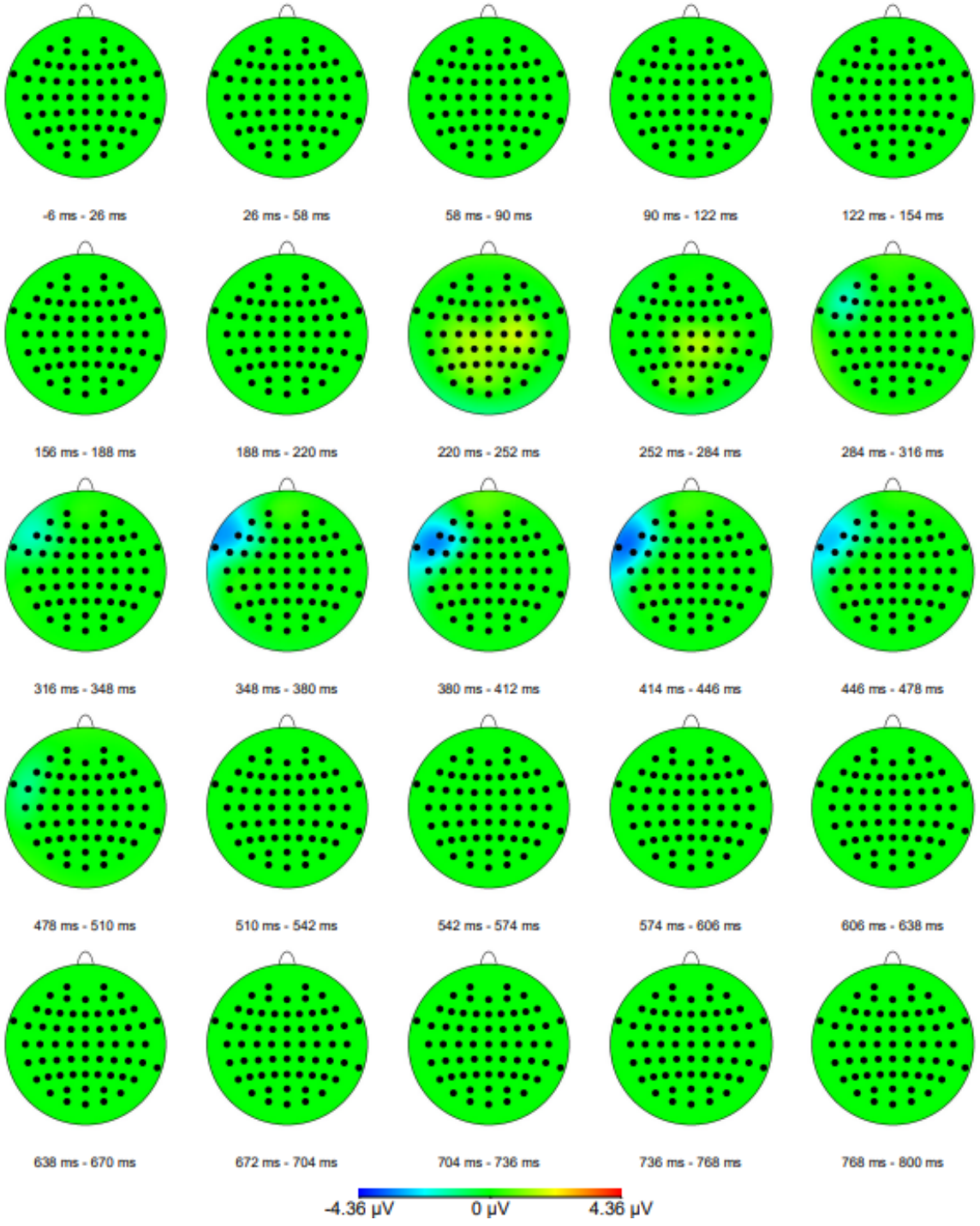

Figure 8: Computer gain – robot gain topographical map.

Human loss – robot loss:
Within this comparison, of human minus robot activation on loss trials, there was one cluster identified. As shown in figure 9, this was a positive difference in activity between 480-574ms across central posterior sites, however, the temporal and spatial significance does not suggest it is an FRN or P300.

