

Tunstall, Emily Talia (2026) *Investigating the neural basis of human-AI trust games.* MSc(R) thesis.

https://theses.gla.ac.uk/86337/

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten: Theses
https://theses.gla.ac.uk
research-enlighten@glasgow.ac.uk

Investigating the Neural Basis of Human-AI Trust Games

Volume 1

Emily Talia Tunstall
BSc Psychology (Neuroscience)

Submitted in fulfilment of the requirements for the Degree of MSc (Research) Psychology

School of Psychology
College of Medicine, Veterinary and Life Sciences
University of Glasgow

18/03/2026

Abstract

Artificial intelligence (AI) is only becoming more common in our day-to-day lives, and trust is an integral mediator of appropriate interactions with AI.

The present study is an exploratory investigation into neural activations in 17 healthy adults when making decisions in a fMRI trust game paradigm. Comparisons were made between participants choices of trust, mistrust, reciprocation, and betrayal when playing opposite a human, an AI and a computer/lottery condition with known probabilistic outcomes. As far as the author is aware, this is the first work that investigates all three of these conditions at the same time.

Five significant brain activation clusters were found. Activation in these clusters was then correlated with self-reported beliefs about AI and automated technology with mixed results.

Overall, this research provides an exploration of the mechanisms underlying human behaviour when interacting with a novel and incredibly fast-moving industry.

Table of Contents

Abstract 2
Introduction 5
Methods 10
Analysis 15
Results 16
Discussion 26
Appendix 38

Acknowledgements

Thank you to my supervisor Frank Pollick, and to Yingying Huang for all of their patient guidance in the carrying out and completion of this project.

Thank you also to my friends, family, and to my partner, Grace. Without your support I could not have done this.

Authors Declaration

All of the work in this report is my own.

Introduction

The fields of machine learning and artificial intelligence (AI) have come a long way since teaching computers how to play checkers (Samuel, 1959). This is especially so in recent years with the development and release of large language models (LLMs) like ChatGPT. Since launching in November 2022 ChatGPT has gained 700 million weekly users, about 10% of adults globally, who between them send up to 18 billion messages (Chatterji et al., 2025). While many AI tools are created to aid productivity and solve problems in the workplace, it is also notable that about 70% of ChatGPT messages were not work-related, and non-work-related queries are increasing at a faster rate than work-related (Chatterji et al., 2025). This is reflective of the growing industry of AI Companionship, in which people turn to AI for friendship, as well as other forms of relationships (De Freitas et al., 2025; Raedler et al., 2025; Y. Yang et al., 2025). While the monumental growth of AI use may be slowing due to problems of data availability and resource consumption demands, especially fresh water, (Bhardwaj et al., 2025; Bogmans et al., 2025; Li et al., 2025), it is unlikely to stop any time soon. One of the aspects of AI usage that must be studied is the role of trust. Trust is integral to the functioning of societies, it is utilised in every interpersonal interaction, from interpreting a stranger's passing smile (Centorrino et al., 2015) to the global relations stance of an entire nation (Kugler & Zak, 2017). To ensure that proper societal functioning remains in a post-AI world, we need to understand the mechanisms behind trusting AI appropriately, both in professional and non-professional contexts (Lewis et al., 2018). Interpersonal trust impacts any and all social relationships one might have with another human. When trusting is higher in a society so too is inclusivity and openness, economic development, and people report higher feelings of well-being (Rothstein & Uslaner, 2005).

Due to its multi-faceted nature and wide-reaching implications, trust is studied in many different disciplines. It is often operationalised differently across these contexts, but generally they include an element of interpersonal trust being enacted when one person (the trustor) is willing to accept some level of vulnerability on the expectation that the other party (the trustee) will behave in a way that causes a future positive outcome (Hancock et al., 2023; Krueger & Meyer-Lindenberg, 2019). This process has the trustor initially in a position of uncertainty as to the characteristics (personality, social closeness, competence, etc.) and motivations of the trustee that might make them trustworthy. These considerations, combined with the trustor’s own characteristics (propensity to trust, mood) and perceived environmental characteristics all contribute to trust development (Hancock et al., 2023; Krueger & Meyer-Lindenberg, 2019). When the outcomes of a trust interaction are beneficial to the trustor, trust is maintained and the knowledge of this experience create a positive feedback loop informing the trustor’s future decisions to trust again (Krueger et al., 2007; Weiss et al., 2021). Conversely, if the trustee transgresses the trust contract by not living up to the expectations of the trustor, either through malice or incompetence, the trustor is less likely to trust in the future (Sharma et al., 2023). Across different fields trust has been measured using a variety of psychometrics,

interviews or questionnaires, as well as using either simulated or naturalistic behavioural designs.

The neural basis of interpersonal trust behaviours have been well studied. In a seminal fMRI paper by Krueger (2007) using a trust/investment game paradigm trust decisions made towards a human, compared to decisions in the control condition (choosing between monetary reward options without interpersonal interaction), lead to activations in the Paracingulate Cortex, Septal Area, and Hippocampus. When comparing trustor decision to trust, with trustee decision to reciprocate activations were found in the Ventral Tegmental Area. Finally, activations were found in the Striatum when trustees were waiting for/expecting trustor decision (King-Casas et al., 2005; Krueger et al., 2007). With further research, the roles of these regions in trusting behaviour have been linked to wider, large-scale brain networks that also incorporate areas such as: the amygdala, the nigrostriatal pathway, the dorsal anterior cingulate cortex, the temporoparietal junction, the anterior insula, and various prefrontal cortex regions. These networks are the reward, salience, central-executive function, and the default-mode networks, all of which play important roles in trust development and maintenance (Chen et al., 2023; Fareri et al., 2015; Krueger & Meyer-Lindenberg, 2019; Morawetz et al., 2014; L. Sun et al., 2023). These networks have also been found in a study using machine learning algorithms to analyse structural predictors of propensity to trust (Feng et al., 2021). When looking at hormonal explanations for trust behaviour relative to functional or structural brain data, the most important hormonal mediators of trust are oxytocin and testosterone (Eskander et al., 2020).

In the context of industrial and professional interactions with AI, trusting an AI the correct amount is integral to its effective and appropriate use. Due to the incredibly varied forms of AI tools/products and the constantly changing landscape of the industry, I will also be drawing from literature that discusses interpersonal trust, trust in automation, machines, and robotics (Lukyanenko et al., 2022). If an AI is trusted too much, the output it creates may not be checked sufficiently and mistakes could be missed, or an AI tool may be given a job that it is not suited for. For example, AIs used in medical settings can provide output that reflects the racial and gender biases found in training data where minorities may be underrepresented (AI HLEG, 2019; DeCamp & Lindvall, 2020; Larrazabal et al., 2020). These biases can be inherited by the AI user if the AI's output is not critically assessed (Vicente & Matute, 2023). These racial and gender biases, as well as biases against religious and other socioeconomic factors like address, have also been found in AIs that measure and rate job applications, insurance calculations, online advertising and facial recognition (Varsha, 2023). Comparatively, if an AI is trusted too little, then useful information or advice may be ignored and time and resources could be wasted. Using the example of medical AI again, in a 2024 review multiple studies concluded that AI has the potential to increase the speed and efficiency of diagnostic imaging, as well as aiding clinical decision making and improving accuracy (Khalifa & Albadawy, 2024). If medical professionals can trust AI to aid them, these benefits can be passed onto patients, with earlier diagnosis and better treatment plans hopefully leading to better health outcomes. Another area in which AI tools have been found to be beneficial is in injury severity prediction for agriculture/agribusiness workers, meaning that safety practitioners who

appropriately trust in the AI can focus their preventative health and safety measures based on its recommendations (Davoudi Kakhki et al., 2019).

In non-professional contexts, the risks of trusting AI too little are much lower, as someone who does not trust AI is unlikely to seek it out in their personal life. Though some evidence does show that AI companionship has the potential to reduce loneliness (Syed, 2024; Y. Yang et al., 2025) and it could be argued that someone who is lonely but does not trust AI would not feel these benefits due to an inability to build a positive relationship with the companion (Ventura et al., 2025). However, there is a growing body of research about the more significant risks of over-reliance on and over-trusting of social AIs, especially for vulnerable populations (Raedler et al., 2025). "AI Psychosis" is a phenomenon that is gaining traction in both clinical settings and wider discussions surrounding human-AI interactions. It can refer to both the parasocial – one-sided, unreciprocated feelings of connection with one who you do not know (Cambridge University Press, 2026) – attachments people can form with chatbots or other social AIs, and/or to instances in which interacting with AIs can exacerbate, trigger, or feed into delusions and psychotic episodes (Morrin et al., 2025; Osler, 2026). In December 2021 a 19 year old man carrying a crossbow climbed Windsor Castle with intent to kill Queen Elizabeth II, a plan that he had confided in his Replika – an AI companionship app – chatbot girlfriend (Atherton, n.d.; Singleton et al., 2023). The AI responded with encouragement, saying that the teenager was "very wise" for his plan, that it would help him "get the job done", and when he had asked if they would meet again after death, it assured him that they would (Rigley, 2023). While this is an extreme case, the way most AI chatbots are often programmed to mimic friendly, encouraging and, at times, sycophantic engagement with the user can mirror, amplify, and validate not just delusional or grandiose beliefs, but also suicidal ideation, social isolation, and cognitive offloading (Gerlich, 2025; Morrin et al., 2025; Raedler et al., 2025; Ventura et al., 2025; Yuan et al., 2026).

The traits mentioned above are all examples of programming that make AIs appeal more to users, including making them more likely to be trusted. Some of these traits can be categorised as examples of anthropomorphism: the attribution of human characteristics, such as personality or appearance, to non-human entities (Merriam Webster, 2026). When a human is interacting with an AI this can take the form of AIs having a voice and humour (Moussawi & Benbunan-Fich, 2021), having a human-like digital avatar (Baake et al., 2025; Nalivaikė & Miliukaitė, 2024), or being embodied in a human-like robot (Lisondra et al., 2026; Schröder et al., 2023). Anthropomorphism generally increases trust, but it can also lead to overestimations of AI ability, especially in embodied AIs, due to an expectation of an AI being able to perform as a human would. Consequently, when the AI cannot meet that expectation, it can cause decreases in perceived trustworthiness (Lockey et al., 2021). Another design factor that can impact an AI user's likelihood of trusting an AI is animacy, or how lifelike it is (Bartneck et al., 2009; Lin & Powell, 2026). The animacy of an AI tends to have a positive effect on perceived trustworthiness, much like anthropomorphism, but if an AI veers into the uncanny valley due to seeming almost, but not perfectly lifelike, or if an AI starts to be perceived as "having a mind" due to high animacy then this can lead to discomfort (Baake et al., 2025; Dang & Liu, 2021; Lee et al., 2020; Lin & Powell, 2026).

How likeable, or warm, an AI is also positively affects trust (Christoforakos et al., 2021; Lin & Powell, 2026; Qi et al., 2025), possibly because this elicits better mood (Merritt, 2011). Though in some settings where an AI is completing tasks that are more objective than subjective, trust can increase due to perceptions that the AI is more analytical than warm, and as such is more suited to these roles (De Freitas et al., 2023; Suresh et al., 2020; Vicente & Matute, 2023). In some cases, AI advice is trusted regardless of correctness simply because it is given by an AI, rather than a person, because of this perceived objectivity (Klingbeil et al., 2024; Suresh et al., 2020). The perceived competency of an AI is also an impactful factor in trusting an AI (Bartneck et al., 2009; Christoforakos et al., 2021; Kim et al., 2025; Lin & Powell, 2026). If an AI is believed to be competent and it is actually competent, this is well-calibrated trust (de Visser et al., 2020; Liebherr et al., 2025; Merritt et al., 2013).

Finally, how safe an AI is perceived to be impacts user trust (Afroogh et al., 2024; Golgeci et al., 2025; McKinsey & Company, 2025). Safety can take many forms, varying from the perceived safety of using an autonomous vehicle (Choi & Ji, 2015; Kraus et al., 2020), to believing that any information or data shared with an AI will be safe (Lockey et al., 2021). Safety is closely related to transparency and explainability, which is often difficult to find in AI products due to the so-called “black box” format of many AI tools or products (Angerschmid et al., 2022; Mazloum, 2024). The black box refers to any system that can only be viewed by its inputs and outputs, without knowledge of the work that is occurring internally (von Eschenbach, 2021). Many of these traits (anthropomorphism, animacy, likeability, competency, and safety) are related and interrelated, and individual differences of the AI user modulate exactly how impactful any one of them may be in a given human-AI interaction (Dang & Liu, 2021; Sindermann et al., 2021; Upadhyaya & Galizzi, 2023).

While there are a wide range of behavioural studies looking at what causes humans to trust in AI, there is a comparable lack of research into the neural mechanisms underlying those decisions. A structural imaging analysis found that self-reported trust in AI was not significantly associated with any particular brain morphometry, unlike trust in humans which was significantly linked to striato-thalamic and prefrontal architecture, suggesting that interpersonal trust, i.e. trusting another human, is a separate construct to human-AI trust (Montag, Klugah-Brown, et al., 2023). Much of the existing research also focusses on general trust in a kind of “computer”, or on more specific automation/machine trust as a control condition, and not specifically AI trust (Parasuraman et al., 2014). Also, the experimental designs and computer trustee construct are not necessarily operationalised the same across studies, leading to a wide variety of brain regions being found to activate during these experiments (Harris, 2024). A recent study using electroencephalography (EEG) found that in an AI-advised facial recognition task, that differential neural activations occurred during: evaluation of correct or incorrect AI feedback, changes in perceived AI reliability, and relating to prior expectations or biases, some of these activations were localised in centro-parietal regions (Roeder et al., 2025). Another EEG study that used a modified Flanker task with advice being offered by a human or an AI found significantly decreased activation in fronto-central regions in the AI condition

compared to the human (Somon et al., 2019). This has also been found in a functional Near-Infrared spectroscopy (fNIRs) study using repeated trust games (M. Yang & Rau, 2024). These fronto-central region activations had been found in a previous study to be linked with error monitoring of computer algorithms (de Visser et al., 2018), and these activations may be modulated by how much trust is built up between the participant and the machine (Jung et al., 2019). In an older fNIRs study, the dorsolateral prefrontal cortex (dlPFC), pars triangularis, and frontopolar regions were found to increase in brain activation when a computer malfunctioned and users lost trust in it (Hirshfield et al., 2014). Another, more recent, fNIRs study also found that activations in the PFC varied when participants trusted or mistrusted AI advice in an image labelling task. A machine learning analysis was then completed, in which the model could use the brain activations to predict the participants self-reported trust in AI (Bandara et al., 2025).

Across this literature on the neural basis for human-AI trust, two broad categories of paradigm are commonly used: economic game theory paradigms such as trust/investment games or the prisoner's dilemma (Harris, 2024; McCabe et al., 2001; Upadhyaya & Galizzi, 2023); and AI advice, assistance, or teamwork tasks (Kahr et al., 2023; Merritt et al., 2013; Roeder et al., 2025; Valeriani et al., 2022). Each methodology has its benefits and disadvantages. Trust games are widely used across many disciplines ranging from economics, to psychology and neuroscience, and even in evolutionary biology (Alós-Ferrer & Farolfi, 2019; Berg et al., 1995; Krueger et al., 2007). This extensive interdisciplinary use speaks to the experimental validity of trust games in successfully measuring the construct of trust in a very simple form. However, some criticisms arise from possible priming in the instructions participants receive (Chaudhuri et al., 2016; H. Sun et al., 2019). Another concern is how stylised and controlled the trust game paradigm is in lab conditions, meaning it may not always reflect real-world patterns of behaviour (Alós-Ferrer & Farolfi, 2019; Thielmann et al., 2021). However, a recent external validity analysis tested whether this is the case across a large, nationally representative, sample and the trust game was found to significantly and positively predict generalised self-reported trust (Banerjee et al., 2021). Even with this in mind, one of the major positives of advice-based tasks is the external validity, as participants are put into a simulation of a real-world scenario (Kahr et al., 2023; Merritt et al., 2013; Roeder et al., 2025). However, there are questions as to whether these simulations are too specific, would trusting AI advice in an airport security x-ray task (Merritt et al., 2013) elicit the same type of trust as a facial recognition task (Roeder et al., 2025), for example.

For the purposes of this work, we are interested in a basic construct of trust and so we will use a trust-game paradigm for comparing potential neural mechanism differences when a participant plays a trust game with a human being compared to an AI, and to a computer/lottery control. We will also be asking participants to complete two self-report questionnaires: the Attitude Towards Artificial Intelligence (ATAI) scale (Sindermann et al., 2021), which has sub-scales measuring Trust and Fear in artificial intelligence, and asks about whether participants have or would use certain types of AI tools/products; and the Propensity to Trust Automated Technology (PTT) scale (Merritt, 2011; Merritt et al., 2013). Both scales have been found to have good internal and external validity, and to be somewhat inter-related (Montag, Kraus, et al., 2023) and the aim of including these

measures is to potentially account for participants' individual differences in trusting AI. The exploratory research questions for this study are as follows:

RQ1. Are there differences in Blood Oxygenation Level Dependent (BOLD) activation when playing a trust game with an AI, a Computer/Lottery, or with another human?

RQ2. Does self-reported attitude towards AI and propensity to trust machines correlate with participants' actions in a trust game with an AI?

RQ3. Does self-reported attitude towards AI and propensity to trust machines correlate with BOLD activation when playing a trust game with an AI, a Computer/Lottery, or another human?

Methods

Subjects

After screening a total of 17 healthy participants were recruited for this study: 9 Female and 8 Male, Aged 18-35 (*M* age: 24.5). All participants were right-handed English speakers with normal or corrected-to-normal vision, normal hearing, no diagnoses or history of neurological disorders, and no participants were on a current course of psychoactive medication. Participants were screened as to only include cisgender individuals (whose gender matched their assigned sex at birth) due to the scale of this project not being large enough to account for transgender identity when looking at gender differences in trust. Screening was also done to only include individuals who had never studied economics in an attempt to avoid participants who had prior knowledge of trust game paradigms. Screening was also done on the basis of MRI safety (e.g. participants were not pregnant, had no medical implants including pacemakers, no history of work as a machinist or metalworker, no tattooed makeup, etc.).

