ANNUAL REVIEWS

Annual Review of Psychology

The Neuroscience of Human and Artificial Intelligence Presence

Lasana T. Harris[1,2]

[1]Department of Experimental Psychology, University College London, London, United Kingdom; email: lasana.harris@ucl.ac.uk

[2]Alan Turing Institute, London, United Kingdom

Annu. Rev. Psychol. 2024. 75:433–66

First published as a Review in Advance on October 31, 2023

The *Annual Review of Psychology* is online at psych.annualreviews.org

https://doi.org/10.1146/annurev-psych-013123-123421

Keywords

social cognition, social presence, artificial intelligence, AI, social perception, social decision making, social context

Abstract

Two decades of social neuroscience and neuroeconomics research illustrate the brain mechanisms that are engaged when people consider human beings, often in comparison to considering artificial intelligence (AI) as a nonhuman control. AI as an experimental control preserves agency and facilitates social interactions but lacks a human presence, providing insight into brain mechanisms that are engaged by human presence and the presence of AI. Here, I review this literature to determine how the brain instantiates human and AI presence across social perception and decision-making paradigms commonly used to realize a social context. People behave toward humans differently than they do toward AI. Moreover, brain regions more engaged by humans compared to AI extend beyond the social cognition brain network to all parts of the brain, and the brain sometimes is engaged more by AI than by humans. Finally, I discuss gaps in the literature, limitations in current neuroscience approaches, and how an understanding of the brain correlates of human and AI presence can inform social science in the wild.

Contents

INTRODUCTION

In the summer of 2022, Google fired software engineer Blake Lemoine for violating employment and data security policies. Lemoine claimed that the artificial intelligence (AI) called Language Model for Dialogue Applications (LaMDA) being developed by Google had gained sentience or consciousness. As evidence, Lemoine cited, from among thousands of messages he had exchanged with the AI, responses to questions like “What do you fear the most?” that seemed to indicate the software was self-aware. As Lemoine famously stated, “I know a person when I talk to it.” Timnit Gebru and Margaret Mitchell, researchers at Google in data ethics, had already warned of the ability of large language models (LLMs) like LaMDA to simulate human consciousness to the point where people could not dissociate them from humans (Bender et al. 2021). Both researchers were fired from Google two years before Lemoine’s claims became public.

This case makes salient human presence and how the brain might instantiate social context. The anthropomorphism described above and Lemoine’s own words suggest he experienced a human presence. A human presence can be experienced in the actual, imagined, or implied presence of another human being. Anthropomorphism captures the imagined aspect of this psychological experience, since people objectively realize that AI is not human. However, this realization does not preclude social interaction with AI. The actual and implied presence of a conspecific (for humans, these are other human beings) is the environment in which the brains of most animals, including humans, evolved. The physical presence or not of a conspecific signaled opportunities for cooperation or competition, as well as potential threat and safety, so being able to detect the presence of others was necessary for survival and reproduction.

Anthropomorphism: attribution of mental states or a mind to a nonhuman entity

Conspecific: a target that is a member of the same species as the perceiver

Most social interactions require people to realize that they are in the presence of another human being. Human presence instantiates a social context, and a successful social interaction requires behavior appropriate to that context. Social rules and norms determine what behavior is appropriate. Such situationalism shapes social psychological theorizing and experimentation, yet we lack a unified understanding of what exactly a social context is. This review attempts to understand the social context by considering human presence and contrasting it with a nonhuman presence (AI) by comparing brain activity when people engage with humans and brain activity during interactions with AIs.

Imagined Human Presence

The imagined presence of conspecifics is perhaps unique to human beings and reflects the flexible nature of human social cognition, distinguishing it from that of other species (Harris 2017; O. Deroy & L.T. Harris, unpublished manuscript). This imaginative process also allows for flexibility in what is considered human irrespective of ontology—that is, regardless of whether biological human beings suffer a dehumanized perception or nonhuman entities trigger anthropomorphism. Human beings are distinct from nonhumans because they have a mind (Fiske & Taylor 2013). Having a mind entails a conscious experience of thought and emotion, which is private and known only to the person experiencing their mind (Wittgenstein 1993). Therefore, human beings never directly perceive other minds; they infer them. In essence, acknowledging that another being in the world has a mind is an imaginative process. For instance, people can imagine themselves as an abstract concept separate from their phenomenological experience—respectively, the "I" and the "me" (James 1890)—suggesting that the self can be perceived as a social target and even dehumanized (Tang & Harris 2015).

Given this special flexibility of human social cognition, it is important to understand where the boundaries lie for what people consider human. Controlling for the imagination component of social cognition would reveal the mechanisms involved in the actual and implied presence of biological human beings, a sort of humanity detector. This begs the question, What psychological and brain processes are necessary and sufficient for something to be considered human? Brain imaging may shed some light on this issue.

Dualism and Embodiment

An alternate way of considering what makes something human relies on Descartes's (1637) mind-body problem. Descartes theorized that the mind and the body were separate entities, and indeed, human beings inherently think of the mind and the body as separate entities. The brain processes human beings' physical forms using feature space mapping in the ventral temporal lobe (Haxby et al. 2001)—the same kind of processing that nonhuman objects trigger—allowing the perceiver to identify the target. In addition, the brain calculates a mental state inference—a Bayesian integration of statistical information—to infer the contents of a target's mind (Amodio & Frith 2006, Frith & Frith 2001, Gallagher & Frith 2003). Minds cannot be directly perceived but are only inferred from behavior. Therefore, this second process is typically reserved for human beings.