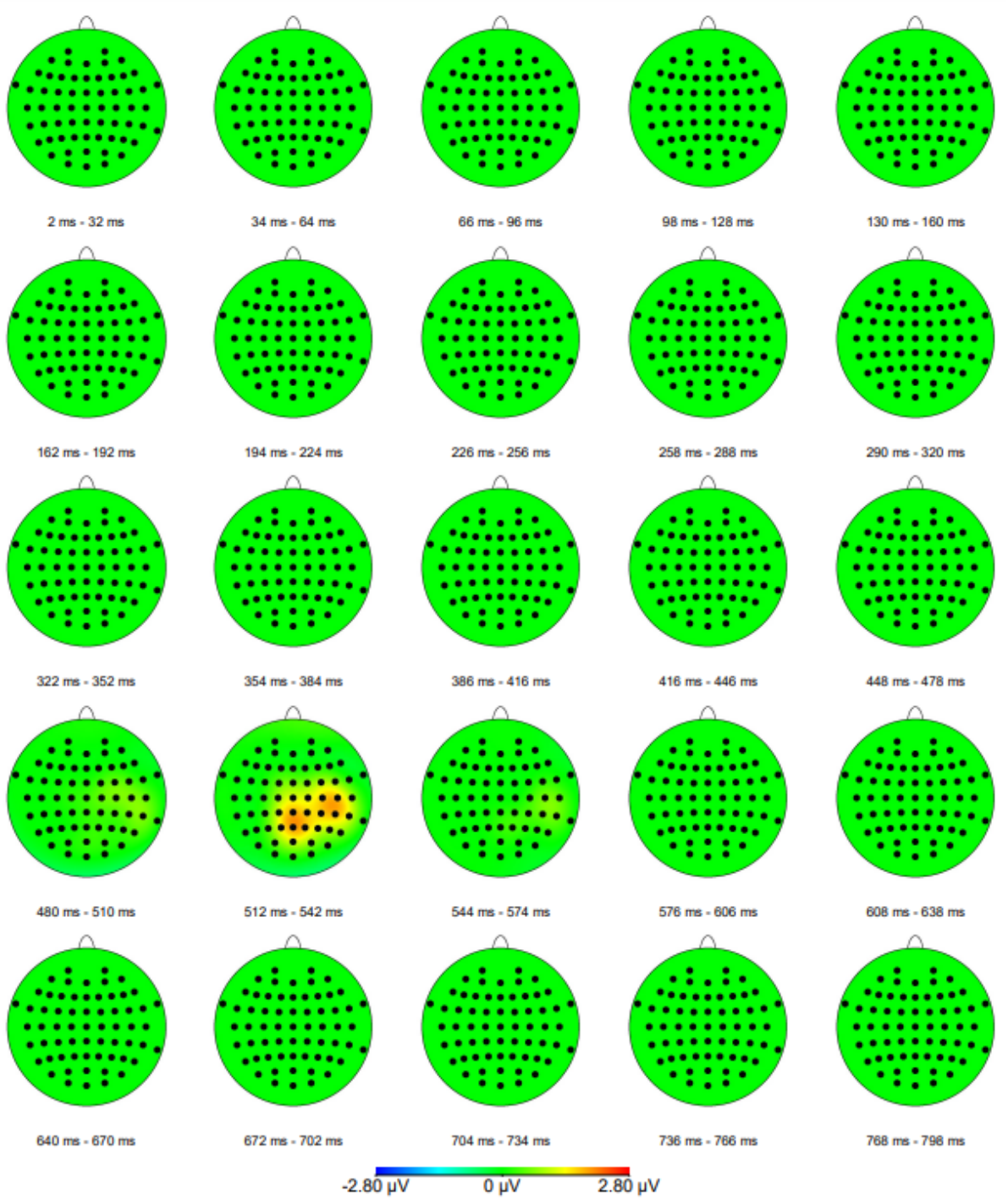

Figure 9: Human loss – robot loss topographical map.

Human loss – Computer loss:
Within this comparison, of human minus computer activation on loss trials, there was one cluster identified. As shown in figure 10, this was a negative difference in activity between 348-412ms across occipital sites, this does not have the temporal or spatial significance to be either an FRN or P300.