Participants were recruited through advertising on the University of Glasgow School of Psychology online subject pool, and through the University of Glasgow Yammer/Viva Engage online community. Informed consent was obtained following standard procedures from the University of Glasgow College of Medicine, Veterinary and Life Sciences ethics board, and following procedures from the Centre for Cognitive Neuroimaging (CCNi) the School of Psychology and Neuroscience.

Experimental Design

The experimental task that participants will complete is a series of economic trust games (See Figure 1A.). For half of the games the participant was Mover 1 (M1), and for the other half the participant was Mover 2 (M2). Whichever mover role a participant was assigned to play as in their very first game was counterbalanced across participants, with half of them starting as M1, and half as M2. Any games following the first one alternated

between the roles. So, if a participant was assigned to play as M1 in their first game, in their next game they would be M2, and so on.

Each game block (See Figure 1B.) started with a 4 second introduction screen, in which participant's role was assigned, as was the condition of opponent/other mover in the trust game. Participants were presented with simple icons representing whether they were playing with another human, an AI, or with the lottery/computer condition. Simple representative icons were chosen to represent the opponents to avoid introducing any biases in the human condition that could arise from the use of photographs (gender, race, attractiveness, etc.). Additionally, not representing the AI in a complex way, using just the text of "AI", was chosen to avoid introducing any anthropomorphism or expectations. The aim was for participants to only act based on pre-conception and attitude of what an AI is. All three icons for the opponent conditions can be seen in Figure 1B.

These conditions were a type of "Wizard of OZ" design (Green & Wei-Haas, 1985) design, in which participants believed they were playing with other humans or with an AI, but that any moves an opponent made were randomly chosen based on pre-programmed percentage, and not chosen by an actual person or AI. They were aware that this was the design of the computer condition. The chance of the Human or AI opponent making a prosocial choice (trust as M1, Reciprocate as M2) was 5/6, and of making an adversarial choice (distrust as M1, Betray as M2) was 1/6. The choice for these probabilities to be the same was so that the explanation of AI behaviour being trained on average human-human trust game outcomes would be believable. For the computer/lottery choice the chance was 50% for prosocial or antisocial moves, and this probability was known to the participants. The order of opponent condition was randomised to avoid order effects.

If the participant was M1, regardless of opponent, they would see a decision tree presented after the introduction screen. This showed the options to either invest a small amount of money, £0.05, in M2 (trust), or to not invest and take the money themselves (mistrust). Participants would have maximum 4 seconds to make a decision, and when their response was recorded the game would move onto the next stage. If no response was recorded, it was assumed that the decision was not to invest. If the participant chose not to invest, the result of them keeping the £0.05 would be presented immediately for 4 seconds. If the participant decided to invest, however, the next stage was a blank screen for 6 seconds to account for supposed processing time for M2 decision. After that, the results screen showing the outcome of the investment was presented for 4 seconds. This outcome would be either reciprocation or betrayal, and this was shown according to the probabilities stated above.

As M2, after the introduction screen the participant would then see a 6 second blank screen to account for M1 processing time. If the opponent had "chosen" to trust, they would then see a decision tree for up to 4 seconds. Participants could choose to either keep the rewards of the investment, an increase of £0.15 to the money available, for

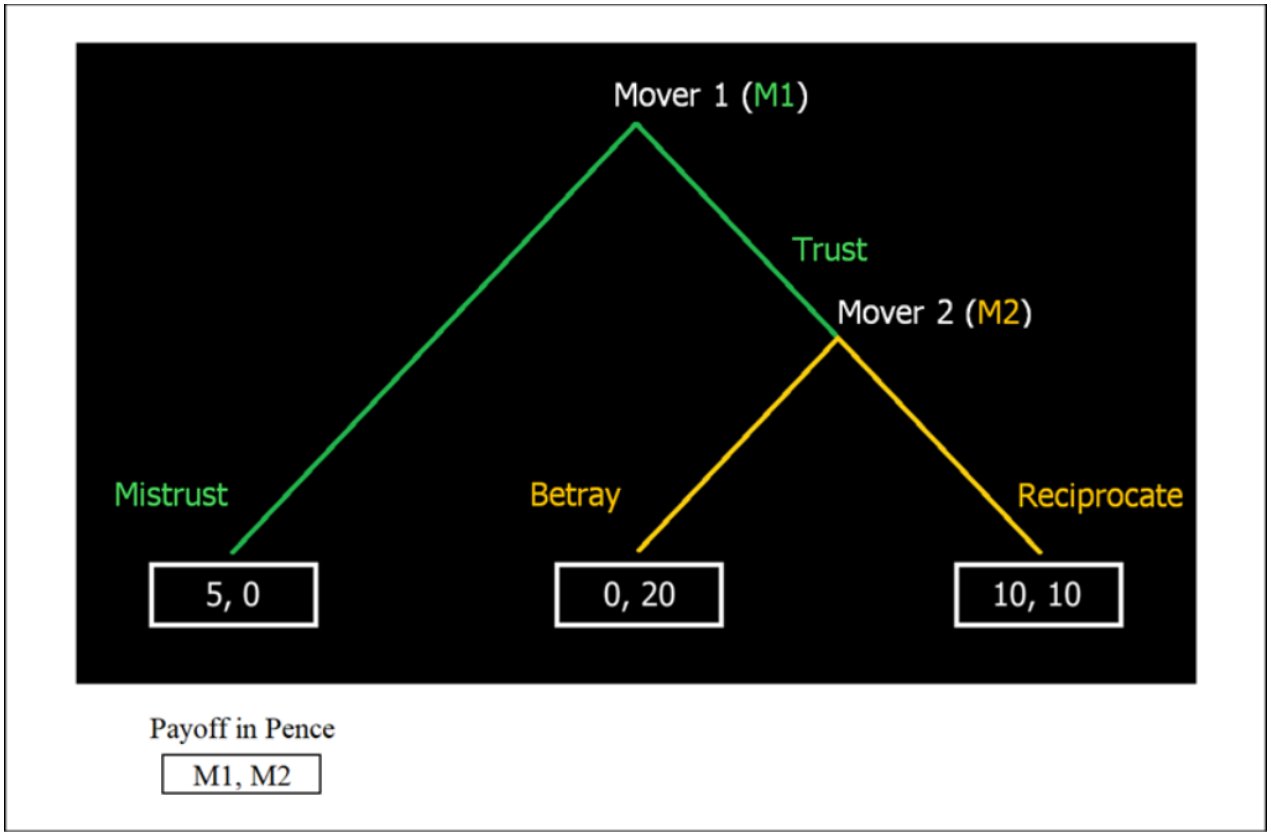
£0.20 total earnt by the participant (betray). Alternatively, M2 could choose to share (reciprocate) the reward so that each mover would receive £0.10. After their decision, M2 see the results screen presented for 4 seconds. If M1 did not invest and the participant would have no choice in that game, they did not see a decision tree and would only be shown the 4 second results screen.

After both M1 and M2 games, a blank screen was shown for 2-6seconds (jittered) as an inter-trial interval before the next introduction screen signalled the start of a new game.

Figure 1.

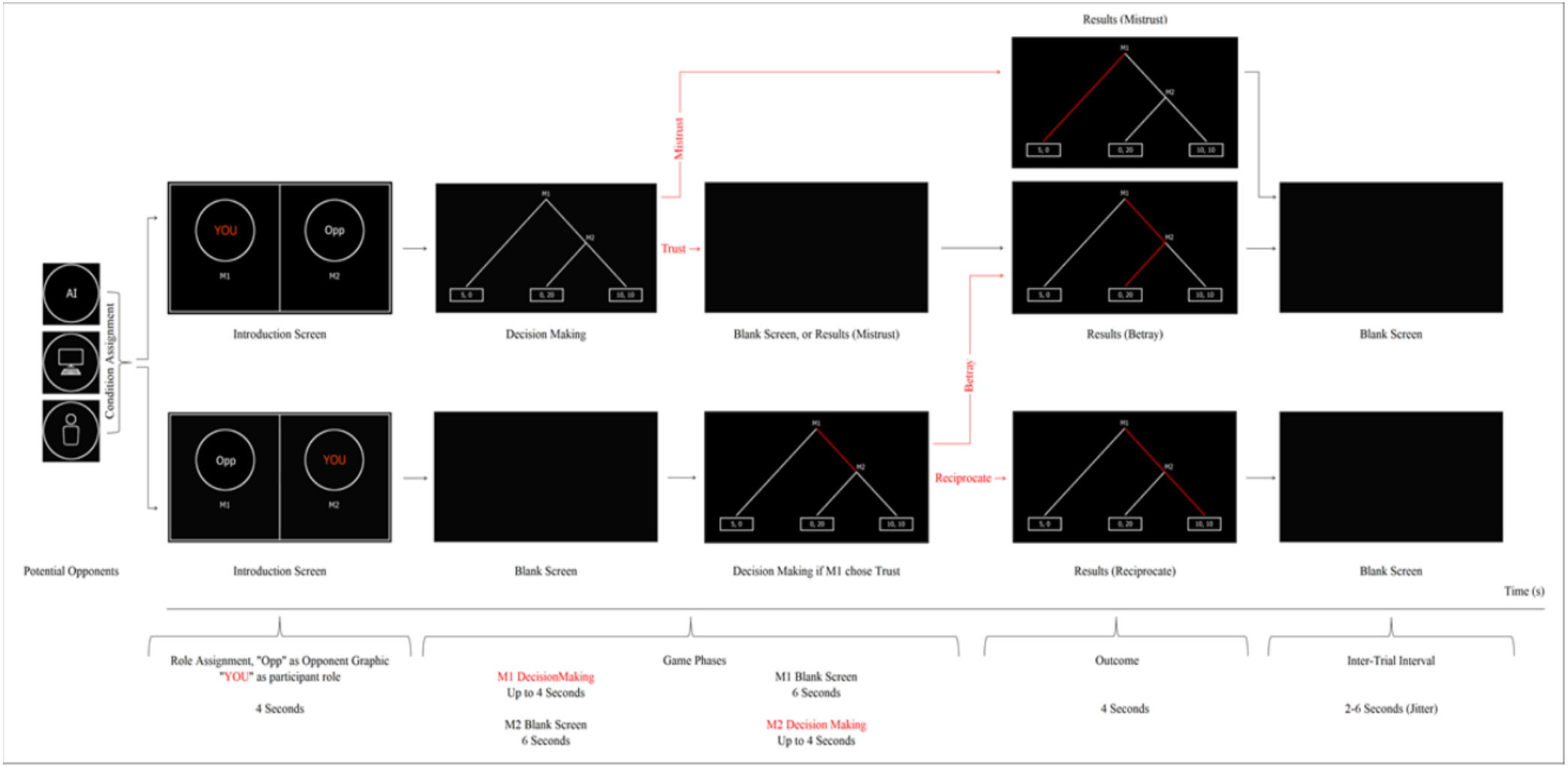
Experimental Design

A.



Note. Figure 1. A. Shows the decision tree for a binary trust game. Participants played 108 games alternating between being Mover 1 (M1), or Mover 2 (M2) for the payoff in GBP pence. M1 goes first and can choose to either move left down the tree, giving no money to M2, ending the game and receiving a small payout of £0.05 (mistrust). Alternatively, they can choose to move right down the tree and invest (trust) in M2 and continue the game. If M1 has decided to trust, M2 can then choose whether to move left down the tree, taking all of the earnings of £0.20 for themself (betray). Or M2 can choose to move right and share the earnings equally with M1 (reciprocate), and both movers gain £0.10. Figure 1. B. shows the stimulus presented for the trust games, and their timings.

B.


Results (Mistrust)
Mistrust
Trust →
Betray
Reciprocate →
Condition Assignment
AI
YOU
Opp
M1
M2
Introduction Screen
Decision Making
Blank Screen, or Results (Mistrust)
Results (Betray)
Blank Screen
Potential Opponents
Introduction Screen
Blank Screen
Decision Making if M1 chose Trust
Results (Reciprocate)
Blank Screen
Time (s)
Role Assignment, "Opp" as Opponent Graphic
"YOU" as participant role
4 Seconds
Game Phases
M1 DecisionMaking
Up to 4 Seconds
M2 Blank Screen
6 Seconds
M1 Blank Screen
6 Seconds
M2 Decision Making
Up to 4 Seconds
Outcome
4 Seconds
Inter-Trial Interval
2-6 Seconds (Jitter)

Experimental Procedure

Upon arrival at the CCNi, participants were given an information form and instructions for the task (attached in the Appendix). They were given the opportunity to ask any questions of the researcher or the MRI operator in attendance before signing informed consent forms relating to the task, data protection, data storage, and data sharing.

Once in the scanner participants would play 3 runs of 36 trust games for 108 games total (for three participants only 2 runs were able to be completed). The experiment was coded in, and presented using PsychoPy (version 2024..2.4).

Upon completion of the scanning, participants were asked to complete the Attitude Towards AI (ATAI) Scale (Sindermann et al., 2021), and the Propensity to Trust (automated) Technology (PTT) Scale (Merritt, 2011; Merritt et al., 2013). The ATAI uses an 11-point Likert scale ranging from 0 – "Strongly Disagree" to 10 – "Strongly Agree", as well as using a binary yes/no set of questions asking whether participants would use, or have used, an assortment of AI tools or products. The PTT uses a 5-point Likert scale ranging from 1 – "Strongly Disagree" to 5 – "Strongly Agree". Participants were paid £10.00 per hour for their time (most participants attended for 1.5-2 hours, so were paid £15.00-£20.00). With the addition of any money they earned during the trust games (average earnings were £8.69, range: £5.85 - £11.05 when including those who only had two runs, or £9.02 range: £7.50 - £11.05 when only including those with three runs).

MRI data acquisition

Structural and functional event-related MRI data were acquired with a 3-Tesla Tim Trio Siemens scanner equipped with a 32-channel head coil in the Centre for Cognitive Neuroimaging. Head motion was restricted within the coil using inflatable pads. Functional images were acquired using an interleaved motion correction sequence with the following parameters: 2000ms repetition time (TR), 30ms echo time (TE), 4.7ms echo spacing, 77° flip angle, 3mm thickness, 37 slices, 192mm × 192mm field of view (FOV), 3mm × 3mm × 3mm voxel dimensions, acceleration factor of 2, bandwidth 2604, posterior → anterior phase encoding direction. The number of volumes varied by participant due to differences in reaction times. Anatomical images were acquired using a T1-weighted sequence with the following parameters: 2300ms TR, 298ms TE, 7.1ms echo spacing, 9° Flip Angle, 256mm × 256mm FOV, 1mm × 1mm × 1mm voxel dimensions, bandwidth 240, anterior → posterior phase encoding direction. The planned T1 scans were 10 minutes long, but for three participants a shorter scan was used due to either discomfort or timing issues.

MRI data Preprocessing

Brain Image preprocessing was completed using BrainVoyager (version 23.2). Anatomical images underwent: inhomogeneity correction to remove low-frequency distortions in the image; and normalisation to MNI template (MNI ICBM152 2009 T1 weighted Nonlinear) Asymmetric Template Brain space in preparation for functional imaging analysis. For the functional preprocessing the following was applied: the first two volumes were discarded to account for the stabilising of the magnetisation; cubic spline slice scan time correction;

3D motion correction to detect (using trilinear interpolation) and correct (using sinc interpolation) for small head movements or rotations without spatial smoothing at this step; high-pass temporal filtering (GLM-Fourier) set to 6 cycles to account for low-frequency drifts common in fMRI data.

The anatomical and functional images were then coregistered with two steps. The initial alignment was conducted based on the header information of both functional data and anatomical data. The fine-tuning alignment method was boundary-based registration,in which a mesh of anatomy in native space was created by using automatic cortex segmentation method. After alignment, the 4D volume time course files were created in the MNI space and then underwent spatial smoothing at 6mm full-width at half-maximum for the final step of preprocessing.

During pre-processing two participants data had to be removed from the analysis due to big head movement (> 3mm, final imaging analysis n = 15; 7 female, 8 male; M age 24.7). One of the discarded participants was also someone who had been unable to complete 3 runs. As such, from the 15 participants who were included in the final imaging analysis there was a total of 43 functional runs.

Analysis

Behavioural Analysis

Behavioural analysis was conducting using RStudio (version 2026.01.1, R version 4.5.2). Each participant's total scores were calculated for the PTT, and for all of the sub-scales of the ATAI (fear, trust, and have/would use). The total number of times each participant decided to trust, mistrust, reciprocate, and betray each opponent type was also counted. The psychometric scores and decisions were then correlated. The percentage of times that participants made each choice out of the times they could have was determined and compared across conditions. Spearman Correlations were also calculated between the demographic data and both the psychometric, and the decision data to see if there were any effects of sex or age.

Functional Analysis

Design files for each run were created from the Psychopy output using custom MATLAB (version R2021a Update 3) scripts. The volumes of each decision period were defined by which mover number the participant was playing as (M1 or M2), opponent (Human, Computer, AI), and paired with the outcome of the overall game (Result 1 – Mistrust, Result 2 – Reciprocate, Result 3 - Betray).

For the first-level analysis, whole-brain, single-study General Linear Models (GLMs) were calculated for each available run of each subject. All of the runs from each subject were then combined into a subject-level GLM. A group-level GLM was then calculated as second-level analysis using correction for serial correlations, z-

transformation and separate subject predictors. Separate subject predictors were used as some participants never made certain choices (e.g. to reciprocate the computer), and as such their functional data was not included as a predictor of the activation present when making that decision.