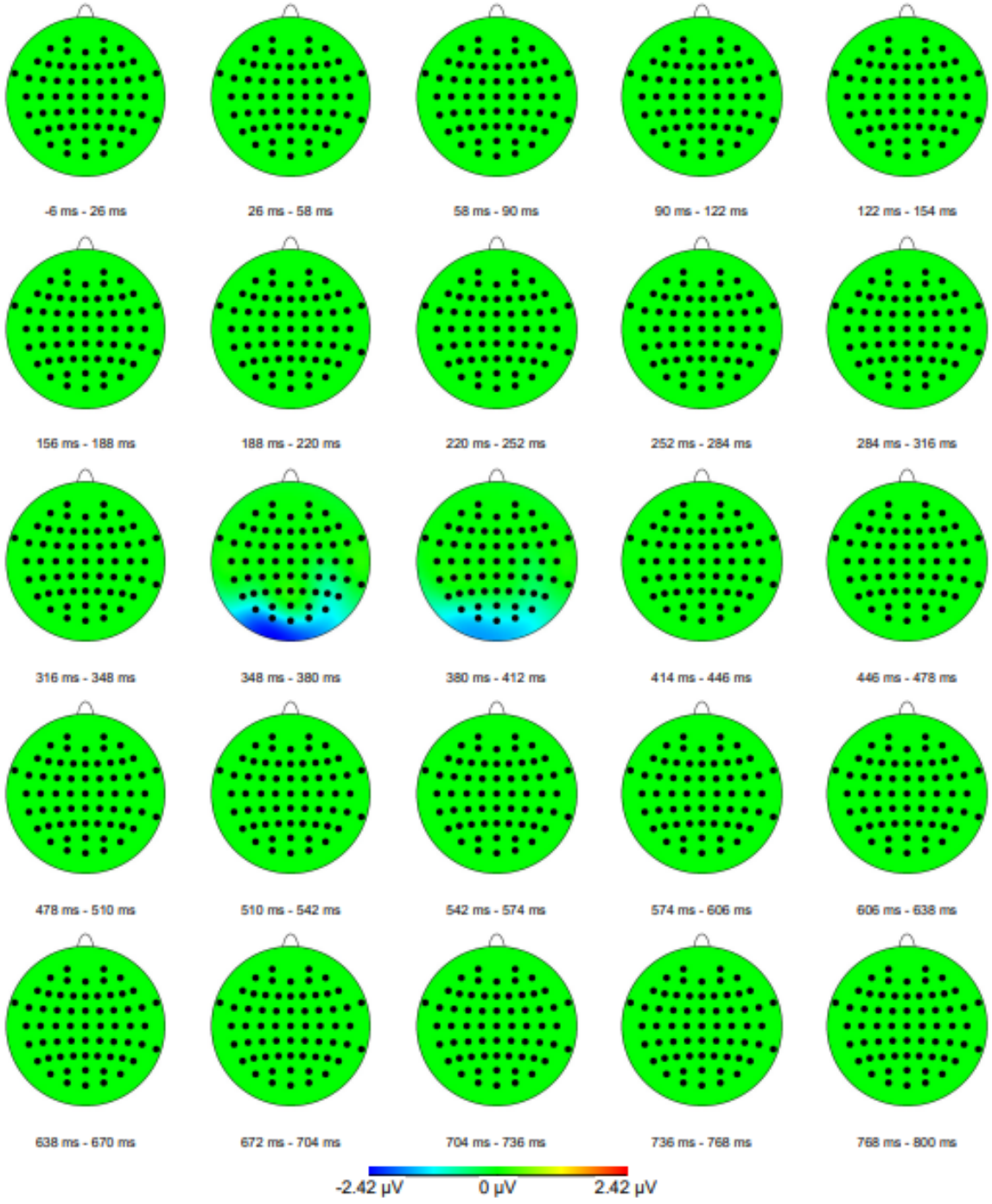

Figure 10: Human loss – computer loss topographical map.

Computer loss – Robot loss:
Within this comparison, of computer minus robot activation on loss trials, there were two clusters identified. As shown in figure 11, the first was a positive difference in activity between 250-385ms posterior sites. The second was a negative difference in activity between 292-386ms across occipital sites. However, neither of these areas of activation have not have the temporal or spatial significance to be an FRN or P300.

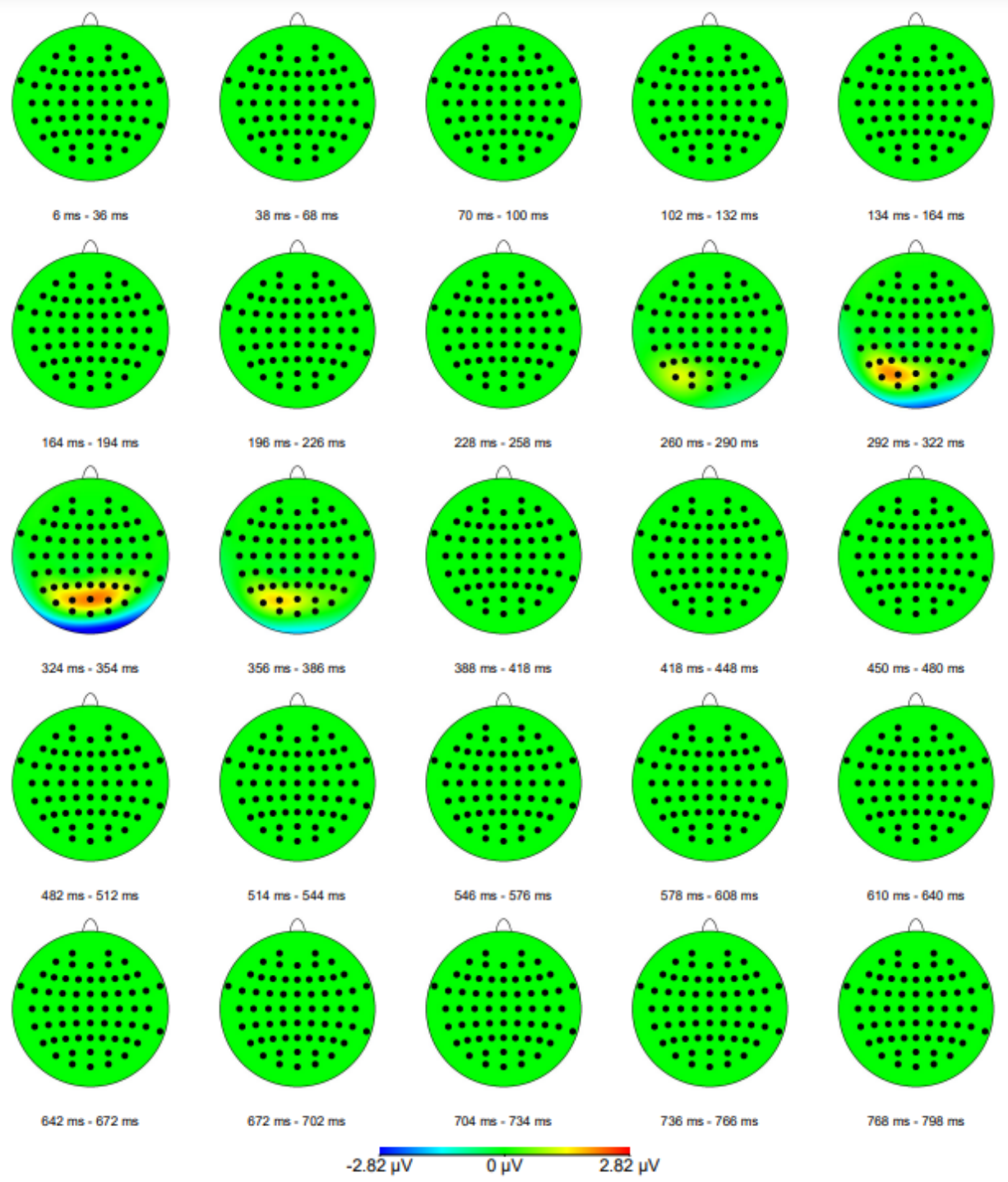

Figure 11: Computer loss – robot loss topographical map.

Discussion

This experiment combined behavioural, self-reported and neurophysiological methodologies culminating in a trust game that was played with three partners varying anthropomorphism while EEG was being recorded, followed by a generalised trust questionnaire for each partner. This allows one to gain a broader understanding of the impact of anthropomorphism on trust and expectation.