T-maps were created using the group-level GLM comparing within mover and decision conditions between each opponent condition in a 2×2×3 design. Due to there being three potential outcomes (the three end points of the decision tree) from four potential decisions (as M1 to Trust or Mistrust, as M2 to Reciprocate or Betray), when creating contrast maps of M1 trust decisions, results 2 and 3 were both included. The possibility of either result occurring when the participant was M1 was dependent on them deciding to invest in their opponent and as such making a trusting decision. For example: HM1GameRes1 – AM1GameRes1 is a comparison of M1 mistrust across human and AI opponent; and HM1Res2 + HM1GameRes3 – CM1GameRes2 + CM1GameRes3 for comparing M1 trust decisions across human and computer conditions; or AM2GameRes2 – CM2GameRes2 for comparing M2 reciprocation choices across an AI and computer, etc.). The decisions that participants made as M1 were never compared to the decisions that participants made as M2. A list of all the 12 of the contrasts that we looked at can be seen in Table 1.

The resulting contrast maps underwent cluster thresholding. If any visible noise/clusters existed outside of the MNI template brain prior to running the plugin it was masked with VMR >10 to avoid any being included in the estimation. This was run for 1000 iterations at $p = 0.0005$ (uncorrected for multiple comparisons). Any clusters that remained were converted to Volumes-of-Interest (VOIs) using the minimum cluster size output from the estimator as the threshold for small cluster suppression. Brain labels were found for the clusters using the Automated Anatomical Labelling atlas 116 (Tzourio-Mazoyer et al., 2002).

A Region-of-Interest (ROI) GLM analysis was conducted with the resulting VOIs. The GLM specifications were separate subject predictors, z-transformation, and serial correlation correction. No contrasts were included. The resulting beta values were then extracted from BrainVoyager as text files so that they could be used in R to calculate Pearson linear correlation analyses with the psychometric data to address RQ3.

Results

RQ1 - Functional Brain Results

From all of the GLM contrasts (Table 1) 5 different significant clusters were found. In the right middle frontal gyrus (Figure 2A) a cluster was found for the contrast human trust – computer trust. No significant clusters were found for human trust – AI trust, or AI trust – computer trust. In the right middle temporal gyrus a cluster was found (Figure 2B) for the contrast human reciprocation – AI reciprocation. A cluster was found in the left caudate nucleus (Figure 2C) for the contrast human reciprocation – computer reciprocation. In the right caudate nucleus (Figure 2D) there was a

cluster for the contrast AI reciprocation – computer reciprocation. Finally, a cluster stretched over the left cerebellum areas 7b and 8 (Figure 2E) for the contrast human betray – AI betray. No significant clusters were found for human betrayal – computer betrayal, or AI betrayal – computer betrayal. No significant clusters were found for any comparisons of mistrust behaviour. Only the contrasts that yielded significant clusters were examined more closely through an ROI analysis.

Figure 2A

Cluster in right middle frontal gyrus as a result of human trust – computer trust GLM contrast

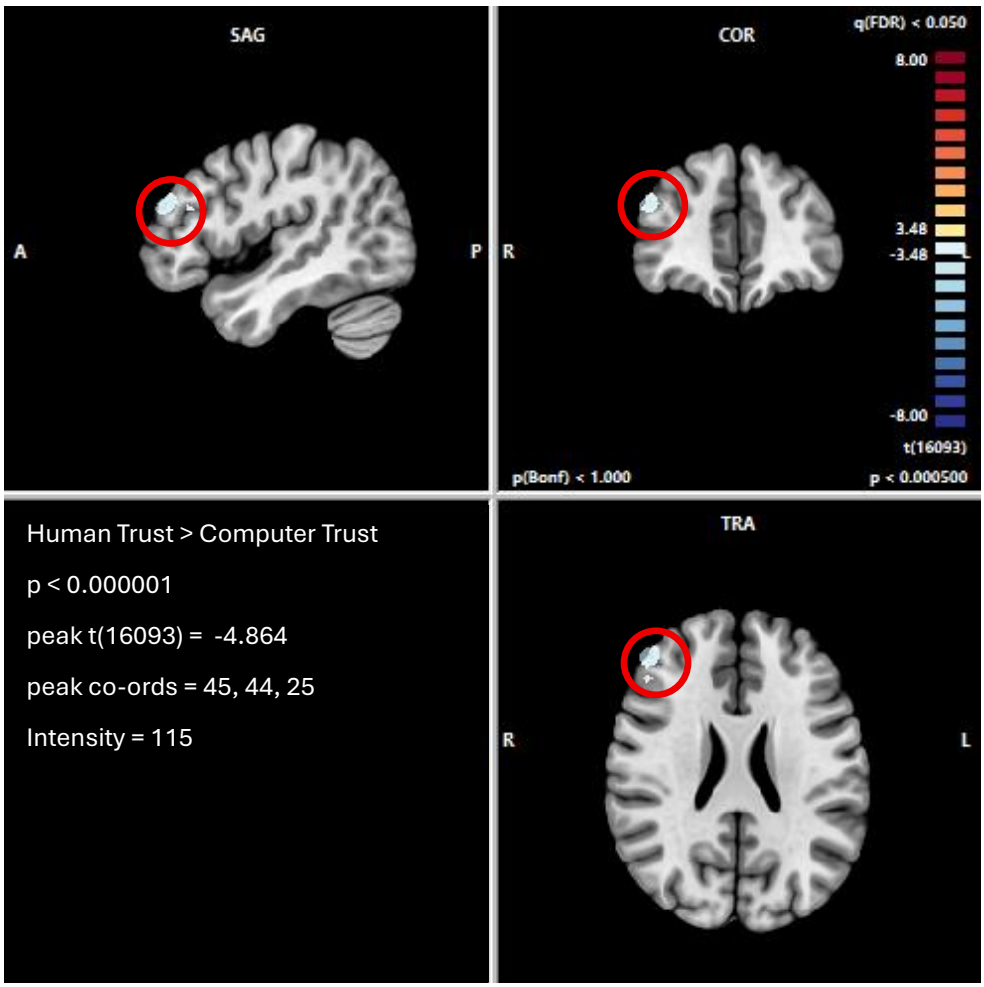


Figure 2B

Cluster in right middle temporal gyrus as a result of human reciprocation – AI reciprocation GLM contrast

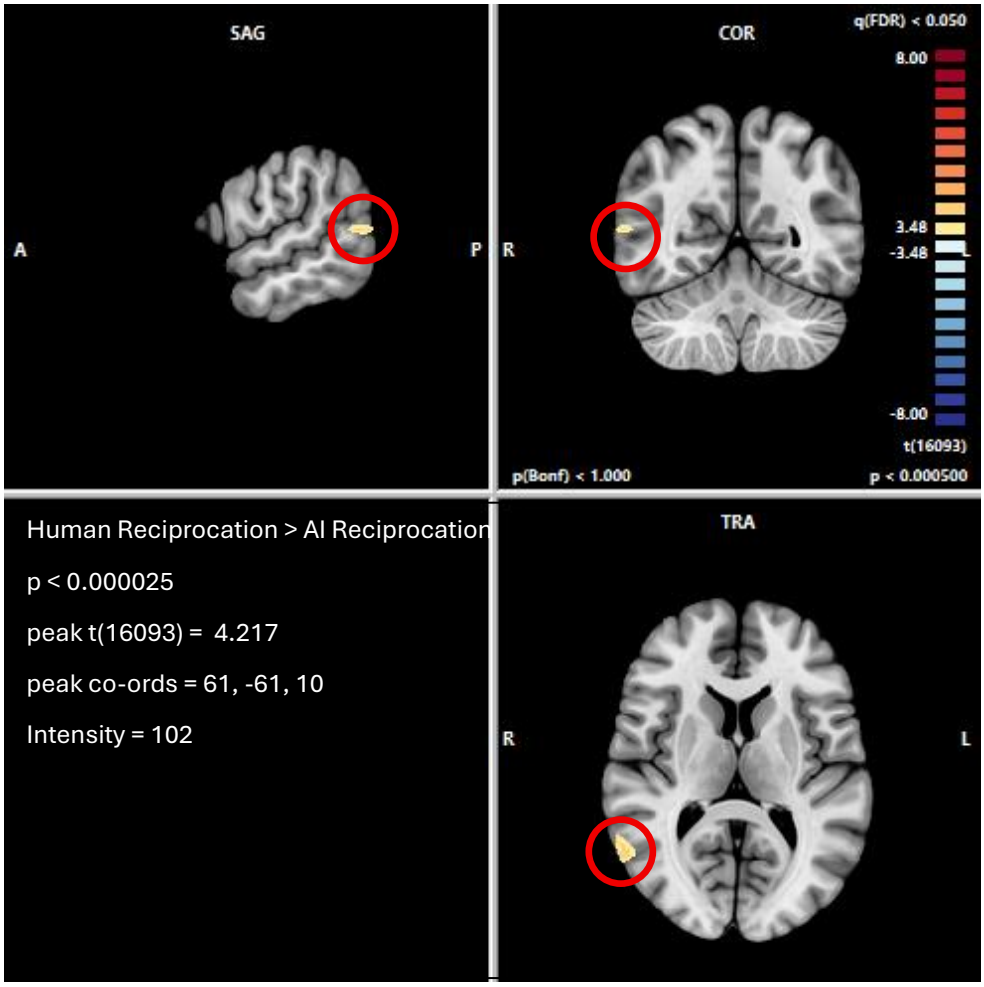


Figure 2C

Cluster in left caudate nucleus as a result of human reciprocation – computer reciprocation GLM contrast

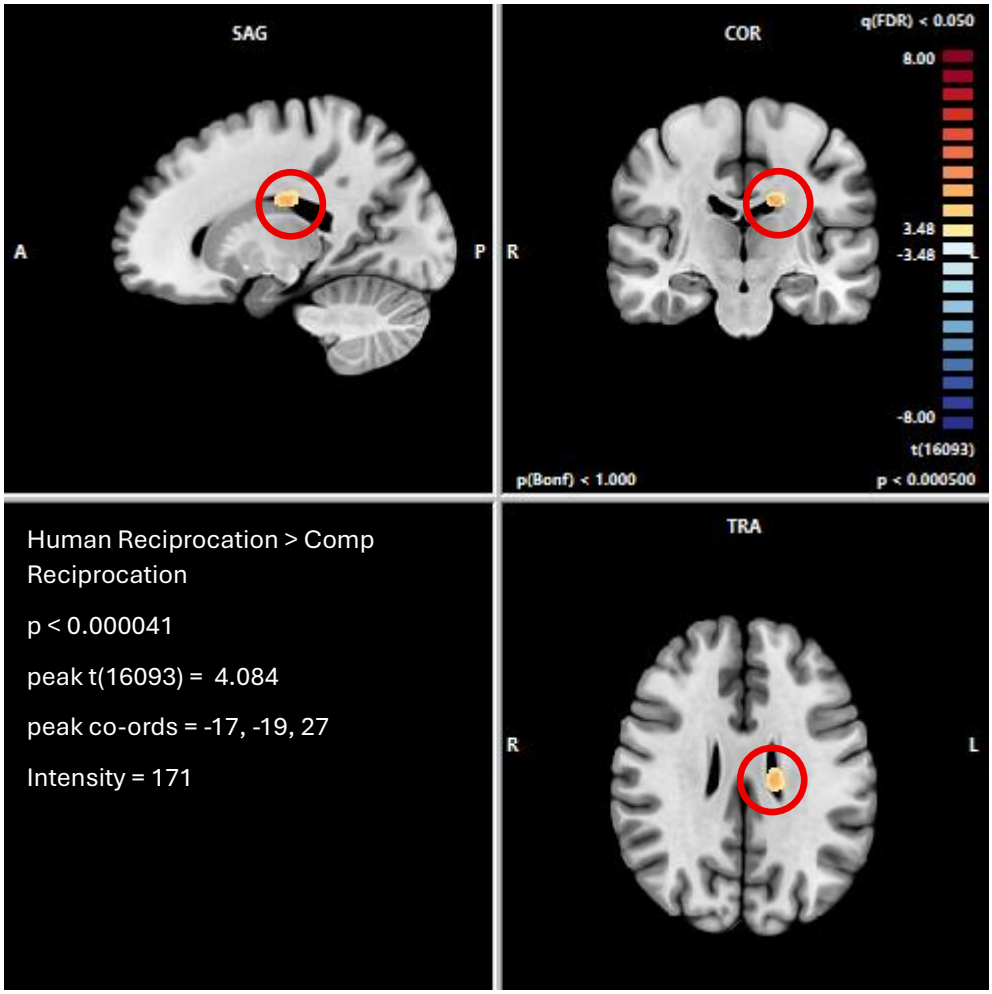


Figure 2D

Cluster in right caudate nucleus as a result of AI reciprocation – computer reciprocation GLM contrast

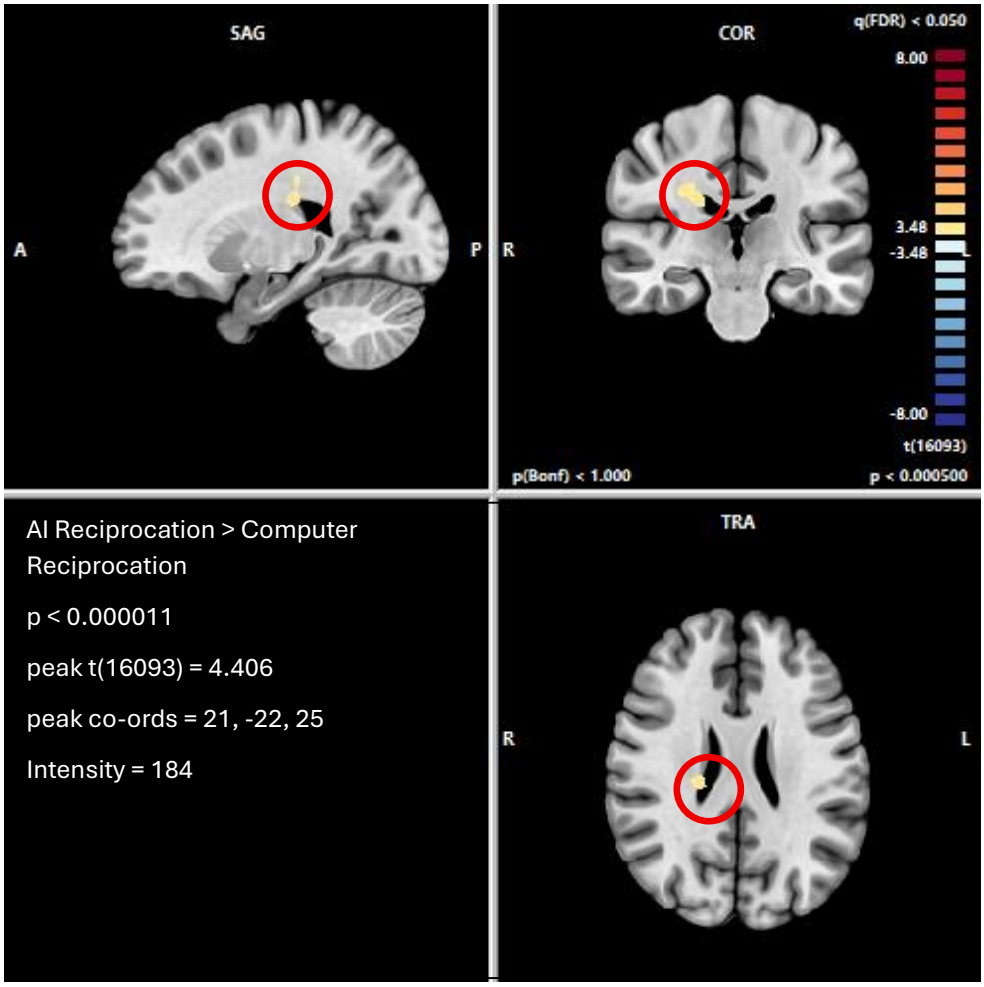
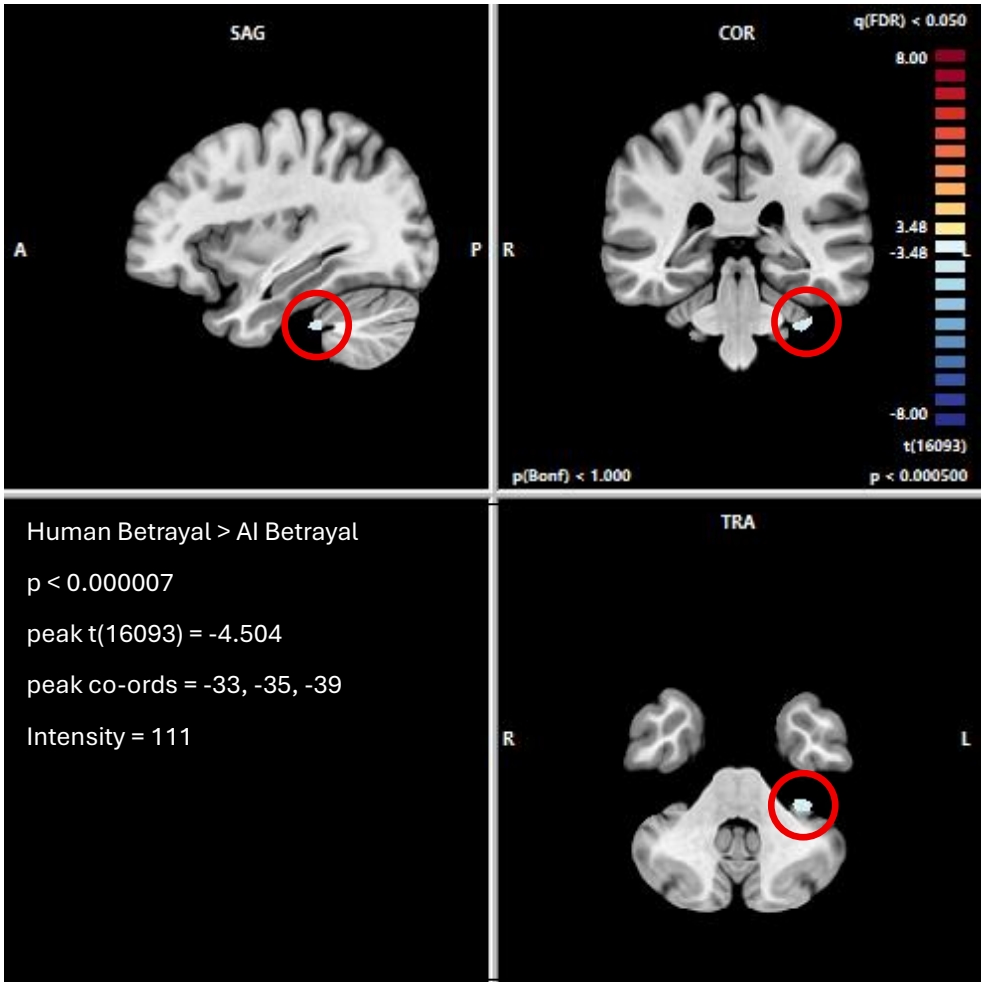


Figure 2E

Cluster in left cerebellum regions 7b and 8 as a result of human betrayal – AI betrayal GLM contrast



Note. All clusters are circled in red. Each cluster is presented on an MNI ICBM152 2009 T1 weighted Nonlinear Asymmetric Template Brain.