Both the behavioural and self-reported data suggested that humans are perceived as more trustworthy than computers and robots, however there is no difference between the trustworthiness perception of robots and computers. This contrasts some of the previous research, for example, using a checkmate game, which is an adaptation of the trust game (Berg et al., 1995), Alarcon et al (2023) identified that there were no differences in trust behaviours between a human and a robot. In addition to this, although self-reported findings have been varied, some researcher suggested that there are no differences in self-reported trust between humans and machines (Jain et al., 2022; Alarcon et al., 2021). The present findings do not support hypothesis 1 suggesting there will be no behavioural differences between the partners. In addition, these findings only partly support hypothesis 2 suggesting more anthropomorphic agents will be perceived as more trustworthy. One factor which may be leading to these findings is the interaction between trustworthiness and dominance perceptions.

Trustworthiness and dominance are the key dimensions of which people base their first impressions (Oosterhof & Todorov, 2008; Todorov et al., 2008; Sutherland et al., 2013). Gurtman (1992) suggested that distrust is closely related to dominance perceptions. In the present study, the human was rated as significantly more trustworthy than the robot and the computer, however there was no significant difference between the robot and computer. In addition, the human was perceived as less dominant than the computer, however there was no difference between the human and robot and robot compared to computer. The significantly increased dominance rating seen for computer, may have impacted the trustworthiness rating. Interestingly, more anthropomorphic robots have been found to be perceived as being more dominant (Kim et al., 2022) and within social interactions more dominant robots are perceived as less trustworthy (Li et al., 2015; Yoo et al., 2022). However, in this study there was no significant difference between human and robot dominance perceptions, so the difference in trustworthy perceptions may be due to other factors.

One theory which may explain this difference is the perfect automation schema (PAS) (Dzindolet et al., 2002). This theory states that users have a schema that suggests if automation is functioning correctly then it has an extremely low, even non-existent, error rate. It is suggested that PAS could be associated with higher trust prior to an error, but lower trust following an error (Dzindolet et al., 2002). Previous research suggests increased anthropomorphism leads to the application of social norms to an agent, and also increased expectations of the agent (Epley et al, 2007; Airenti, 2015). Participants may have perceived the correct behaviour to be to reciprocate, as this is a social norm (Gouldner, 1960), leading them to expect the partner to return more than they had originally invested. The computer as the least anthropomorphic agent does not have these social norms placed on it. Whereas, the robot is a more anthropomorphic agent and therefore there may be greater

expectations of reciprocity placed on this agent. When the robot fails to reciprocate, the PAS is violated and trust in the agent decreases. There is no perfect human schema therefore the humans error is attributed to factors such as momentary inattention or fatigue, meaning the participants may be more forgiving to human errors (Dzindolet et al., 2002). Expectation violation can be reflected by ERP components such as FRN and P300.

The ERP results suggest that there is a robot driven difference in activation between the three conditions. In the gain feedback conditions there was a larger FRN present for the human-robot comparison and the computer-robot comparison compared to no difference between FRN in the human-computer comparison. A larger FRN is associated with feedback that is worse than expected (Holroyd & Coles 2002; Nieuwenhuis et al., 2004; Schultz et al., 1997). This suggests that there were differences in violation of expectation between these agents, such that the robot elicited the largest expectation violation followed by the computer. These findings do not support hypothesis 3, which suggested the FRN would be larger for more anthropomorphic agents on loss trials as the human did not elicit the largest FRN. However, it does suggests anthropomorphism played a role in modulating trust and expectations of the participants. This could be explained by the PAS (Dzindolet et al., 2002). The robot has the expectation of performing perfectly by conforming the social norm of reciprocity (Gouldner, 1960), hence when it fails to meet this norm there is a large expectation violation suggesting the individual had over-trusted the robot. However, the computer, which is not anthropomorphic, may still be expected to perform perfectly and return more, however this expectation may not be as high as it is with the anthropomorphic robot due to the lack of imposing social norms onto the computer.

The lack of a difference in the human-computer comparison could be due to the fact that humans understand other humans can make mistakes and they are willing to forgive them. Whereas automation, such as the computer or the robot anthropomorphic robots, are expected to perform perfectly and the robot is expected conform to social norms and when they do not perform perfectly, consistently returning more of the investment this violates the expectation and trust quickly decreases (Dzindolet et al., 2002). Such violations require the updating of mental models regarding partners in order to correctly calibrate trust for the next round. The results of the present study do not support hypotheses 4, suggesting P300 was modulated by anthropomorphism, such that computers would have a larger P300 on gain trials compared to humans or robots. The only P300 component identified was in the human gain – robot gain comparison, and it seemed to follow the FRN. As suggested by previous research, P300 is associated with expectation violation, attentional resources and mental model updating when an event violates an expectation in response to positive feedback (Holroyd & Coles, 2002; Barto, 1995; Montague et al., 1996; Schultz, 2002; Yeung & Sanfey, 2004; Hajcak et al., 2005; Wu & Zhou, 2009; Donchin, 1981; Heslenfeld, 2003). This implies that the P300 reflected an unexpected outcome and therefore the updating of a mental model regarding the violation of expectation, as reflected by FRN, for the robot.

Conclusions

Together these findings suggest there is an impact of anthropomorphism on trust and expectation, however, it was not what was hypothesised. The behavioural and self-reported data suggested that humans are perceived as more trustworthy than both a robot and a computer. The EEG results revealed there were differences in expectation for the robot compared to the computer and the human, such that on the gain trials the participants were expecting more from a robot than it actually returned, as shown but the modulation of FRN. These findings supports the PAS theory which suggests that automation is expected to be perfect, and within this game the perfect response would be to reciprocate. This expectation can be seen to be violated by the modulation of the FRN, and therefore trust in the automation decreased dramatically which is seen in the behavioural and self-reported data. However, further research into this relationship would be needed in order to understand the development and decline of these expectations and trust.

Future work

This study investigated trust though behavioural, self-reported and neurological measures. However, in analysis it only looked at the final measures of trust, rather than the development of trust throughout the game and the impact of trust violations on the following trusting behaviours and perceptions. Previous research has highlighted the importance of past experience and outcome evaluation in the development of trust (Sanders et al., 2017; Mayer et al., 1995; King-Casas et al., 2005). To gain a more holistic view of how anthropomorphism impacts trust formation it is important to understand the impact of trust violations, as this would allow one to draw stronger conclusions regarding why the robot was perceived as less trustworthy compared to the human and whether this is related to the PAS.

Secondly, another limitation of this study is that only one pitch was used for the robot, this pitch may have engendered the perceptions of dominance or trustworthiness, therefore masking any potential impacts of anthropomorphism. Dominance and trustworthiness perceptions are influenced by a range of factors, one of which being voice pitch, such that lower pitch is associated with higher dominance and higher pitched voice are perceived as more trustworthy (McAleer et al., 2014; Hodges-Simeon et al., 2010, Elkins & Derrick, 2013). This has also been noted in research involving autonomous vehicles, such that more submissive voice was perceived as being more trustworthy (Yoo et al., 2022). For future studies, one could include multiple voice pitches for the pepper robot to investigate dominance and trustworthiness perceptions alongside the impact of anthropomorphism.

Acknowledgements

Firstly, I would like to thank my supervisor and personal tutor Dr Jeremy Goslin. I am extremely grateful for his consistent support, invaluable insight, and knowledge, throughout the preparational, experimental, and analytical stages of this research. Secondly, I would like to thank my friends for tolerating and supporting me throughout this research. Finally, I would like to thank my family, specifically my mum and my sister, for constantly providing me with unmatched advice and support throughout my time at university.

References:

actiCHamp Plus (64 channels) [Apparatus]. (2019). Gilching, Germany: Brain Products GmbH.

Airenti, G. (2015). The cognitive bases of anthropomorphism: from relatedness to empathy. *International Journal of Social Robotics*, 7(1), 117-127.

Alarcon, G. M., Capiola, A., Hamdan, I. A., Lee, M. A., & Jessup, S. A. (2023). Differential biases in human-human versus human-robot interactions. *Applied Ergonomics*, 106, 103858.

Alarcon, G. M., Gibson, A. M., Jessup, S. A., & Capiola, A. (2021). Exploring the differential effects of trust violations in human-human and human-robot interactions. *Applied Ergonomics*, 93, 103350.

Barto, A. G. (1995). 1 '1 Adaptive Critics and the Basal Ganglia,.'. Models of information processing in the basal ganglia, 215.

Bell, R., Sasse, J., Möller, M., Czernochowski, D., Mayr, S., & Buchner, A. (2016). Event-related potentials in response to cheating and cooperation in a social dilemma game. *Psychophysiology*, 53(2), 216-228.

Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., & Tanaka, F. (2018). *Social robots for education: A review. Science robotics*, 3(21), eaat5954.

Bemelmans, R., Gelderblom, G. J., Jonker, P., & De Witte, L. (2012). Socially assistive robots in elderly care: a systematic review into effects and effectiveness. *Journal of the American Medical Directors Association*, 13(2), 114-120.

Berg, J., Dickhaut, J., & McCabe, K. (1995). *Trust, reciprocity, and social history. Games and economic behavior*, 10(1), 122-142.

Blais, C., Ellis, D. M., Wingert, K. M., Cohen, A. B., & Brewer, G. A. (2019). Alpha suppression over parietal electrode sites predicts decisions to trust. *Social neuroscience*, 14(2), 226-235.

BrainVision Analyzer (Version 2.2.2) [Software]. (2021). Gilching, Germany: Brain Products GmbH.

British Psychological Society. (2022). Standards and guidelines. https://www.bps.org.uk/our-members/standards-and-guidelines

Brynjolfsson, E., & McAfee, A. (2014). The second machine age: Work, progress, and prosperity in a time of brilliant technologies. WW Norton & Company.

Buchan, N. R., Croson, R. T., & Solnick, S. (2008). Trust and gender: An examination of behavior and beliefs in the Investment Game. *Journal of Economic Behavior & Organization*, 68(3-4), 466-476.

Calvo-Barajas, N., Perugia, G., & Castellano, G. (2020, August). The effects of robot's facial expressions on children's first impressions of trustworthiness. In 2020 *29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 165-171). IEEE.

Chang, C. W., Lee, J. H., Chao, P. Y., Wang, C. Y., & Chen, G. D. (2010). Exploring the possibility of using humanoid robots as instructional tools for teaching a second language in primary school. *Journal of Educational Technology & Society*, 13(2), 13-24.

Chang, W. H., & Kim, Y. H. (2013). Robot-assisted therapy in stroke rehabilitation. *Journal of stroke*, 15(3), 174.

Chen, Q. Q., & Park, H. J. (2021). How anthropomorphism affects trust in intelligent personal assistants. *Industrial Management & Data Systems*.

De Visser, E. J., Krueger, F., McKnight, P., Scheid, S., Smith, M., Chalk, S., & Parasuraman, R. (2012, September). The world is not enough: Trust in cognitive agents. *In Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 56, No. 1, pp. 263-267). Sage CA: Los Angeles, CA: Sage Publications.

De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3), 331.

Dimoka, A. (2010). What does the brain tell us about trust and distrust? Evidence from a functional neuroimaging study. *Mis Quarterly*, 373-396.