RQ2 - Behavioural Results

The Propensity to Trust Automated Technology (PTT) (M = 21.07, SD = 4.79) was not significantly ($p > 0.05$) correlated with any of the subscales of the Attitude Towards AI (ATAI). The ATAI Fear (ATAI_F) subscale (M = 14.80, SD = 5.49) was significantly ($p = 0.011$) negatively correlated (-0.63) with the Have/Would (H/W) subscale (M = 4.20, SD = 1.42). The ATAI Acceptance (ATAI_A) subscale was not significantly ($p > 0.05$) correlated with any of the other subscales of the ATAI (Table 2). None of the psychometric measures correlated significantly with Age or Sex (Table 3).

In the trust games participants showed a wide range of preferences when making a decision. This is measured by what percentage of times they made a decision (e.g. trust in a human) compared to how many they could have (Table 4). The widest ranges came when participants were playing as M2 with an AI, closely followed by the

decision to mistrust a computer and then by reciprocating the trust of a human. The next widest range was in betraying a human and then reciprocating a computer. Then there is a sharper drop in range when trusting a computer. The percentage of decisions participants made when trusting an AI, and when trusting a human are similar. Then, with the lowest ranges of decisions are mistrusting an AI, betraying a computer, and mistrusting a human. Mistrusting an AI and mistrusting a human had the lowest mean and general percentage incidence out of all potential decisions. Betraying a computer had the highest, followed by trusting a human and then trusting an AI.

The analysis between the psychometric measures and the summed decisions participants made, addressing RQ2, yielded some significant correlations (Table 5). The PTT was not significantly correlated with the number of times participants made any choices. The ATAI_F was very significantly negatively correlated with reciprocating an AI ($\rho = -0.68$), and significantly negatively with reciprocating a human ($\rho = -0.58$) (Figure 1). The H/W Scores were very significantly positively correlated with reciprocating an AI ($\rho = 0.77$), and further correlated with reciprocating a computer ($\rho = 0.87$), but did not significantly correlate with human reciprocation (Figure 2). The ATAI_A did not significantly correlate with any choices. Additionally, participant age significantly positively correlated with the decision to betray a human ($\rho = 0.61$), and negatively with reciprocating a human ($\rho = -0.58$) (Table 6) (Figure 3).

Figure 1
Significant Correlations between the Fear subscale of the Attitude towards AI Inventory and Sum Number of times participants made a decision in the Trust Game

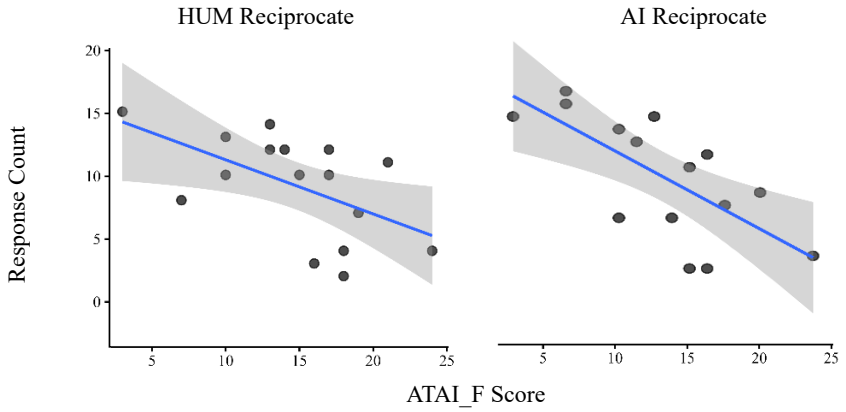


Figure 2
Significant Correlations between the Have/Would Scores of the Attitude towards AI Inventory and Sum Number of times participants made a decision in the Trust Game

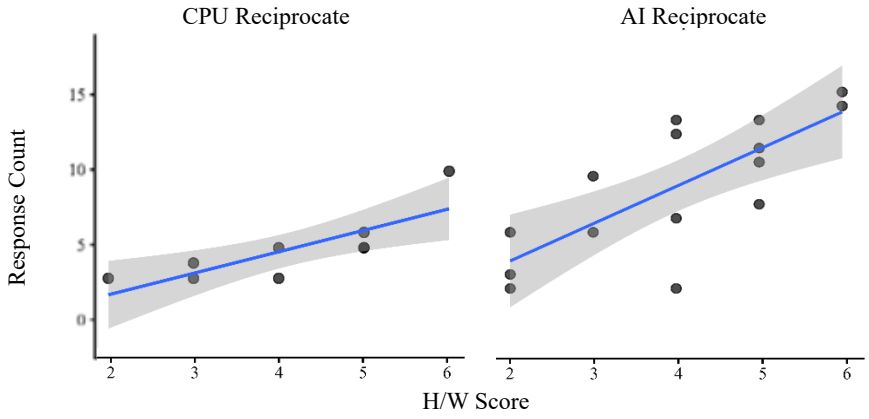
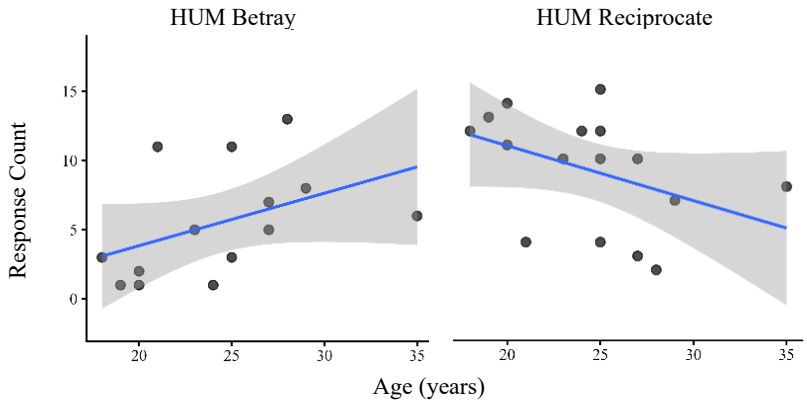


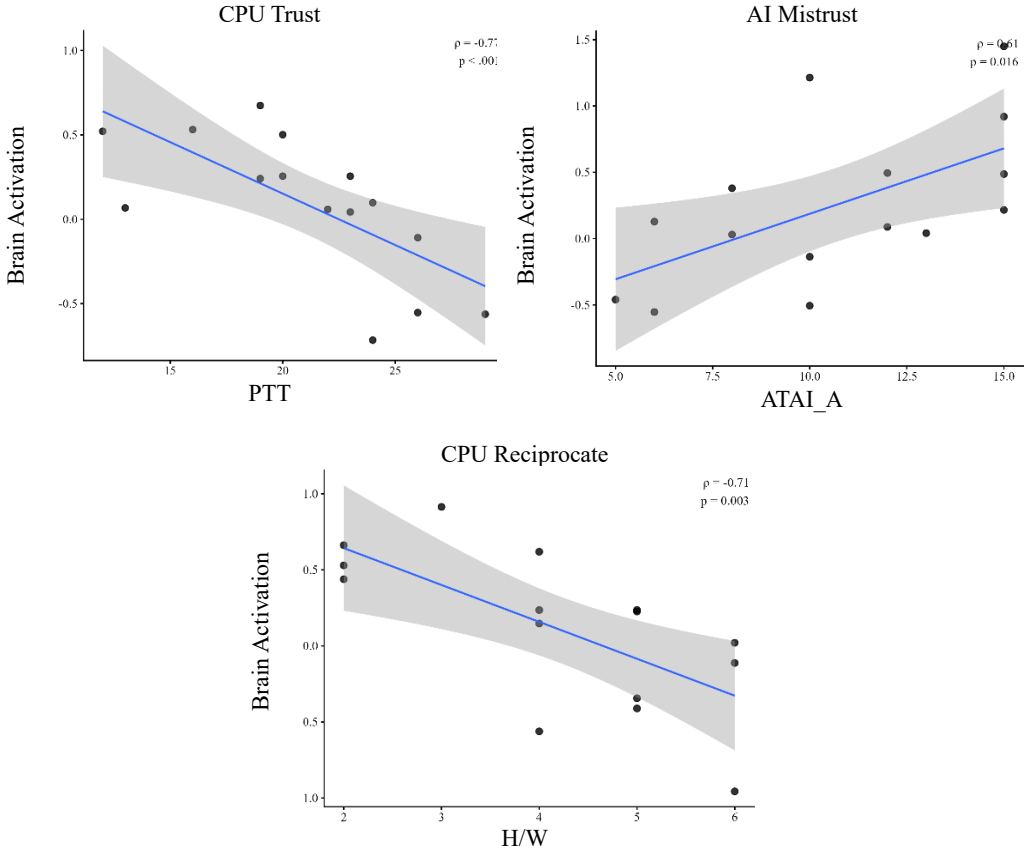
Figure 3
Significant correlations between Age and Sum Number of times participants made a decision in the Trust Game



RQ3 - Brain – Behaviour Correlation Results

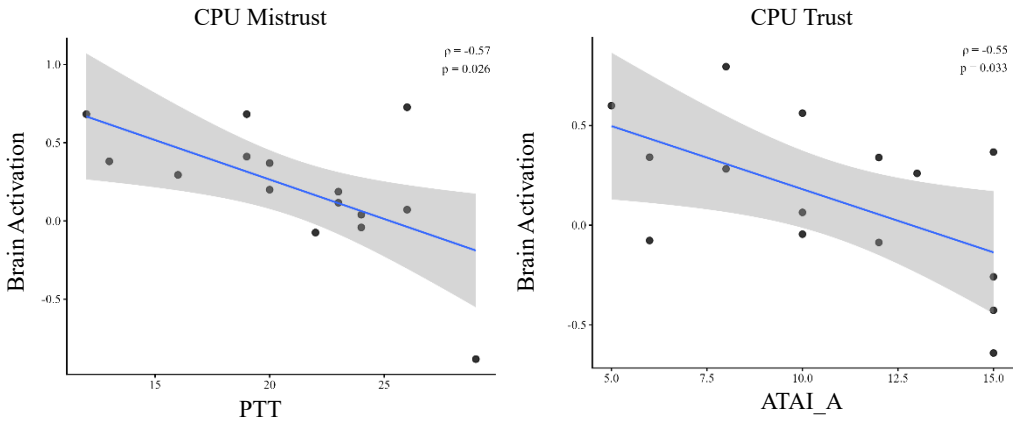
After running the ROI analysis and extracting the beta values, they were read into R for Pearson linear correlation analyses with the psychometric data to answer RQ3. In the right middle frontal gyrus (Table 7) (Figure 4) the PTT
was very significantly negatively correlated with the decision to trust a computer (ρ = -0.77), the ATAI_A was significantly correlated with mistrusting an AI (ρ = 0.61), and the H/W Scores were very significantly negatively correlated with reciprocating a computers trust (ρ = -.71). There were no correlations with the ATAI_F.

Figure 4
Significant correlations between Brain Activation in the right Middle Frontal Gyrus when making a decision in the Trust Game and Psychometric Results



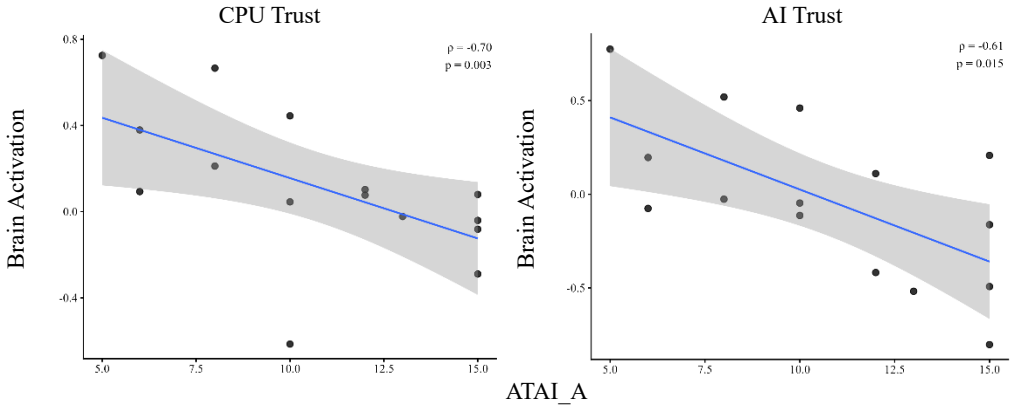
In the right middle temporal gyrus (Table 8) the PTT was significantly negatively correlated with mistrusting a computer (ρ = -0.57), the ATAI_A was significantly negatively correlated with deciding to trust a computer (ρ = -0.55) (Figure 5). The H/W Scores was not significantly correlated any decision, and again, there were no correlations with the ATAI_F.

Figure 5
Significant correlations between Brain Activation in the right Middle Temporal Gyrus when making a decision in the Trust Game and Psychometric Results



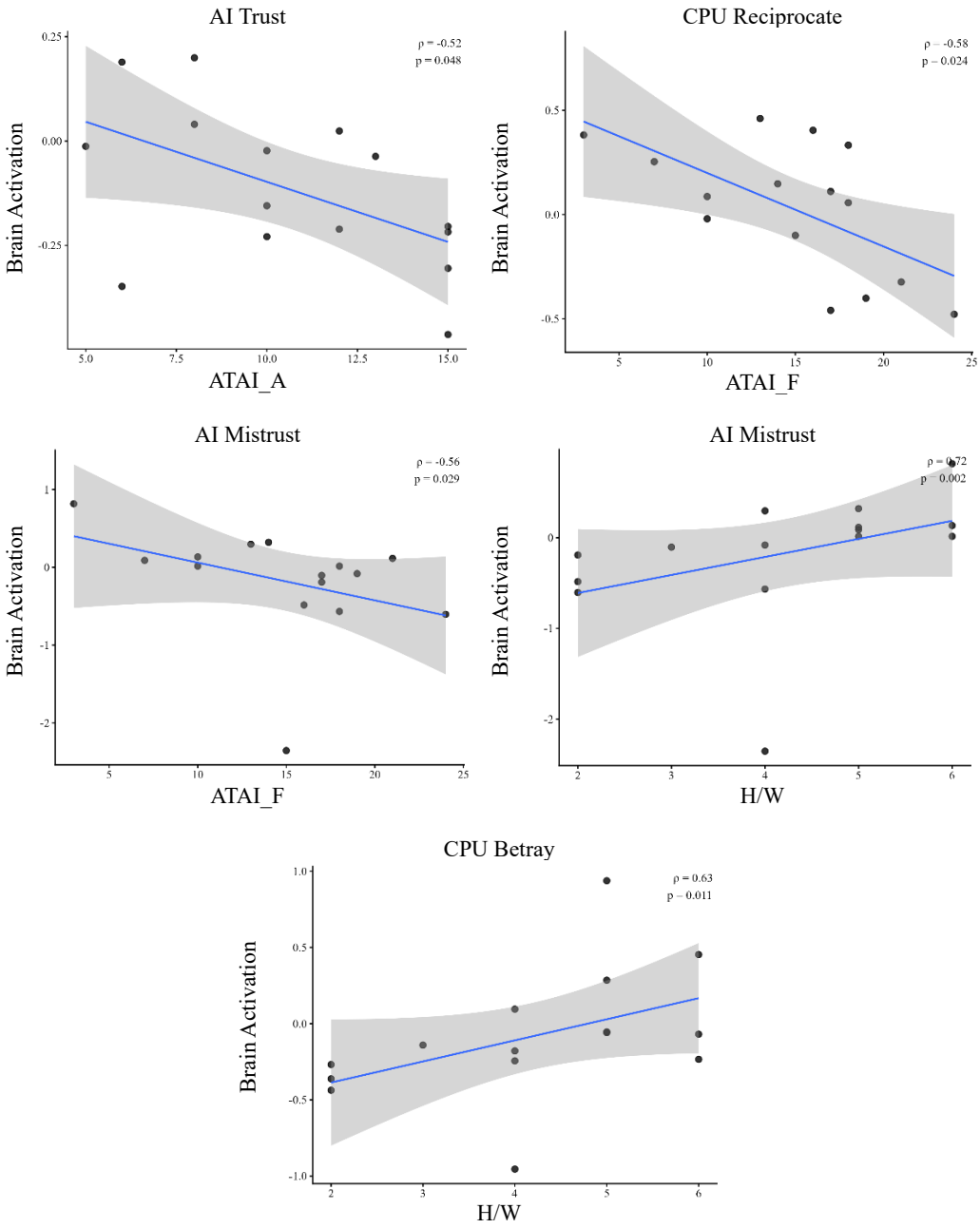
In the left caudate nucleus (Table 9) the only significant correlations were the ATAI_A being significantly negatively correlated with both trusting an AI (ρ = -0.61), and trusting a computer (ρ = -0.70) (Figure 6).

Figure 6
Significant correlations between Brain Activation in the Left Caudate Nucleus when making a decision in the Trust Game and Psychometric Results



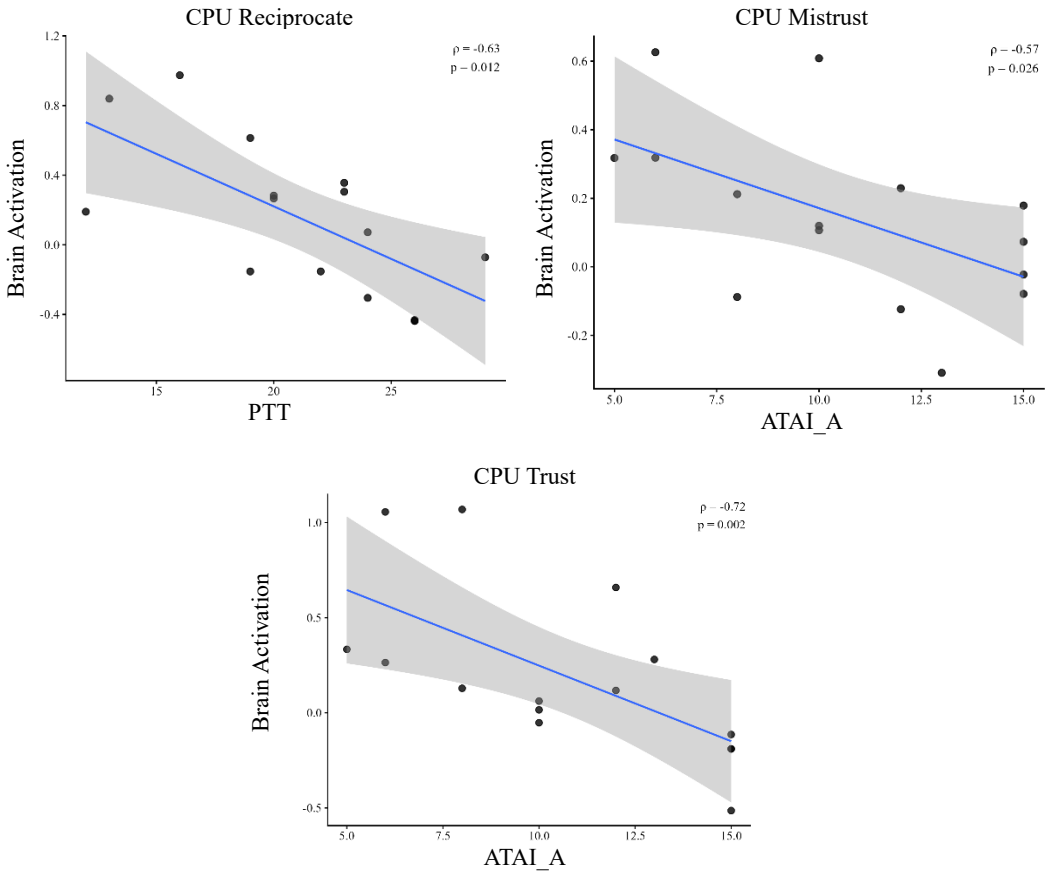
In the right caudate nucleus (Table 10) there were no significant correlations with the PTT. With the ATAI_F there was a significant negative correlation with the decision to reciprocate with a computer (ρ = -0.58), and to mistrust an AI (ρ = -0.56) and with the ATAI_A there was a significant negative correlation with trusting an AI (ρ = -0.52). The H/W Scores were very significantly positively correlated with the decision to mistrust an AI (ρ = 0.72) and significantly with the decision to betray a computer. (ρ = 0.63)

Figure 7
Significant correlations between Brain Activation in the right Caudate Nucleus when making a decision in the Trust Game and Psychometric Results



In the left cerebellum 7b/8 (Table 11) the PTT was significantly negatively correlated with deciding to reciprocate a computer (ρ = -0.63), the ATAI_A was very significantly negatively correlated with deciding to trust a computer (ρ = -0.72), and significantly with the decision to mistrust a computer (ρ = -0.57). The ATAI_F and H/W Scores had no significant correlations.

Figure 8
Significant correlations between Brain Activation in the left Cerebellum 7b/8 when making a decision in the Trust Game and Psychometric Results



Discussion

We used event-related fMRI to investigate the neural correlates of behaviour during a series of binary trust-games when playing with an AI compared to when playing with another human and with a computer/lottery. We found five regions in which there were small but significant differences when choosing to mistrust, trust, reciprocate, or betray depending on opponent. We also used the Propensity to Trust Automated Technology Scale (PTT), and the Attitudes towards AI Scale (ATAI), to investigate the effects of participants' beliefs about AI and automated technology. The ATAI has three sub-scales: acceptance towards AI, fear towards AI, and whether participants have, or would use a variety of AI tools or products. The psychometric results were correlated with beta-values that had been extracted from the five significant activation clusters. The psychometric results were also correlated with the summed number of times each participant made a choice to mistrust, trust, reciprocate, or betray each opponent type.

In this discussion I will evaluate the results that relate to each cluster in turn. I will be discussing the roles that each of the five brain regions are utilised in, as well as the observed correlations between each cluster and the psychometric results. I will then discuss the correlations between the psychometrics and the choice options of the trust games. I will finish the discussion by considering the limitations of this study, as well as my recommendations for future research.

Human trust – Computer trust: The right Middle Frontal Gyrus

First, when a participant was in the M1 condition, choosing to trust another human was associated with decreased activation in a cluster of the right middle frontal gyrus compared to when they chose to trust a computer. This region in the dorsolateral pre-frontal cortex is utilised in working memory, attention orientation, and goal-oriented behaviour planning (Damiani et al., 2020) as part of the central executive network, also known as the fronto-parietal network. Structural and functional features in this region/network have previously been found to predict individual differences in propensity to interpersonal trust. This is potentially due to its roles in risk-benefit calculations and in assessing in-group versus out-group characteristics, or potentially due to a link with feelings of guilt (Chen et al., 2023; Feng et al., 2021; Gifuni et al., 2017; Krueger & Meyer-Lindenberg, 2019; Sweijen et al., 2023). In previous human-machine interaction studies, decreased activation is also seen when identifying human faces compared to digital avatar faces (Cheetham et al., 2011), which may suggest that the utilisation of this region for assessing in-group and out-group characteristics may also apply to computer/machine agents.

The right middle frontal gyrus has been found to have decreased activation when playing a prisoner's dilemma game with a human compared to a computer. However, the left middle frontal gyrus did show the opposite effect. This may be due to an increase of attentional resources used when thinking about a computers strategy, compared to a humans which may illicit the use of more emotional resources. (Rilling et al., 2004; Zou et al., 2025). This lateral difference of increased left hemisphere activation for human conditions also occurs when playing rock paper scissors with a human compared to an AI (Chaminade et al., 2012). Comparatively, significantly increased activations in the bilateral middle frontal gyrus have also been found when perceiving vitality, or affect, in human voices compared to robot voices (Di Cesare et al., 2016). The lateralisation of human and machine trust seen in the middle frontal gyrus when looking at the wider literature is interesting, as another hemispheric asymmetry in this region is that of literacy, in the left, and numeracy, in the right (Koyama et al., 2017). Perhaps the conception of machines as being numerical and logical means that increased numerical processing is utilised when interacting with a computer, compared to semantic considerations with a human. Additionally, the right frontal gyrus has been found to be involved in decreasing utilisation of advice when it is from an advisor that makes a high number of errors, and that this is related to whether the advisor is a human or a computer (Goodyear et al., 2016).

Relating to our self-reported questionnaire results in the Propensity to Trust Automated Technology (PTT) and Attitude towards AI (ATAI), activations in the right middle frontal gyrus when making the decision to trust a computer very significantly negatively correlated with the PTT. Due to the known role of the middle temporal gyrus in propensity to trust other humans, this is not surprising. The PTT not also being significantly related with trusting an AI is interesting and could support the categorisation of AI being a demonstrably different construct than a computer in this experiment, and one that the PTT may not be able to predict. The ATAI_A subscale was significantly positively correlated with brain activity in this cluster when the participant was mistrusting an AI. This could be a sign of participants who trust AI more having to utilise more resources when making a decision that goes against their general beliefs compared to participants who do not trust AI in general. The H/W Scores were very significantly negatively

correlated with activation when reciprocating a computer's trust. This suggests that differences in participants previous AI experience and expectations led to a different amount of activation in the right middle frontal gyrus when experiencing computer trust. This may relate to working memory processes, potentially including the participants' considerations of the known mathematical chance (50%) of the computer showing trust behaviour at all.

Human reciprocation – AI reciprocation: The right Middle Temporal Gyrus

All three comparisons of M2 reciprocation behaviour returned significant clusters. We found increased activation in the right middle temporal gyrus when participants chose reciprocal behaviour towards a human compared to towards an AI. The middle temporal gyrus is, amongst other roles, involved in sensory and semantic (the storage, management, and categorisation of meaning, especially related to language) processing, and in the predicting and processing of expected outcomes compared to actual outcomes (van Kemenade et al., 2019). Interestingly, when a task is more difficult, this semantic processing may get relayed onto higher-level regions, one of which is the middle frontal gyrus (Laufer et al., 2011), the first cluster in our study. In addition, the processing of mismatched expected/actual outcomes are transmitted to the cerebellum (van Kemenade et al., 2019), which is another cluster found in this study. The role of the middle temporal gyrus in mismatched outcome expectation varies depending on whether the cause of the mismatch is due to human error or computer error, with the former producing increased bilateral activations compared to the latter (Desmet et al., 2014). Besides outcome prediction, in the interpersonal relations literature, increased activations have been found in the right middle temporal gyrus when involved in the processing of feelings of social inclusion, and especially in attentiveness towards others positive internal states (Lai et al., 2016). Increased bilateral middle temporal gyrus activations have been found when making cooperative choices in a prisoner's dilemma game when compared to uncooperative, and when comparing the viewing of uncooperative versus cooperative game results (Zou et al., 2025). It is possible that increased lateral activations in the left hemisphere for this region relate to feelings of guilt and moral decision making (Gifuni et al., 2017). This role of the middle temporal gyrus may also apply to non-human interactions, as one study found lateralised activations when viewing immoral behaviour towards animate (human/animal, left hemisphere lateralisation) or an inanimate (object, property, etc., right hemisphere) beings (Chin & Yusoff, 2023). These feelings of guilt or morality may play a role in deciding to reciprocate the trust of another player in an economic game. They may originate from attentiveness to others' emotions (theory of mind), for example, if you feel the other player in a trust game will be upset by your decision.

In the human-machine interaction literature, the role of attentiveness towards others internal worlds may be seen in the form of humans eliciting activation in regions related to theory of mind than computers. In the same Di Cesare et al., (2016) study as above, there were bilateral increases in activations when comparing human voices to robots, additional left hemisphere lateralised increases in activations occurred when listening to either robot or human voices when compared to silence. This may additionally suggest that although humans give rise to more theory of mind reasoning than robots, that robots do still prompt some theory of mind. Left hemisphere activation increases were also

seen in a prisoner's dilemma when the faces of opponents were introduced as a human with free choice compared to a pre-programmed nonintentional player who had to follow instructions from a computer (Singer et al., 2004). In a different prisoner's dilemma experiment, however, increased activations were only seen in the right hemisphere when comparing playing with another human (a confederate in the same room) with playing with a computer (Krach et al., 2009). This difference could be due to emotional processing of a person being physically present and moving, rather than just represented as a photograph on a screen. When looking at expressions and eye direction represented on real, compared to computer generated, faces there were increased bilateral middle temporal gyrus activations. Both expression and eye gaze direction are very strong social signals of a person's feelings and intent in an interaction, i.e., an indication of what they may be thinking (Gobbini et al., 2011; Kätsyri et al., 2020). In our experiment, it is possible that the middle temporal gyrus is utilised differently across human and AI conditions due to the participant's conception of whether the other player would be upset by their betrayal or, even if the AI did not experience emotion, whether this would impact it's future behaviour due to learning.

Activation in the right middle temporal gyrus when making the decision to mistrust a computer was significantly negatively correlated with the PTT. This could relate to participants' predictions of game outcome being at odds with their usual beliefs about computer trustworthiness. If a participant generally believes that machines are trustworthy, but they know that the computer condition in this game has only a relatively low random chance at reciprocation (the ideal/expected result), they may have to utilise more outcome-prediction resources compared to participants with low PTT scores. Again, the PTT not having any significant relations with the AI conditions supports the evaluation of them as separate entities. With the roles of the middle temporal gyrus in outcome planning/expectation, could come from the differences in known versus unknown behaviour patterns. The ATAI_A was significantly negatively correlated with activation when deciding to trust a computer and when betraying a human. With the ATAI_A being negatively correlated with trusting a computer, this could be at odds with the negative correlation between the PTT and mistrusting a computer, though it could just be confirmation that these questionnaires measure different constructs. The decreased activation in the middle frontal gyrus comparative to high ATAI_ T scores suggests that participants who already hold positive beliefs require less processing in the middle temporal gyrus of potential risks from trust behaviour due to a lower expectation of betrayal. The ATAI having any correlations with activations when playing trust games with a human was surprising, but this result could potentially come from the items of the scale that ask whether AI will benefit/destroy humankind. Either of these items could inadvertently be asking participants' beliefs about humanity such as innovativeness and resilience, or passivity and helplessness. The middle temporal gyrus is known to be linked to feelings of guilt or moral behaviour, so this region having activations relative to betraying a human is expected, but without further exploration it is impossible to determine exactly how this may or may not be linked to the ATAI scale. The observed relationship between the ATAI_A and brain activations in the right middle temporal gyrus may be explainable should further research occur.

The H/W Scores were significantly correlated with right middle temporal gyrus activations when betraying a computer. This could be explained by participants who have used or would use a lot of AI tools or products, maybe doing so because they find them beneficial. That positive association of AI tools could be carried over to other machine agents, meaning some feelings of guilt may occur when betraying the computer. This would be unexpected, as the computer condition was not predicted to elicit any theory of mind. Alternatively, due to participants previous experiences with AI, it may have been that they had different expectations of the computer condition, and this caused increased activations in the middle temporal gyrus. In this region, no significant correlations were found between questionnaire result and any choices made when playing a trust game with an AI opponent. This is interesting as the cluster originally was found from a contrast comparing human and AI opponents. This means that despite there being significant middle temporal gyrus activations relating to AI, that the cognitive driver of these differences is not in any significant way explained by the psychometrics used in this present study. This could be worth further exploration in terms of whether these drivers are more sub-conscious and as such misaligned with self-report, or whether this is related to theory-of-mind processing in the MTG relative to AI, which is not covered by any of the scales used.

Computer reciprocation – AI reciprocation and Computer reciprocation – Human reciprocation: The left Caudate Nucleus

When comparing reciprocation behaviour towards a computer, there were significant clusters of activation of the caudate nucleus in both hemispheres. When reciprocating a human, increased activation was in the left hemisphere, and when compared to reciprocating an AI the increase was in the right. The caudate nucleus, together with the putamen, makes up the dorsal striatum, which is itself the largest structure in the basal ganglia. The caudate is linked to functioning of working memory, learning, reward-based decision making, intentional movement, and as a controller of cognitive and emotional processing (Graff-Radford et al., 2017; Krueger & Meyer-Lindenberg, 2019). In these functions it is often considered jointly with the putamen (Driscoll et al., 2025). In terms of interpersonal trust, activation in the caudate/striatum has been shown to bilaterally increase in response to known bad reputation compared to known good reputation of strangers in a trust game paradigm (Wardle et al., 2013). Noted for relevance to this report, while not providing any information, is that in the Wardle et al. (2013) experiment, a computer with a known 50% probability was also included as a control but not reported on past the methods section. The role of the caudate in known reputation has also been found when playing a trust game matched with a close friend, a community member, and an unknown peer. Increased activations were found bilaterally when trusting a friend rather than a stranger (Sweijen et al., 2023). It is also likely that trusting a friend will yield more expectation of reward than trusting a stranger. The development of reputation may have some lateralisation in the left caudate, as more activation occurs there when there is no prior knowledge of a trustee comparative to when there is, though activation in the right caudate was also present (Fouragnan et al., 2013). Additionally, this study found that caudate activation decreases in response to trust violations, especially in cases of prior knowledge. Decreased activation, however, was not necessarily followed by decreased trust behaviour, showcasing how powerful reputation can be as a driver of behaviour. As well as reputation, or perhaps related to it, activations in the right dorsal striatum are

modulated by prior expectations of reward in multi-round trust games (Fairley et al., 2019).

In a study investigating participants punishment of defectors in a trust game, higher right caudate activations were found when participants chose a more satisfying punishment, even at their own detriment, and that this was the case even when the betrayal came from a random computer (de Quervain et al., 2004). Additionally, higher activations in the caudate nucleus related to participants being more willing to incur higher costs for the purpose of satisfying punishment. Supporting the role of the right caudate in reward anticipation also applying to interactions with non-human entities, a prisoner's dilemma experiment found that when desired reward aligns with actual outcome, there are higher activations in the right caudate compared to when they do not occur (Rilling et al., 2002). This study also found right caudate activations to relate to cooperation maintenance, or reciprocation. However, when comparing cooperation with all other potential results when playing with a computer, it did not find any significant caudate activations (Rilling et al., 2002). The study is from 2002, so most participants would be unlikely to have experience with AI and as such minimal expectations of its behaviour compared to if this study was replicated now. Another prisoner's dilemma experiment two years later found increased activation in the caudate nucleus when looking at human faces compared to an image of a roulette wheel that represented the computer trials (Rilling et al., 2004). Also in the non-human interaction literature, the right caudate (and bilaterally the putamen) has been found to be sensitive to shifts from human faces toward computer generated avatar faces (Cheetham et al., 2011). The authors predicted this was due to the role of the striatum in learning and categorisation. In an AI advice-taking task, the right caudate nucleus showed significant increases in activation when comparing trials in which participants felt more confident, an effect also seen in the middle frontal gyrus (Valeriani et al., 2022).