Donchin, E. (1981). Surprise!… surprise?. Psychophysiology, 18(5), 493-513

Drnec, K., Marathe, A. R., Lukos, J. R., & Metcalfe, J. S. (2016). From trust in automation to decision neuroscience: applying cognitive neuroscience methods to understand and improve interaction decisions involved in human automation interaction. *Frontiers in human neuroscience*, 10, 290.

Duffy, B. R. (2003). Anthropomorphism and the social robot. Robotics and autonomous systems, 42(3-4), 177-190.

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human factors*, 44(1), 79-94.

Edwards, M. (1984). Robots in industry: An overview. *Applied ergonomics*, 15(1), 45-53.

Elkins, A. C., & Derrick, D. C. (2013). The sound of trust: voice as a measurement of trust during interactions with embodied conversational agents. *Group decision and negotiation*, 22(5), 897-913.

Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: a three-factor theory of anthropomorphism. *Psychological review*, 114(4), 864.

Evans, A. M., & Revelle, W. (2008). Survey and behavioral measurements of interpersonal trust. *Journal of Research in Personality*, 42(6), 1585-1593.

Gombolay, M.C., Yang, X.J., Hayes, B., Seo, N., Liu, Z., Wadhwania, S., & Shah, J.A. (2016). Robotic Assistance in Coordination of Patient Care. In Robotics: Science and Systems.

Gouldner, A. W. (1960). The norm of reciprocity: A preliminary statement. *American sociological review*, 161-178.

Guizzo, E., & Goldstein, H. (2005). The rise of the body bots [robotic exoskeletons]. *IEEE spectrum*, 42(10), 50-56.

Gurtman, M. B. (1992). Trust, distrust, and interpersonal problems: a circumplex analysis. *Journal of personality and social psychology*, 62(6), 989.

Guzman, A. L. (2016). Making AI safe for humans: A conversation with Siri. In Socialbots and their friends (pp. 85-101). Routledge.

Hajcak, G., Holroyd, C. B., Moser, J. S., & Simons, R. F. (2005). Brain potentials associated with expected and unexpected good and bad outcomes. *Psychophysiology*, 42(2), 161-170.

Hajcak, G., Moser, J. S., Holroyd, C. B., & Simons, R. F. (2007). It's worse than you thought: The feedback negativity and violations of reward prediction in gambling tasks. *Psychophysiology*, 44(6), 905-912

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. Human Factors: *The Journal of the Human Factors and Ergonomics Society*, 53(5), 517–527. doi: 10.1177/0018720811417254.

Hancock, P. A., Kessler, T. T., Kaplan, A. D., Brill, J. C., & Szalma, J. L. (2021). Evolving trust in robots: specification through sequential and comparative meta-analyses. *Human factors*, 63(7), 1196-1229.

Heerink, M., Kröse, B., Evers, V., & Wielinga, B. (2010). Assessing acceptance of assistive social agent technology by older adults: the almere model. *International journal of social robotics*, 2(4), 361-375.

Heslenfeld, D. J. (2003). Visual mismatch negativity. *In Detection of change* (pp. 41-59). Springer, Boston, MA.

Hinds, P. J., Roberts, T. L., & Jones, H. (2004). Whose job is it anyway? A study of human-robot interaction in a collaborative task. *Human–Computer Interaction*, 19(1-2), 151-181

Hodges-Simeon, C. R., Gaulin, S. J., & Puts, D. A. (2010). Different vocal parameters predict perceptions of dominance and attractiveness. *Human Nature*, 21, 406-427.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human factors, 57(3), 407-434.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychological review*, 109(4), 679.

Hu, X., Xu, Z., Li, Y., & Mai, X. (2018). The impact of trust decision-making on outcome processing: Evidence from brain potentials and neural oscillations. *Neuropsychologia*, 119, 136-144.

Jain, R., Garg, N., & Khera, S. N. (2022). Comparing differences of trust, collaboration and communication between human-human vs human-bot teams: an experimental study. CERN IdeaSquare *Journal of Experimental Innovation*.

Jensen, T., Khan, M. M. H., Fahim, M. A. A., & Albayram, Y. (2021, June). Trust and anthropomorphism in tandem: the interrelated nature of automated agent appearance and reliability in trustworthiness perceptions. *In Designing interactive systems conference 2021* (pp. 1470-1480).

Jian, J. Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International journal of cognitive ergonomics*, 4(1), 53-71.

Johnson Jr, R. A. Y. (1993). On the neural generators of the P300 component of the event-related potential. *Psychophysiology*, 30(1), 90-97.

Kessler, T. T., Larios, C., Walker, T., Yerdon, V., & Hancock, P. A. (2017). A comparison of trust measures in human–robot interaction scenarios. In Advances in human factors in robots and unmanned systems (pp. 353-364). Springer, Cham.

Kessler, T., Stowers, K., Brill, J.C., & Hancock, P.A. (2017). Comparisons of human-human trust with other forms of human-technology trust. Proceedings of the Human Factors and Ergonomics Society Annual Meeting.

Kim, L. H., Domova, V., Yao, Y., Huang, C. M., Follmer, S., & Paredes, P. E. (2022). Robotic presence: The effects of anthropomorphism and robot state on task performance and emotion. *IEEE Robotics and Automation Letters*, 7(3), 7399-7406.

Kim, W., Kim, N., Lyons, J. B., & Nam, C. S. (2020). Factors affecting trust in high-vulnerability human-robot interaction contexts: A structural equation modelling approach. *Applied ergonomics*, 85, 103056.