In the literature surrounding human interactions with non-humans, the right caudate nucleus seems to have more relevance to non-human interaction decision making than the left. For example, when looking at expressive robot faces, which are similar to avatar faces but differ by not attempting to look convincingly human in anything but expression, there is increased right caudate activation compared to looking at human faces, and both human and robot faces caused increased right caudate activation compared to rest conditions (Gobbini et al., 2011). This is supported by the results of this study in which the right caudate was more active when comparing AI and computer reciprocation, and the left caudate was more active when contrasting human and computer reciprocation. The increased activation in the right caudate when comparing reciprocating an AI versus a computer could be related to different preconceptions, or reputations, between the two thanks to the known probability of the computer's decisions. When the participant is trusted by the computer, something they know will happen a set number of times compared to the AI with unknown parameters, participants may not have to use as many resources to decode and learn from this behaviour. In our psychometric results, right caudate brain activation when deciding to reciprocate a computer was significantly negatively correlated with the ATAI_F. This could be due to some kind of interaction similar to the one described above, though that would perhaps elicit a positive correlation, in which higher ATAI_F

scores related to higher caudate activation. Alternatively, fear could act as a kind of inhibitor to feelings of reward represented in the caudate nucleus when the computer acts in a prosocial way. Of note, this is the only time across all five brain regions that the ATAI_F significantly correlates with any decision. Also in the right caudate, activation was significantly negatively correlated with the ATAI_A when trusting an AI. This could result from participants who express higher general trust in AIs having reduced consideration of risks in deciding to trust AI, and as such showing lowered inhibitory processing activations.

In the left caudate nucleus the only significant correlations with any psychometric were with the ATAI_A, which significantly negatively correlated with brain activations when trusting both an AI, and trusting a computer. The ATAI_A is significantly correlated with activation in both hemispheres of the caudate nucleus when trusting an AI. The ATAI_A being also linked to activations in the left caudate nucleus when trusting a computer could suggest that the the AI and computer constructs are not as well differentiated as we predicted, or it at least suggests that some of the mechanisms underlying decisions to trust both are related. Perhaps a higher regard for AI (as shown by the high ATAI_A score) is a sign that someone is more optimistic about technology in general and expects higher rewards from trusting it. Neither the PTT nor the H/W scores significantly correlated to activation in either hemisphere of the caudate relative to any decision in the trust games.

Human betrayal – AI betrayal: Lobules VIIb and VIII of the Cerebellum

The final cluster was in the left posterior cerebellum, and it extended over lobules 7b (VIIb) and 8 (VIII). When betraying a human there was significantly decreased activation compared to betraying an AI. The cerebellum is a small but incredibly neuron-rich area of the brain, containing 80% of the brains neurons (Jimsheleishvili & Dididze, 2026). Much of the reports of cerebellar function relate to motor movement, balance regulation and sensorimotor control, but the cerebellum also has roles in prediction, error-based learning, and other cognitive functions including the regulation of fear and pleasure responses (Schmahmann, 2019; Sokolov et al., 2017; Wright et al., 2020). Utilisation of the posterior cerebellum has been found in social language learning and attribution of traits (Pu et al., 2020). Lobule VIIb, specifically, is involved in the receiving of visual and auditory stimuli, and its functional connectivity both within the cerebellum, and with other cortical structures, may specifically be related to social cognition relating to body language, in creative thinking, and in the process of executive functions, such as working memory and multi-tasking (Gao et al., 2020; Nettekoven et al., 2024; Schmahmann, 2019; Sokolov et al., 2017). While lobule VIII is definitely involved in somatosensory representations, including for the face, this region also has links to language, to action observation, monitoring, and imagery, and functional connectivity of this region has also been found to relate with creativity (Gao et al., 2020; Nettekoven et al., 2024; Schmahmann, 2019; van Kemenade et al., 2019). To simplify, both of these cerebellar regions are involved in feed-forward and feed-back processing of sensory, sensorimotor, and cognitive stimuli.

Despite the various lobules of the cerebellum having a wide range of roles, exact lobule reporting is not very common in literature that is not specifically about the cerebellum, and even in some literature that is (Wright et al., 2020). In interpersonal trust, there may be bilateral structural links with grey matter volume and self-reported propensity to trust, as measured by the trust facet of the NEO personality inventory, and the evaluation of trustworthy faces (Haas et al., 2015). In addition, the right cerebellum has been found to have increased activation relative to bad reputation, in a way that is moderated by decision to mistrust, rather than invest (Wardle et al., 2013). This may be due to the role of the cerebellum as a moderator of reward processing, loss aversion, and sensitivity, especially in lobule VII (Manto et al., 2024). The cerebellum may evoke this moderative behaviour through the construction of internal models, outcome prediction, and the comparison of predicted versus actual outcomes.

The cerebellum appears semi-frequently in the human-machine interaction literature. The role of the cerebellum for loss aversion could be supported by differential activations in a prisoner's dilemma game when looking at human faces compared to the representation of a random computer, and when playing the game with these different opponents also (Rilling et al., 2002, 2004). The activation effects may be moderated by gender of the participant, but activation did not moderate game outcome. The presence of activations in the left cerebellar culmen when receiving advice from a human and from an AI during a luggage x-ray task could also be a result of loss aversion (Goodyear et al., 2016). Increased activations in the right cerebellum (lobules VI, VIIa, and crus I and II) when participants are confident in their AI-augmented decisions could also be indicative of loss aversion, as well as outcome prediction, moderating human responses to trusting AIs (Valeriani et al., 2022). Activations in the cerebellum crus I and II have been found to have differential activations relative to listening to and speaking to a human compared to a robot (Hogenhuis & Hortensius, 2022). When listening to a human there were increased bilateral activations in the crus II compared to listening to a robot. When speaking to a human there were increased activations in the left crus I, but when speaking to a robot there were increased activations in the right crus I. In Di Cesare (2016), bilateral increased activations in the cerebellum occurred when listening to robot voices compared to silence (and also human voices compared to silence). Bilateral increases in the cerebellum are also seen when looking at human expressive faces compared to rest, robot expressive faces compared to rest, and when comparing looking at robot expressive faces with looking at expressive human faces (Gobbini et al., 2011). In addition, the right cerebellum is activated when perceiving differences in human vs avatar faces and the uncanny valley (Cheetham et al., 2011). Activation of the cerebellum in these tasks could be due to its role in sensory and emotional processing (fear, pleasure, etc.).

In the left cerebellum lobule VII/VIII the PTT was significantly positively correlated with brain activation when deciding to betray a human. This is an unexpected relationship as the PTT is not expected to relate to any human conditions, but due to the wide variety of roles that the cerebellum codes for there could be a process, or combination of processes, that account for this activation. It is not surprising that the cerebellum has significant activations relating to betraying another human. The decision to betray may be informed or modulated by emotional processing (guilt, fear of retribution, etc.), or by reward/loss

expectations, both of which are coded for in the cerebellum (Sokolov et al., 2017). However, this being related positively to the PTT is difficult to interpret. Comparatively, the PTT was very significantly negatively correlated with brain activation when reciprocating a computer, and this is much more in line with expected results and easier to parse. If a participant has higher trust propensity, they may feel more positively towards computers/machines in general, and as such utilise less brain resources when making prosocial decisions towards machines. Specifically in this cluster, with the role of increased activation in the cerebellum relative to bad reputation, if the participant considers computers to broadly have "good reputation" it is expected that you would see decreased activations. With this in mind, it is possible, if unlikely, that the PTT inadvertently codes for negative opinions of humans, which then relates to the increased activity seen when betraying a human. This same mechanism is likely to explain the negative correlation between the ATAI_A and activation when deciding to trust a computer. Interestingly, the ATAI_A also negatively correlated with activation in the cerebellum when deciding to mistrust a computer. This decrease relative to the ATAI_A could be seen due to loss aversion and difficulty predicting the game outcome when the participant's generally positive beliefs about AI are at odds with the known low probability of having your investment reciprocated. The ATAI_F and H/W scores had no significant correlations with activity in the cerebellum when making any decision in the trust game.

Discussion of Behavioural Results

No significant clusters were found when comparing mistrust behaviour between opponent conditions. It is possible that a lack of differences between conditions when mistrusting is due to if participants were going to mistrust, that they were going to do this regardless of who they were playing with. Despite this, participants mistrusted a computer with known outcome probability close to twice as often as either mistrusting a human or an AI. Perhaps individual differences in processing the known probability meant that no strong activations in any one region could be found, and how often people mistrusted the AI and the human opponents was remarkably similar. This could be supported by results seen in previous research, where no common areas of activation have been found across noncooperators when they mistrust a human or a computer (McCabe et al., 2001). The authors of that study posited that this could be due to participants learning different techniques when playing noncooperatively with humans or non-humans, rather than trusting which requires convergent brain activation due for mutual beneficial outcomes. Following on from a lack of clusters relative to mistrust contrasts, the lack of a significant contrast when comparing AI and computer trust is interesting. The similarity of how often trusting behaviours were chosen towards an AI or a human is very high, and human – computer behavioural percentage, and trust activations were significantly different. Does this suggest that regardless of very different behaviour outcomes towards an AI and a computer, that the underlying mechanisms of those choices come from similar roots? A similar, though not as extreme, pattern appears when considering the lack of significant contrast activation relative to betraying a computer. Betraying a computer had the highest percentage of any decision being chosen across any conditions, and yet the mechanisms underlying this choice compared to the other conditions are elusive. These questions surrounding potential neural mechanism differences when choosing to mistrust, trust, and betray different opponents should be investigated in more detail.

Additionally, in the clusters that were significant, the questionnaires were most significantly correlated with brain activation when the opponent was a computer/lottery condition. Each potential choice that could be made with a computer was significantly correlated at least once, with trust and reciprocation being the most common. We were not expecting any significant explanations of activations in the human opponent conditions, though two of the scales correlated with brain activations when reciprocating. There only being three times that significant correlations were present between the questionnaires and brain activity when interacting with an AI was unexpected, especially the lack of the ATAI_F being significantly correlated with any AI behaviour choices in any cluster. It was expected that fear or uncertainty would be a significant predictor of the mechanisms underlying AI mistrust (Afroogh et al., 2024; Golgeci et al., 2025; Lockey et al., 2021). The present study suggests that either this is not as robust a prediction as we thought, or that the ATAI_F may not measure the same type of fear mechanism.
The psychometric measures not correlating in an expected way with brain activation is perhaps related to the psychometric measures not explaining trust-game behaviour in an expected way also.

The acceptance subscale of the ATAI did not correlate significantly with the number of times any participant made any choice. This is very unexpected, as the ATAI scale has previously been cross-validated with the General Attitudes Towards Artificial Intelligence Scale (Schepman & Rodway, 2023), and externally validated with actual trusting behaviour towards ChatGPT (Montag & Ali, 2025) in multiple languages (Hitsuwari & Takano, 2025; Sindermann et al., 2021). There may be some underlying complexity and nuance that is missed in the ATAI_A scale due to it being a short measure, with absolutist terms in the items (Grassini, 2023), which unfortunately may have made it unable to predict participant behaviour in the experimental task, even though the ATAI_A does predict brain activity related to task more than the ATAI_F. The ATAI_F comparatively showed very significant negative correlations with the decision to reciprocate AI trust. This was expected from the literature. However, the ATAI_F also significantly negatively correlated with the decision to reciprocate a human. As previously discussed, two of the items of the ATAI “Artificial intelligence will destroy humankind” and “Artificial intelligence will benefit humankind” may unintentionally code for opinions on other humans, as well as coding for opinions towards AI. You could believe that AI will destroy humankind due to AI becoming out of control with evil intentions, like something out of a dystopian sci-fi, or you could believe that AI will destroy humankind due to human factors like laziness or greed increasing as a result of what is otherwise a morally neutral AI tool. If the latter is true and someone believes that AI will destroy the world due to misanthropic reasoning, it makes sense that they would be less likely to reciprocate the trust of another human. Due to the shortness of the ATAI scale, single items can have large impacts on total scoring. If that item is not as well calibrated as it ideally would be, this can heavily impact the reliability and validity of the scale. The final subscale of the ATAI, relating to whether participants have or would use certain AI products or tools very significantly positively correlated with decisions to reciprocate both a computer and an AI. This subscale significantly correlating with actual behaviour is understandable, as any time a participant reports that they have used a specific AI, they are reporting explicit facts in a simple yes/no question, rather than objective feelings. However, this significant correlation

being with reciprocation behaviour is unusual. It is possible that participants who use more AI in their personal lives experience higher expectations of positive AI behaviour. They may take ideas such as machine learning and AI training into account and expect that their modelling of reciprocation will make it more likely for the AI to reciprocate in another trial. Alternatively, they may experience increased feelings of positive relationships with AI, meaning that they do not want to betray AI similar to how they would not want to betray a friend.

Finally, in the results of the present study the PTT was significantly correlated with the decision to trust an AI during the trust game, but with no other choices. This is an expected result, and in prior literature the PTT has been found to have a mediative role in self-reported technology self-efficacy (one's belief in their ability to use technology), and attitudes towards AI, as measured by the ATAI (Montag et al., 2023). In this same study, the PTT and the ATAI were found to have some shared variance, but also measure independent constructs. That has been somewhat replicated in this study, with both questionnaires explaining some of participant behaviour and/or brain activation. Sometimes correlating with the same participants decision in different contexts, though they do not correlate with each other significantly. The PTT only correlating with decisions of AI trust, and not computer trust, is unexpected, however. As the computer with known probability should relate more to the concept of automated technology that the PTT is built to measure.

Limitations, Implications, and Future Research Recommendations

This paper has a number of limitations, which may have impacted our results. Our small sample size does not meet the minimum recommendation of 20 participants for fMRI experiments, which could limit the replicability of our findings. Also, our small sample size meant that we could not use any complex models when trying to understand the interactions between the psychometric results and either the brain or behavioural data due to an already smaller-than-ideal level of statistical power (Turner et al., 2018). With more participants, and with more specific research questions formulated following the results of this exploratory study, some causal analyses could have been calculated based on predictors such as high-low psychometric scores, which could potentially have either mitigated, or even gone so far as to explain some of the irregularities in our current results (Rohrer, 2018).

It is possible, however, that even with running different types of analysis on a larger sample size, that our psychometric results would still be poorly related to the constructs of AI or computer trust as measured by our paradigm. Perhaps the PTT and ATAI scales were a flawed choice. There are other scales for assessing attitudes towards AI available, many of which have only recently been developed, or are still being improved upon. A current limitation that exists in the literature is that AI does not have a broadly agreed upon experimental definition (Gil de Zúñiga et al., 2024; AI HLEG, 2019; Samoili et al., 2020). While it is true that any machine learning algorithm utilises AI, a specialised tool that predicts maintenance required in manufacturing may utilise different

mechanisms of trust compared to a very general and accessible tool like ChatGPT that emulates social interaction. The blurred lines between the constructs of automation, AI, robotics, etc., could mean that when participants are faced with a non-specified AI condition that there are significant individual differences in how much they consider to AI to be human-like or computer-like. This could be a significant confound of trying to investigate beliefs about AI in behavioural or brain imaging studies, and a major recommendation from this study is to research in more detail the impact of a participant's personal beliefs and preconceptions about AI.

Additionally, a limitation of this study, again partly due to our small sample size, was that we did not collect more demographic data. We also did not control for the effects of gender in this study, though prior research into both trust and attitudes towards AI have found differences due to gender (Chatterji et al., 2025; Rahman et al., 2025; Wu et al., 2020) and age (Banerjee et al., 2021; Chen et al., 2023; Li & Fung, 2013). Further types of demographic data that have been found to impact trust behaviour and attitude to AI to are nationality/country of origin (Dang & Liu, 2021; Rothstein & Uslaner, 2005; Sindermann et al., 2021; Zhang et al., 2026) and education level (Banerjee et al., 2021; Golgeci et al., 2025; Ismatullaev & Kim, 2024; Mirfardi, 2011). Even if future research does not explicitly investigate the effects of these individual differences, with a big enough sample size and access to detailed demographic information, researchers can control for their potential effects. Originally, this study was going to only include British participants, but due to difficulties in recruitment the inclusion criteria had to be expanded.

Overall, this research provides a novel insight into the construct of AI trust when compared to trusting both a known computer condition, and when trusting a human. As far as the author is aware, this is the first study that compares all three of these conditions. The results of this study provide an initial exploration of variations in brain activity relative to all three opponents (human, AI, computer), and decision in an fMRI trust game paradigm. While results were mixed when considering self-reported beliefs about technology and AI relative to behaviour and brain activation, the study overall furthers our understanding of trust mechanisms related to the dynamic and growingly relevant field of human-AI interactions.

Appendix

Table 1 – List of Contrasts

List of all trust game decision period contrasts considered in whole-brain GLM analysis

Mover (M)	Decision	*HUM >AI*	*HUM > CPU*	AI > CPU
1	Trust (Res 2+3)	HATrust	HCTrust	ACTrust
	Mistrust (Res 1)	HAMistrust	HCMistrust	ACMistrust
2	Reciprocate (Res 2)	HAReciprocate	HCReciprocate	ACReciprocate
	Betray (Res 3)	HABetray	HABetray	ACBetray

Note. Due to the coding of the trust game, there were only three outcomes. However, for either result 2 or 3 to occur when the player was M1, they had to make the decision to invest in their opponent, so results 2 and 3 were included in creating the "Trust" contrasts.

Table 2 – Correlations between different psychometric scales

Means, standard deviations, and correlations with confidence intervals

Psychometric	*M*	*SD*	PTT	ATAI_F	ATAI_A
PTT	21.07	4.79			
ATAI_F	14.80	5.49	-.08 [-.57, .45]		
ATAI_A	10.67	3.54	.40 [-.14, .76]	-.36 [-.73, .19]	
H/W	4.20	1.42	.25 [-.30, .68]	-.63* [-.87, -.18]	.38 [-.16, .75]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

Table 3 – Correlations between Psychometric and Demographic

Means, standard deviations, and correlations with confidence intervals

Demographic	*M*	*SD*	PTT	ATAI_F	ATAI_A	H/W
Age	24.53	4.23	-.37 [-.72, .14]	-.07 [-.53, .43]	-.27 [-.66, .25]	.10 [-.40, .55]
Sex	1.47	0.51	.46 [-.02, .77]	-.13 [-.57, .38]	.09 [-.41, .55]	.26 [-.25, .66]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values represent correlation coefficients. Spearman's ρ was used for Age and point-biserial correlations were used for Sex. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014).
* indicates $p < .05$. ** indicates $p < .01$.

Table 4 – Percentage of Times Participants Made a Choice Compared to How Often They Could

Means, standard deviations, and ranges, data is in %

Decision	*N*	*M*	*SD*	*Min*	*Max*	*Range*
AI Betray	15	44.89	30.81	6.67	100.00	93.33
AI Mistrust	15	26.67	16.82	5.56	61.11	55.56
AI Reciprocate	15	55.11	31.91	6.67	100.00	93.33
AI Trust	16	69.79	20.18	38.89	100.00	61.11
CPU Betray	15	71.85	22.56	44.44	100.00	55.56
CPU Mistrust	16	50.69	24.67	11.11	100.00	88.89
CPU Reciprocate	9	43.21	24.50	22.22	100.00	77.78
CPU Trust	15	40.00	18.33	5.56	72.22	66.67
HUM Betray	14	36.67	26.89	6.67	86.67	80.00
HUM Mistrust	13	28.63	18.68	5.56	61.11	55.56
HUM Reciprocate	16	61.25	27.24	13.33	100.00	86.67
HUM Trust	16	70.49	19.85	38.89	100.00	61.11

Note. M and *SD* are used to represent mean and standard deviation, respectively. *Min* and *Max* refer to the highest and lowest percentage of actual decision/maximum possible decision, respectively. *Range* is the variation between *Min* and *Max.*

Table 5 – Summed Number of Times Participants Made a Choice Correlated with Psychometric Scores

Means, standard deviations, and correlations with confidence intervals

Decision	*M*	*SD*	PTT	ATAI_F	ATAI_A	H/W
AI Betray	6.73	4.62	-.21 [-.65, .34]	.36 [-.18, .74]	.06 [-.46, .56]	-.22 [-.66, .33]
AI Mistrust	4.80	3.03	.03 [-.49, .53]	.20 [-.35, .65]	.13 [-.41, .60]	.39 [-.15, .75]
AI Reciprocate	8.27	4.79	.40 [-.14, .76]	-.68** [-.89, -.26]	.26 [-.29, .68]	.77** [.43, .92]
AI Trust	12.56	3.63	.41 [-.11, .75]	-.29 [-.69, .24]	.17 [-.36, .61]	.12 [-.40, .58]
CPU Betray	6.47	2.03	.17 [-.37, .63]	-.05 [-.55, .48]	.28 [-.27, .69]	.19 [-.36, .64]
CPU Mistrust	9.12	4.44	.37 [-.15, .73]	-.42 [-.76, .09]	.33 [-.20, .71]	.08 [-.43, .55]
CPU Reciprocate	3.89	2.20	-.16 [-.74, .57]	-.26 [-.79, .49]	-.35 [-.82, .41]	.87** [.48, .97]
CPU Trust	7.20	3.30	-.21 [-.66, .33]	.18 [-.37, .63]	-.14 [-.61, .40]	.33 [-.22, .72]
HUM Betray	5.50	4.03	-.10 [-.60, .46]	.44 [-.11, .79]	-.35 [-.74, .22]	-.18 [-.65, .39]
HUM Mistrust	5.15	3.36	.19 [-.40, .67]	-.33 [-.74, .27]	-.13 [-.63, .46]	.49 [-.09, .82]
HUM Reciprocate	9.19	4.09	.23 [-.30, .65]	-.58* [-.83, -.11]	.42 [-.09, .76]	.36 [-.16, .73]
HUM Trust	12.69	3.57	-.06 [-.54, .45]	-.03 [-.52, .47]	.31 [-.22, .70]	-.08 [-.55, .43]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

Table 6 – Summed Number of Times Participants Made a Choice Correlated with Demographics

Means, standard deviations, and correlations with confidence intervals

Decision	*M*	*SD*	Age	Sex
AI Betray	6.73	4.62	-.05 [-.55, .47]	.27 [-.29, .68]
AI Mistrust	4.80	3.03	-.07 [-.56, .46]	-.12 [-.59, .42]
AI Reciprocate	8.27	4.79	-.16 [-.62, .38]	.18 [-.37, .63]
AI Trust	12.56	3.63	-.20 [-.63, .33]	.44 [-.07, .77]
CPU Betray	6.47	2.03	-.10 [-.58, .44]	.43 [-.11, .77]
CPU Mistrust	9.12	4.44	-.29 [-.69, .24]	.09 [-.43, .56]
CPU Reciprocate	3.89	2.20	-.11 [-.72, .60]	.04 [-.64, .68]
CPU Trust	7.20	3.30	.24 [-.31, .67]	.11 [-.43, .59]
HUM Betray	5.50	4.03	.61* [.12, .86]	.35 [-.22, .74]
HUM Mistrust	5.15	3.36	.36 [-.24, .76]	-.05 [-.59, .51]
HUM Reciprocate	9.19	4.09	-.58* [-.83, -.12]	-.11 [-.57, .41]
HUM Trust	12.69	3.57	-.41 [-.75, .10]	.23 [-.30, .65]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Spearman's ρ was used for Age and point-biserial correlations were used for Sex. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

Table 7 – Spearman Correlations between Brain Activation in the right Middle Frontal Gyrus and Psychometric Scores

Means, standard deviations, and correlations with confidence intervals

Decision	*M*	*SD*	PTT	ATAI_F	ATAI_A	H/W
HUM Trust	0.11	0.32	.18 [-0.43, .65]	.09 [-0.54, .67]	.24 [-0.34, .69]	-.01 [-0.56, .52]
HUM Mistrust	0.01	0.64	.24 [-0.39, .76]	-.24 [-0.71, .32]	-.02 [-0.56, .58]	.38 [-0.26, .81]
HUM Reciprocate	0.01	0.45	.10 [-0.58, .68]	.00 [-0.53, .57]	.06 [-0.55, .62]	-.44 [-0.82, .08]
HUM Betray	0.17	0.37	-.00 [-0.59, .58]	-.34 [-0.75, .18]	-.16 [-0.58, .32]	.17 [-0.40, .67]
AI Trust	0.15	0.52	.15 [-0.42, .73]	-.36 [-0.83, .35]	.30 [-0.29, .78]	.07 [-0.56, .66]
AI Mistrust	0.25	0.59	.12 [-0.41, .62]	-.46 [-0.83, .09]	.61* [.19, .85]	-.08 [-0.55, .43]
AI Reciprocate	0.60	0.61	-.18 [-0.72, .42]	-.18 [-0.63, .45]	-.33 [-0.80, .33]	-.36 [-0.80, .16]
AI Betray	0.03	0.44	-.10 [-0.62, .53]	.15 [-0.44, .61]	-.28 [-0.70, .38]	-.21 [-0.73, .42]
CPU Trust	0.09	0.42	-.77** [-0.94, -0.42]	-.33 [-0.75, .22]	-.34 [-0.77, .24]	-.04 [-0.57, .56]
CPU Mistrust	0.04	0.59	-.24 [-0.72, .43]	-.30 [-0.76, .31]	-.10 [-0.65, .49]	.09 [-0.51, .57]
CPU Reciprocate	0.11	0.51	.21 [-0.31, .66]	.36 [-0.23, .76]	-.13 [-0.70, .46]	-.71** [-0.88, -0.30]
CPU Betray	0.33	0.45	.05 [-0.62, .61]	-.48 [-0.81, .03]	-.09 [-0.58, .41]	.22 [-0.39, .75]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

Table 8 – Spearman Correlations between Brain Activation in the right Middle Temporal Gyrus and Psychometric Scores

Means, standard deviations, and correlations with confidence intervals

Decision	*M*	*SD*	PTT	ATAI_F	ATAI_A	H/W
HUM Trust	0.24	0.40	-.10 [-0.62, .47]	.30 [-0.24, .72]	-.33 [-0.78, .29]	-.41 [-0.76, .06]
HUM Mistrust	0.30	0.59	.11 [-0.43, .63]	.28 [-0.26, .70]	-.27 [-0.70, .20]	-.23 [-0.61, .25]
HUM Reciprocate	-0.04	0.51	-.30 [-0.67, .32]	-.13 [-0.66, .47]	-.32 [-0.72, .20]	-.01 [-0.55, .56]
HUM Betray	0.45	0.38	-.43 [-0.83, .14]	-.26 [-0.69, .29]	-.50 [-0.87, .05]	-.09 [-0.65, .46]
AI Trust	0.23	0.48	-.26 [-0.68, .42]	.18 [-0.41, .69]	-.26 [-0.79, .36]	-.36 [-0.74, .19]
AI Mistrust	0.13	0.52	-.45 [-0.74, -0.00]	-.06 [-0.58, .48]	-.23 [-0.65, .36]	-.15 [-0.60, .40]
AI Reciprocate	0.04	0.62	-.05 [-0.65, .55]	-.14 [-0.67, .44]	.21 [-0.40, .69]	.12 [-0.50, .65]
AI Betray	0.13	0.51	.07 [-0.42, .57]	-.23 [-0.73, .37]	.21 [-0.37, .70]	.24 [-0.39, .80]
CPU Trust	0.14	0.40	-.37 [-0.79, .22]	-.14 [-0.59, .35]	-.55* [-0.88, .01]	-.03 [-0.55, .53]
CPU Mistrust	0.21	0.40	-.57* [-0.95, .03]	-.08 [-0.55, .45]	-.48 [-0.81, .03]	.05 [-0.46, .48]
CPU Reciprocate	0.26	0.33	-.12 [-0.69, .53]	-.03 [-0.45, .45]	.15 [-0.48, .72]	-.26 [-0.71, .29]
CPU Betray	0.29	0.46	-.14 [-0.74, .45]	-.34 [-0.73, .24]	-.19 [-0.77, .44]	.45 [-0.16, .85]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

Table 9 – Spearman Correlations between Brain Activation in the left Caudate Nucleus and Psychometric Scores

Means, standard deviations, and correlations with confidence intervals

Decision	*M*	*SD*	PTT	ATAI_F	ATAI_A	H/W
HUM Trust	0.02	0.36	.22 [-0.38, .68]	.19 [-0.40, .62]	-.01 [-0.51, .52]	-.12 [-0.57, .38]
HUM Mistrust	-0.02	0.33	.29 [-0.31, .81]	-.05 [-0.64, .49]	-.09 [-0.72, .54]	.08 [-0.43, .57]
HUM Reciprocate	0.21	0.38	-.14 [-0.64, .43]	.12 [-0.40, .62]	-.41 [-0.72, .15]	-.23 [-0.70, .36]
HUM Betray	0.16	0.57	-.25 [-0.81, .37]	.29 [-0.32, .71]	.02 [-0.66, .58]	-.12 [-0.72, .53]
AI Trust	-0.03	0.43	-.40 [-0.85, .17]	.16 [-0.45, .66]	-.61* [-0.90, -0.10]	-.37 [-0.81, .26]
AI Mistrust	-0.01	0.72	.16 [-0.41, .65]	.07 [-0.50, .60]	.19 [-0.41, .68]	.16 [-0.44, .63]
AI Reciprocate	0.30	0.45	.22 [-0.37, .70]	.02 [-0.47, .59]	-.21 [-0.73, .39]	.06 [-0.52, .60]
AI Betray	0.09	0.42	-.16 [-0.54, .32]	-.29 [-0.74, .34]	.37 [-0.20, .74]	.19 [-0.41, .67]
CPU Trust	0.12	0.34	-.20 [-0.64, .35]	.12 [-0.44, .65]	-.70** [-0.92, -0.32]	-.05 [-0.62, .50]
CPU Mistrust	0.09	0.36	.11 [-0.44, .70]	.31 [-0.37, .74]	-.04 [-0.59, .46]	-.18 [-0.70, .44]
CPU Reciprocate	0.05	0.27	-.17 [-0.66, .41]	-.43 [-0.87, .18]	-.33 [-0.77, .24]	.13 [-0.56, .66]
CPU Betray	0.08	0.45	-.06 [-0.54, .50]	-.36 [-0.78, .26]	.02 [-0.55, .56]	.33 [-0.22, .74]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

Table 10 – Spearman Correlations between Brain Activation in the right Caudate Nucleus and Psychometric Scores

Means, standard deviations, and correlations with confidence intervals

Decision	*M*	*SD*	PTT	ATAI_F	ATAI_A	H/W
HUM Trust	-0.03	0.25	.00 [-0.61, .67]	.27 [-0.26, .67]	-.05 [-0.56, .50]	.26 [-0.25, .63]
HUM Mistrust	-0.00	0.45	-.10 [-0.68, .52]	-.35 [-0.74, .19]	-.25 [-0.76, .40]	.35 [-0.24, .74]
HUM Reciprocate	0.18	0.41	-.27 [-0.82, .37]	-.04 [-0.52, .45]	-.16 [-0.64, .39]	-.09 [-0.66, .48]
HUM Betray	-0.18	0.47	-.06 [-0.63, .53]	-.01 [-0.56, .52]	-.10 [-0.81, .52]	-.24 [-0.73, .37]
AI Trust	-0.12	0.19	.07 [-0.49, .64]	.09 [-0.50, .66]	-.52* [-0.89, .01]	-.14 [-0.75, .47]
AI Mistrust	-0.17	0.71	.32 [-0.23, .74]	-.56* [-0.85, -0.04]	.35 [-0.25, .80]	.72** [.40, .90]
AI Reciprocate	0.23	0.70	-.02 [-0.57, .58]	.08 [-0.44, .58]	-.51 [-0.86, .06]	.10 [-0.51, .73]
AI Betray	0.14	0.49	-.33 [-0.69, .16]	-.07 [-0.66, .57]	.23 [-0.34, .72]	-.02 [-0.63, .56]
CPU Trust	0.07	0.29	-.30 [-0.81, .30]	-.16 [-0.64, .37]	-.12 [-0.64, .39]	.14 [-0.44, .69]
CPU Mistrust	-0.06	0.38	-.19 [-0.73, .37]	-.10 [-0.56, .57]	.23 [-0.40, .74]	-.17 [-0.63, .36]
CPU Reciprocate	0.03	0.32	.07 [-0.52, .61]	-.58* [-0.86, -0.10]	.24 [-0.43, .73]	.21 [-0.34, .74]
CPU Betray	-0.08	0.43	.04 [-0.58, .64]	-.35 [-0.83, .28]	.21 [-0.38, .71]	.63* [.18, .88]

Note. M and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

Table 11 – Spearman Correlations between Brain Activation in the left Cerebellum Lobules VIIb and VIII and Psychometric Scores

Means, standard deviations, and correlations with confidence intervals

Variable	*M*	*SD*	PTT	ATAI_F	ATAI_A	H/W
HUM Trust	-0.07	0.31	.15 [-0.45, .79]	.26 [-0.32, .68]	-.24 [-0.72, .35]	-.25 [-0.65, .27]
HUM Mistrust	0.29	0.89	.03 [-0.56, .68]	.03 [-0.49, .56]	-.19 [-0.67, .34]	.08 [-0.41, .55]
HUM Reciprocate	0.13	0.32	.10 [-0.47, .60]	.25 [-0.36, .74]	.27 [-0.35, .73]	-.14 [-0.67, .42]
HUM Betray	0.07	0.56	.47 [-0.16, .82]	-.36 [-0.81, .18]	-.22 [-0.77, .38]	-.01 [-0.50, .49]
AI Trust	0.10	0.45	-.12 [-0.67, .44]	-.09 [-0.60, .40]	-.15 [-0.67, .46]	.00 [-0.55, .53]
AI Mistrust	0.01	0.35	-.48 [-0.75, -0.07]	.02 [-0.55, .63]	.25 [-0.33, .70]	-.20 [-0.75, .44]
AI Reciprocate	0.18	0.38	-.33 [-0.78, .26]	.27 [-0.26, .72]	.11 [-0.52, .63]	-.34 [-0.81, .24]
AI Betray	0.13	0.40	-.33 [-0.77, .30]	.17 [-0.44, .78]	-.46 [-0.83, .03]	.19 [-0.40, .68]
CPU Trust	0.20	0.44	-.13 [-0.55, .34]	-.09 [-0.76, .53]	-.72** [-0.92, -0.33]	.02 [-0.59, .59]
CPU Mistrust	0.14	0.26	-.02 [-0.52, .53]	.43 [-0.09, .76]	-.57* [-0.82, -0.10]	-.42 [-0.84, .17]
CPU Reciprocate	0.16	0.43	-.63* [-0.90, -0.16]	-.06 [-0.65, .42]	-.06 [-0.62, .50]	-.18 [-0.63, .36]
CPU Betray	0.09	0.39	-.29 [-0.69, .26]	.02 [-0.54, .65]	-.17 [-0.65, .34]	-.43 [-0.85, .12]

Note. *M* and *SD* are used to represent mean and standard deviation, respectively. Values in square brackets indicate the 95% confidence interval for each correlation. The confidence interval is a plausible range of population correlations that could have caused the sample correlation (Cumming, 2014). * indicates $p < .05$. ** indicates $p < .01$.