King-Casas, B., Tomlin, D., Anen, C., Camerer, C. F., Quartz, S. R., & Montague, P. R. (2005). Getting to know you: reputation and trust in a two-person economic exchange. *Science*, 308(5718), 78-83.

Klüber, K., & Onnasch, L. (2022). Appearance is not everything-Preferred feature combinations for care robots. *Computers in Human Behavior*, 128, 107128.

Krain, A. L., Wilson, A. M., Arbuckle, R., Castellanos, F. X., & Milham, M. P. (2006). Distinct neural mechanisms of risk and ambiguity: a meta-analysis of decision-making. Neuroimage, 32(1), 477-484.

Kraus, M., Kraus, J., Baumann, M., & Minker, W. (2018, May). Effects of gender stereotypes on trust and likability in spoken human-robot interaction. *In Proceedings of the eleventh international conference on language resources and evaluation* (LREC 2018).

Kulms, P., & Kopp, S. (2019). More human-likeness, more trust? The effect of anthropomorphism on self-reported and behavioral trust in continued and interdependent human-agent cooperation. *In Proceedings of mensch und computer* 2019 (pp. 31-42).

Lee, J. D., & Seppelt, B. D. (2009). Human factors in automation design. *Springer handbook of automation*, 417-436.

Lee, J.D, & Moray, N. (1992). Trust, control strategies and allocation of function in human-machine systems. *Ergonomics*, 35(10), 1243-1270.

Lee, J.D., & See, K.A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, 46(1), 50-80.

Lenth, R. (2020). emmeans: Estimated Marginal Means, aka Least-Squares Means. [R package]. Retrieved from https://cran.r-project.org/package=emmeans.

Li, J., Ju, W., & Nass, C. (2015, March). Observer perception of dominance and mirroring behavior in human-robot relationships. *In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction* (pp. 133-140).

Long, Y., Jiang, X., & Zhou, X. (2012). To believe or not to believe: trust choice modulates brain responses in outcome evaluation. *Neuroscience*, 200, 50-58.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, 8(4), 277-301. doi:10.1080/14639220500337708

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, 164, 177e190. https://doi.org/10.1016/j.jneumeth.2007.03.024

Mayer, R.C., Davis, J.H., & Schoorman, F.D. (1995). An integrative model of organizational trust. *Academy of management review*, 20(3), 709-734
.
McAleer, P., Todorov, A., & Belin, P. (2014). How do you say ‘Hello’? Personality impressions from brief novel voices. *PloS one*, 9(3), e90779.

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the national academy of sciences*, 98(20), 11832-11835.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of neuroscience*, 16(5), 1936-1947.

Muir, B.M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics*, 37, pp. 1905–1922

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of social issues*, 56(1), 81-103.

Natarajan, M., & Gombolay, M. (2020). Effects of anthropomorphism and accountability on trust in human robot interaction. *In Proceedings of the 2020 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 33-42).

Nieuwenhuis, S., Holroyd, C. B., Mol, N., & Coles, M. G. (2004). Reinforcement-related brain potentials from medial frontal cortex: origins and functional significance. *Neuroscience & Biobehavioral Reviews*, 28(4), 441-448.

Nieuwenhuis, S., Slagter, H. A., Von Geusau, N. J. A., Heslenfeld, D. J., & Holroyd, C. B. (2005). Knowing good from bad: differential activation of human cortical areas by positive and negative outcomes. *European Journal of Neuroscience*, 21(11), 3161-3168.

Oh, S., Seong, Y., & Yi, S. (2017). Preliminary study on neurological measure of human trust in autonomous systems. In IIE Annual Conference. Proceedings (pp. 1066-1072). Institute of Industrial and Systems Engineers (IISE).

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, 105(32), 11087-11092.

Pak, R., Fink, N., Price, M., Bass, B., & Sturre, L. (2012). Decision support aids with anthropomorphic characteristics influence trust and performance in younger and older adults. *Ergonomics*, 55(9), 1059-1072.

Parasuraman, R. (2003). Neuroergonomics: Research and practice. *Theoretical issues in ergonomics science*, 4(1-2), 5-20.

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors: *The Journal of the Human Factors and Ergonomics Society*, 39(2), 230- 253.

Pennisi, P., Tonacci, A., Tartarisco, G., Billeci, L., Ruta, L., Gangemi, S., & Pioggia, G. (2016). Autism and social robotics: A systematic review. *Autism Research,* 9(2), 165-183.

Perrin, F., Pernier, J., Bertrand, O., & Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalography Clinical Neurophysiology*, 72, 184e187. https://doi.org/10.1016/0013-4694(89)90180-6.

R Core Team. (2021). R: A Language and environment for statistical computing. (Version 4.1) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from MRAN snapshot 2022-01-01).

Rai, T. S., and Diermeier, D. (2015). Corporations Are Cyborgs: Organizations Elicit Anger but Not Sympathy when They Can Think but Cannot Feel. Organizational *Behav. Hum. Decis. Process*. 126, 18–26. doi:10.1016/j.obhdp.2014.10.001

Salem, M., Lakatos, G., Amirabdollahian, F., & Dautenhahn, K. (2015). Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. In Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (pp. 141-148). ACM.

Sambrook, T. D., & Goslin, J. (2015). A neural reward prediction error revealed by a meta-analysis of ERPs using great grand averages. *Psychological bulletin*, 141(1), 213.

Sanders, T. L., MacArthur, K., Volante, W., Hancock, G., MacGillivray, T., Shugars, W., & Hancock, P. A. (2017, September). Trust and prior experience in human-robot interaction. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 61, No. 1, pp. 1809-1813). Sage CA: Los Angeles, CA: SAGE Publications.