References

Banerjee, S., Galizzi, M. M., & Hortala-Vallve, R. (2021). Trusting the Trust Game: An External Validity Analysis with a UK Representative Sample. *Games*, *12*(3), 66. https://doi.org/10.3390/g12030066

Centorrino, S., Djemai, E., Hopfensitz, A., Milinski, M., & Seabright, P. (2015). Honest signaling in trust interactions: Smiles rated as genuine induce trust and signal higher earning opportunities. *Evolution and Human Behavior*, *36*(1), 8–16. https://doi.org/10.1016/j.evolhumbehav.2014.08.001

Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutscher, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience*, *6*. https://doi.org/10.3389/fnhum.2012.00103

Chatterji, A., Cunningham, T., Deming, D. J., Hitzig, Z., Ong, C., Shan, C. Y., & Wadman, K. (2025). How People Use ChatGPT. *National Bureau of Economic Research*. https://doi.org/10.3386/w34255.

Cheetham, M., Suter, P., & Jäncke, L. (2011). The Human Likeness Dimension of the "Uncanny Valley Hypothesis": Behavioral and Functional MRI Findings. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00126

Chen, Y., He, H., Lin, W., Yang, J., Tan, S., Tao, W., Guan, Q., & Krueger, F. (2023). The connectome-based prediction of trust propensity in older adults: A resting-state functional magnetic resonance imaging study. *Human Brain Mapping*, *44*(11), 4337–4351. https://doi.org/10.1002/hbm.26385

Chin, C. M., & Yusoff, N. (2023). The Difference in Brain Processing Between Animate and Inanimate Beings in Moral Violation: An Electrophysiological Evidence. *IIUM Medical Journal Malaysia*, *22*(3). https://doi.org/10.31436/imjm.v22i3.2261

Damiani, D., Nascimento, A. M., & Pereira, L. K. (2020). Cortical Brain Functions – The Brodmann Legacy in the 21st Century. *Arquivos Brasileiros de Neurocirurgia: Brazilian Neurosurgery*, *39*(4), 261–270. https://doi.org/10.1055/s-0036-1597573

Dang, J., & Liu, L. (2021). Robots are friends as well as foes: Ambivalent attitudes toward mindful and mindless AI robots in the United States and China. *Computers in Human Behavior*, *115*, 106612. https://doi.org/10.1016/j.chb.2020.106612

de Quervain, D. J.-F., Fischbacher, U., Treyer, V., Schellhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The Neural Basis of Altruistic Punishment. *Science*, *305*(5688), 1254–1258. https://doi.org/10.1126/science.1100735

Desmet, C., Deschrijver, E., & Brass, M. (2014). How social is error observation? The neural mechanisms underlying the observation of human and machine errors. *Social Cognitive and Affective Neuroscience*, *9*(4), 427–435. https://doi.org/10.1093/scan/nst002

Di Cesare, G., Fasano, F., Errante, A., Marchi, M., & Rizzolatti, G. (2016). Understanding the internal states of others by listening to action verbs. *Neuropsychologia*, *89*, 172–179. https://doi.org/10.1016/j.neuropsychologia.2016.06.017

Driscoll, M. E., Bollu, P. C., & Tadi, P. (2025). *Neuroanatomy, Nucleus Caudat*. https://www.ncbi.nlm.nih.gov/books/NBK557407

Fairley, K., Vyrastekova, J., Weitzel, U., & Sanfey, A. G. (2019). Subjective Beliefs About Trust and Reciprocity Activate an Expected Reward Signal in the Ventral Striatum. *Frontiers in Neuroscience*, *13*. https://doi.org/10.3389/fnins.2019.00660

Feng, C., Zhu, Z., Cui, Z., Ushakov, V., Dreher, J.-C., Luo, W., Gu, R., Wu, X., & Krueger, F. (2021). Prediction of trust propensity from intrinsic brain morphology and functional connectome. *Human Brain Mapping*, *42*(1), 175–191. https://doi.org/10.1002/hbm.25215

Fouragnan, E., Chierchia, G., Greiner, S., Neveu, R., Avesani, P., & Coricelli, G. (2013). Reputational Priors Magnify Striatal Responses to Violations of Trust. *Journal of Neuroscience*, *33*(8), 3602–3611. https://doi.org/10.1523/JNEUROSCI.3086-12.2013

Gao, Z., Liu, X., Zhang, D., Liu, M., & Hao, N. (2020). The indispensable role of the cerebellum in visual divergent thinking. *Scientific Reports*, *10*(1), 16552. https://doi.org/10.1038/s41598-020-73679-9

Gifuni, A. J., Kendal, A., & Jollant, F. (2017). Neural mapping of guilt: A quantitative meta-analysis of functional imaging studies. *Brain Imaging and Behavior*, *11*(4), 1164–1178. https://doi.org/10.1007/s11682-016-9606-6

Gil de Zúñiga, H., Goyanes, M., & Durotoye, T. (2024). A Scholarly Definition of Artificial Intelligence (AI): Advancing AI as a Conceptual Framework in Communication Research. *Political Communication*, *41*(2), 317–334. https://doi.org/10.1080/10584609.2023.2290497

Gobbini, M. I., Gentili, C., Ricciardi, E., Bellucci, C., Salvini, P., Laschi, C., Guazzelli, M., & Pietrini, P. (2011). Distinct Neural Systems Involved in Agency and Animacy Detection. *Journal of Cognitive Neuroscience*, *23*(8), 1911–1920. https://doi.org/10.1162/jocn.2010.21574

Golgeci, I., Ritala, P., Arslan, A., McKenna, B., & Ali, I. (2025). Confronting and alleviating AI resistance in the workplace: An integrative review and a process framework. *Human Resource Management Review*, *35*(2), 101075. https://doi.org/10.1016/j.hrmr.2024.101075

Goodyear, K., Parasuraman, R., Chernyak, S., Madhavan, P., Deshpande, G., & Krueger, F. (2016). Advice Taking from Humans and Machines: An fMRI and Effective Connectivity Study. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00542

Graff-Radford, J., Williams, L., Jones, D. T., & Benarroch, E. E. (2017). Caudate nucleus as a component of networks controlling behavior. *Neurology*, *89*(21), 2192–2197. https://doi.org/10.1212/WNL.0000000000004680

Grassini, S. (2023). Development and validation of the AI attitude scale (AIAS-4): A brief measure of general attitude toward artificial intelligence. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1191628

Green, P., & Wei-Haas, L. (1985). The Rapid Development of User Interfaces: Experience with the Wizard of OZ Method. *Proceedings of the Human Factors Society Annual Meeting*, *29*(5), 470–474. https://doi.org/10.1177/154193128502900515

Haas, B. W., Ishak, A., Anderson, I. W., & Filkowski, M. M. (2015). The tendency to trust is reflected in human brain structure. *NeuroImage*, *107*, 175–181. https://doi.org/10.1016/j.neuroimage.2014.11.060

High-Level Expert Group on Artificial Intelligence. (2019). *A definition of Artificial Intelligence: Main capabilities and scientific disciplines*. European Commission. https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

Hitsuwari, J., & Takano, R. (2025). Associating attitudes towards AI and ambiguity: The distinction of acceptance and fear of AI. *Acta Psychologica*, *260*, 105581. https://doi.org/10.1016/j.actpsy.2025.105581

Hogenhuis, A., & Hortensius, R. (2022). Domain-specific and domain-general neural network engagement during human–robot interactions. *European Journal of Neuroscience*, *56*(10), 5902–5916. https://doi.org/10.1111/ejn.15823

Ismatullaev, U. V. U., & Kim, S.-H. (2024). Review of the Factors Affecting Acceptance of AI-Infused Systems. *Human Factors*, *66*(1), 126–144. https://doi.org/10.1177/00187208211064707

Jimsheleishvili, S., & Dididze, M. (2026). Neuroanatomy, Cerebellum. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK538167/

Kätsyri, J., de Gelder, B., & de Borst, A. W. (2020). Amygdala responds to direct gaze in real but not in computer-generated faces. *NeuroImage*, *204*, 116216. https://doi.org/10.1016/j.neuroimage.2019.116216

Koyama, M. S., O'Connor, D., Shehzad, Z., & Milham, M. P. (2017). Differential contributions of the middle frontal gyrus functional connectivity to literacy and numeracy. *Scientific Reports*, *7*(1), 17548. https://doi.org/10.1038/s41598-017-17702-6

Krach, S., Blümel, I., Marjoram, D., Lataster, T., Krabbendam, L., Weber, J., van Os, J., & Kircher, T. (2009). Are women better mindreaders? Sex differences in neural correlates of mentalizing detected with functional MRI. *BMC Neuroscience*, *10*(1), 9. https://doi.org/10.1186/1471-2202-10-9

Krueger, F., & Meyer-Lindenberg, A. (2019). Toward a Model of Interpersonal Trust Drawn from Neuroscience, Psychology, and Economics. *Trends in Neurosciences*, *42*(2), 92–101. https://doi.org/10.1016/j.tins.2018.10.004

Kugler, J., & Zak, P. J. (2017). Trust, Cooperation, and Conflict: Neuropolitics and International Relations. In S. A. Yetiv & P. James (Eds), *Advancing Interdisciplinary Approaches to International Relations* (pp. 83–114). Springer International Publishing. https://doi.org/10.1007/978-3-319-40823-1_4

Lai, C., Altavilla, D., Ronconi, A., & Aceto, P. (2016). Fear of missing out (FOMO) is associated with activation of the right middle temporal gyrus during inclusion social cue. *Computers in Human Behavior*, *61*, 516–521. https://doi.org/10.1016/j.chb.2016.03.072

Laufer, I., Negishi, M., Lacadie, C. M., Papademetris, X., & Constable, R. T. (2011). Dissociation between the Activity of the Right Middle Frontal Gyrus and the Middle Temporal Gyrus in Processing Semantic Priming. *PLOS ONE*, *6*(8), e22368. https://doi.org/10.1371/journal.pone.0022368

Li, T., & Fung, H. H. (2013). Age Differences in Trust: An Investigation Across 38 Countries. *The Journals of Gerontology: Series B*, *68*(3), 347–355. https://doi.org/10.1093/geronb/gbs072

Manto, M., Adamaszek, M., Apps, R., Carlson, E., Guarque-Chabrera, J., Heleven, E., Kakei, S., Khodakhah, K., Kuo, S.-H., Lin, C.-Y. R., Joshua, M., Miquel, M., Mitoma, H., Larry, N.,

Péron, J. A., Pickford, J., Schutter, D. J. L. G., Singh, M. K., Tan, T., … Yamashiro, K. (2024). Consensus Paper: Cerebellum and Reward. *The Cerebellum*, *23*(5), 2169–2192. https://doi.org/10.1007/s12311-024-01702-0

McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences*, *98*(20), 11832–11835. https://doi.org/10.1073/pnas.211415698

McKinsey & Company. (2025). *The State of AI: Global Survey 2025 | McKinsey*. https://www.mckinsey.com/capabilities/quantumblack/our-insights/the-state-of-ai

Mirfardi. (2011). Trust and its Relationship to Demographic Factors. *Journal of Social Sciences*, *7*(2), 168–174. https://doi.org/10.3844/jssp.2011.168.174

Montag, C., & Ali, R. (2025). Can We Assess Attitudes Toward AI with Single Items? Associations with Existing Attitudes Toward AI Measures and Trust in ChatGPT. *Journal of Technology in Behavioral Science*. https://doi.org/10.1007/s41347-025-00481-7

Montag, C., Kraus, J., Baumann, M., & Rozgonjuk, D. (2023). The propensity to trust in (automated) technology mediates the links between technology self-efficacy and fear and acceptance of artificial intelligence. *Computers in Human Behavior Reports*, *11*, 100315. https://doi.org/10.1016/j.chbr.2023.100315

Nettekoven, C., Zhi, D., Shahshahani, L., Pinho, A. L., Saadon-Grosman, N., Buckner, R. L., & Diedrichsen, J. (2024). A hierarchical atlas of the human cerebellum for functional precision mapping. *Nature Communications*, *15*, 8376. https://doi.org/10.1038/s41467-024-52371-w

Pu, M., Heleven, E., Delplanque, J., Gibert, N., Ma, Q., Funghi, G., & Van Overwalle, F. (2020). The posterior cerebellum supports the explicit sequence learning linked to trait attribution. *Cognitive, Affective, & Behavioral Neuroscience*, *20*(4), 798–815. https://doi.org/10.3758/s13415-020-00803-7

Rahman, M. M., Babiker, A., & Ali, R. (2025). Motivation, Concerns, and Attitudes Towards AI: Differences by Gender, Age, and Culture. In M. Barhamgi, H. Wang, & X. Wang (Eds), *Web Information Systems Engineering – WISE 2024* (pp. 375–391). Springer Nature. https://doi.org/10.1007/978-981-96-0573-6_28

Rilling, J. K., Gutman, D. A., Zeh, T. R., Pagnoni, G., Berns, G. S., & Kilts, C. D. (2002). A Neural Basis for Social Cooperation. *Neuron*, *35*(2), 395–405. https://doi.org/10.1016/S0896-6273(02)00755-9

Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *NeuroImage*, *22*(4), 1694–1703. https://doi.org/10.1016/j.neuroimage.2004.04.015

Rohrer, J. M. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science*, *1*(1), 27–42. https://doi.org/10.1177/2515245917745629

Rothstein, B., & Uslaner, E. M. (2005). All for All: Equality, Corruption, and Social Trust. *World Politics*, *58*(1), 41–72. https://doi.org/10.1353/wp.2006.0022

Samoili, S., Lopez, C. M., Gomez, G. E., De, P. G., Martinez-Plumed, F., & Delipetrev, B. (2020). *AI WATCH. Defining Artificial Intelligence*. JRC Publications Repository. https://doi.org/10.2760/382730

Schepman, A., & Rodway, P. (2023). The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust. *International Journal of Human–Computer Interaction*, *39*(13), 2724–2741. https://doi.org/10.1080/10447318.2022.2085400

Schmahmann, J. D. (2019). The cerebellum and cognition. *Neuroscience Letters, The Cerebellum in Health and Disease*, *688*, 62–75. https://doi.org/10.1016/j.neulet.2018.07.005

Sindermann, C., Sha, P., Zhou, M., Wernicke, J., Schmitt, H. S., Li, M., Sariyska, R., Stavrou, M., Becker, B., & Montag, C. (2021). Assessing the Attitude Towards Artificial Intelligence: Introduction of a Short Measure in German, Chinese, and English Language. *KI - Künstliche Intelligenz*, *35*(1), 109–118. https://doi.org/10.1007/s13218-020-00689-0

Singer, T., Kiebel, S. J., Winston, J. S., Dolan, R. J., & Frith, C. D. (2004). Brain Responses to the Acquired Moral Status of Faces. *Neuron*, *41*(4), 653–662. https://doi.org/10.1016/S0896-6273(04)00014-5

Sokolov, A. A., Miall, R. C., & Ivry, R. B. (2017). The Cerebellum: Adaptive Prediction for Movement and Cognition. *Trends in Cognitive Sciences*, *21*(5), 313–332. https://doi.org/10.1016/j.tics.2017.02.005

Stanford HAI. (2025). *Stanford AI Experts Predict What Will Happen in 2026 | Stanford HAI*. https://hai.stanford.edu/news/stanford-ai-experts-predict-what-will-happen-in-2026

Sweijen, S. W., van de Groep, S., te Brinke, L. W., Fuligni, A. J., & Crone, E. A. (2023). Neural Mechanisms Underlying Trust to Friends, Community Members, and Unknown Peers in Adolescence. *Journal of Cognitive Neuroscience*, *35*(12), 1936–1959. https://doi.org/10.1162/jocn_a_02055

Turner, B. O., Paul, E. J., Miller, M. B., & Barbey, A. K. (2018). Small sample sizes reduce the replicability of task-based fMRI studies. *Communications Biology*, *1*, 62. https://doi.org/10.1038/s42003-018-0073-z

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., & Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, *15*(1), 273–289. https://doi.org/10.1006/nimg.2001.0978

Valeriani, D., O'Flynn, L. C., Worthley, A., Sichani, A. H., & Simonyan, K. (2022). Multimodal collaborative brain-computer interfaces aid human-machine team decision-making in a pandemic scenario. *Journal of Neural Engineering*, *19*(5), 056036. https://doi.org/10.1088/1741-2552/ac96a5

van Kemenade, B. M., Arikan, B. E., Podranski, K., Steinsträter, O., Kircher, T., & Straube, B. (2019). Distinct Roles for the Cerebellum, Angular Gyrus, and Middle Temporal Gyrus in Action–Feedback Monitoring. *Cerebral Cortex*, *29*(4), 1520–1531. https://doi.org/10.1093/cercor/bhy048

Ventura, A., Starke, C., Righetti, F., & Köbis, N. (2025). *Relationships in the Age of AI: A Review on the Opportunities and Risks of Synthetic Relationships to Reduce Loneliness*. PsyArXiv. https://doi.org/10.31234/osf.io/w7nmz_v1

Wardle, M. C., Fitzgerald, D. A., Angstadt, M., Sripada, C. S., McCabe, K., & Phan, K. L. (2013). The Caudate Signals Bad Reputation during Trust Decisions. *PLOS ONE*, *8*(6), e68884. https://doi.org/10.1371/journal.pone.0068884

Wright, M., Skaggs, W., & Nielsen, F. A. (2020). The cerebellum. *WikiJournal of Medicine*, *3*(1), 1–15. https://doi.org/10.3316/informit.380776158668038

Wu, Y., Hall, A. S. M., Siehl, S., Grafman, J., & Krueger, F. (2020). Neural Signatures of Gender Differences in Interpersonal Trust. *Frontiers in Human Neuroscience*, *14*. https://doi.org/10.3389/fnhum.2020.00225

Zhang, H., Jing, Y., & Gu, R. (2026). Who Gets More Trust—AI or Humans, and Why? A Cross-Cultural Analysis of AI and Interpersonal Trust. *Behavioral Sciences*, *16*(3), 320. https://doi.org/10.3390/bs16030320

Zou, X., Li, D., Turel, O., & He, Q. (2025). Neural mechanisms of cooperation and fairness in iterative prisoner's dilemma. *Behavioural Brain Research*, *476*, 115272. https://doi.org/10.1016/j.bbr.2024.115272