Sato, A., Yasuda, A., Ohira, H., Miyawaki, K., Nishikawa, M., Kumano, H., & Kuboki, T. (2005). Effects of value and reward magnitude on feedback negativity and P300. *Neuroreport*, 16(4), 407-411.

Schaefer, K.E. (2013). The perception and measurement of human robot trust. Doctoral Dissertation, University of Central Florida, Orlando, Fl.

Schaefer, K.E., Billings, D.R., Szalma, J.L., Adams, J.K., Sanders, T.L., Chen, J.Y., & Hancock, P. A. (2014). A meta-analysis of factors influencing the development of trust in automation: Implications for human-robot interaction (No. ARL-TR-6984). ARMY RESEARCH LAB ABERDEEN PROVING GROUND MD HUMAN RESEARCH AND ENGINEERING DIRECTORATE
.

Schäfer, M. B., Stewart, K. W., & Pott, P. P. (2019). Industrial robots for teleoperated surgery–a systematic review of existing approaches. *Current Directions in Biomedical Engineering*, 5(1), 153-156.

Schultz, W. (2002). Getting formal with dopamine and reward. Neuron, 36(2), 241-263.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275(5306), 1593-1599.

Simpson, B., McGrimmon, T., & Irwin, K. (2007). Are blacks really less trusting than whites? Revisiting the race and trust question. *Social Forces*, 86(2), 525-552.

Sin, J., & Munteanu, C. (2020). An empirically grounded sociotechnical perspective on designing virtual agents for older adults. *Human–Computer Interaction*, 35(5-6), 481-510.

Singmann, H. (2018). afex: Analysis of Factorial Experiments. [R package]. Retrieved from https://cran.r-project.org/package=afex.

SoftBank. (2014). SoftBank Mobile and Aldebaran Unveil “Pepper” – the World’s First Personal Robot That Reads Emotions | *SoftBank.* Webpage received 13/12/2023 https://www.softbank.jp/en/corp/group/sbm/news/press/2014/20140605_01/

Sutherland, C. A., Oldmeadow, J. A., Santos, I. M., Towler, J., Burt, D. M., & Young, A. W. (2013). Social inferences from faces: Ambient images generate a three-dimensional model. *Cognition*, 127(1), 105-118.

Systems, S. (2022). Sona Mobile. IN: Sona Systems.

The jamovi project. (2022). jamovi. (Version 2.3) [Computer Software]. Retrieved from https://www.jamovi.org.

Todorov, A., Baron, S. G., & Oosterhof, N. N. (2008). Evaluating face trustworthiness: a model based approach. *Social cognitive and affective neuroscience*, 3(2), 119-127.

Tolmeijer, S., Zierau, N., Janson, A., Wahdatehagh, J. S., Leimeister, J. M. M., & Bernstein, A. (2021). Female by Default? – Exploring the Effect of Voice Assistant Gender and Pitch on Trait and Trust Attribution. In Extended abstracts of the 2021 *CHI conference on human factors in computing systems* (pp. 1-7). DOI:https://doi.org/10.1145/3411763.3451623

Tolmeijer, S., Zierau, N., Janson, A., Wahdatehagh, J. S., Leimeister, J. M. M., & Bernstein, A. (2021, May). Female by default?–exploring the effect of voice assistant gender and pitch on trait and trust attribution. *In Extended abstracts of the 2021 CHI conference on human factors in computing systems* (pp. 1-7). https://doi.org/10.1145/3411763.3451623

Tulk, S., & Wiese, E. (2018). Trust and approachability mediate social decision making in human-robot interaction. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (Vol. 62, No. 1, pp. 704-708). Sage CA: Los Angeles, CA: SAGE Publications.

Wang, L., Zheng, J., Huang, S., & Sun, H. (2015). P300 and decision making under risk and ambiguity. *Computational intelligence and neuroscience*, 2015.

Wang, Y., & Quadflieg, S. (2015). In our own image? Emotional and neural processing differences when observing human–human vs human–robot interactions. *Social Cognitive and Affective Neuroscience*, 10(11), 1515-1524.

Wang, Y., Zhang, Z., Jing, Y., Valadez, E. A., and Simons, R. F. (2016). How do we trust strangers? The neural correlates of decision making and outcome evaluation of generalized trust. *Soc. Cogn. Affect Neurosci*. 11, 1666–1676. doi: 10.1093/scan/nsw079

Waytz, A., Heafner, J., & Epley, N. (2014). The mind in the machine: Anthropomorphism increases trust in an autonomous vehicle. *Journal of experimental social psychology*, 52, 113-117.

World Robotics Report 2021. (2021, October 28) International Federation of Robotics. Retrieved January 13, 2021 from https://ifr.org/ifr-press-releases/news/robot-sales-rise-again

Wu, Y., & Zhou, X. (2009). The P300 and reward valence, magnitude, and expectancy in outcome evaluation. *Brain research*, 1286, 114-122.

Yang, Y., Liu, Y., Lv, X., Ai, J., & Li, Y. (2022). Anthropomorphism and customers' willingness to use artificial intelligence service agents. Journal of Hospitality Marketing & Management, 31(1), 1-23.

Yeung, N., & Sanfey, A. G. (2004). Independent coding of reward magnitude and valence in the human brain. *Journal of Neuroscience*, 24(28), 6258-6264.

Yoo, Y., Yang, M. Y., Lee, S., Baek, H., & Kim, J. (2022). The effect of the dominance of an in-vehicle agent's voice on driver situation awareness, emotion regulation, and trust: A simulated lab study of manual and automated driving. Transportation research part F: traffic psychology and behaviour, 86, 33-47.

Appendices are provided separately as supplementary files (see additional downloads for this article).