However, anthropomorphism suggests that mental state inferences are not reserved for human targets. Anthropomorphism of disembodied (without a physical form; e.g., algorithms) AI, therefore, does not trigger feature space mapping for visual perception but only the mental state inference processes. As a consequence, differences in brain activation when engaging with humans and with disembodied AIs reflect activation unique to human beings beyond mental state inferences, not feature space mapping. Anthropomorphism of embodied AIs (e.g., robots, avatars, androids) triggers feature space mapping because robots, avatars, and so on present a physical form than can be processed. Therefore, differences in brain activation when engaging with humans and with embodied AIs reflect activation unique to human beings beyond both feature space mapping and mental state inferences.

Situationalism: the view that the situation or social context determines people's thoughts, emotions, and behavior, regardless of person-specific factors

Dehumanized perception: a failure to consider another human being's mind, indexed by reduced mentalizing or mental state inferences

Algorithm: a computer program written for a specific purpose, such as calculating, predicting, learning, and recommending based on data

Avatar: a three-dimensional computer-generated representation of a person or other being that is controlled by either a human or an algorithm

Android: a robot that closely resembles a human in appearance

SOCIAL COGNITION AND CONTEXT

Animacy: movement that suggests the target is alive

Agent: an entity that initiates its own behavior and therefore must be guided by internal processes

Definitional confusion abounds, however, when researchers discuss the social cognition of human beings (see the sidebar titled Clarifying Jargon). The literature conflates related social cognitive phenomena, including agency, anthropomorphism, animacy, and person perception, considering them all as social processes. Such an all-encompassing definition blurs the distinction between perceiving biological human beings and perceiving nonhuman social targets, and it emphasizes the social inference component of person perception, ignoring the physical perception dimension. Moreover, it obscures more interesting research questions around human presence and the triggers for related processes, which are the focus of this review. Here I define social presence as distinct from human presence, since the former includes anthropomorphism of nonhuman entities such as AIs. I adopt the term "human presence" from the social presence jargon used in the behavioral neuroscience literature that describes the behavioral and physiological reactions of nonhuman animals to conspecifics. This comparison of the presence and the absence of a human being constrains the literature reviewed and is consistent with the evolutionary context in which the brain evolved.

Flexible social cognition and the sheer enormous variety of humanity make defining a social context difficult. Social contexts therefore poorly generalize across experiments because participants may interpret the social context differently in different cultures (Markus & Kitayama 1991), and even differently on days with different weather (Schwarz & Clore 1983). The all-encompassing nature of the social context is therefore difficult to fully understand, and social and behavioral sciences have done an enormous amount of descriptive work regarding how different social contexts impact specific types of people's behavior.

A definition of the social context must involve agents that are not human, such as well-loved pets, anthropomorphized avatars, social robots, and interactive disembodied AIs. However, the fact that these are primarily modern inventions—except for domesticated animals, who share fantastic social cognition abilities with humans (Hare & Tomasello 2005)—suggests that the social context was primarily instantiated in the brain by the presence of other human beings.

AIs as Nonhuman Controls

There is a substantial body of brain imaging research that considers AI as a nonhuman control. AIs in these experiments are agents that can engage in social interaction but are not human.

CLARIFYING JARGON

The terms "person," "social," "agent," and "human" have been used interchangeably throughout the literature, but clarifying their difference is important for understanding the argument guiding this review. I suggest the following clarification. Human beings consider other human beings as people, so "person" captures this folk understanding and is often used to consider the point of view of participants in the experiment (e.g., participants encountered another person when they entered the lab). As such, person perception is the study of how participants perceive other people, and it represents a thriving subfield of social psychology. "Social" is reserved for the context to allow it to be applied to ontologies broader than just human (e.g., chimpanzee society), and therefore it describes a set of behaviors in which others are considered or involved. "Agent" captures any entity that is self-propelled and behaves nonrandomly, suggesting that such entity is goal directed and has a minimal level of mind. "Human" captures the perceptual process of categorizing a conspecific, and it is at best a stereotype or category description subject to the same subjective and cultural laws of all social categories.

Neuroscientists have used AIs as baselines for comparing responses to human behavior and brain activities in studies of social perception and decision making, revealing insight into human psychology. AIs do not preclude social processing since they can be considered agents with intentions and other mental capacities. However, being considered human involves engaging an abstract concept or stereotype of what it means to be human that can be applied to dissociate humans from nonhumans. Detecting how the brain differentiates biological humans from nonhumans provides insight into the evolved mechanisms of human presence, elucidating triggers of the social context.

Stereotype: a set of traits and behaviors associated with a social target based on cultural narratives about their perceived social group

Differences between interacting with human and with nonhuman agents substantiate the claim that people define the social context. First, moral rules and legal canon consider primarily people as perpetrators and victims, and they are relevant primarily for people. As such, moral reasoning is reserved for people and is not usually considered during exchanges with nonhuman agents (O. Deroy & L.T. Harris, unpublished manuscript). Second, people hold reputational concerns when interacting with others, making impression management relevant (Goffman 1959). People are less inclined to be concerned about the impression a nonhuman agent may form of them. Finally, there is an opportunity for biological and behavioral resonance between human interaction partners but not between nonhuman interaction partners. Such synchronicity facilitates social interaction (McNeill 1997). Here I review the social neuroscience and neuroeconomic literatures to discover the behavioral and brain activity patterns that differentiate between humans and AI during social perception and decision-making paradigms.

BEHAVIORAL NEUROSCIENCE AND SOCIAL PRESENCE

I first consider the social presence literature in nonhuman animals. Importantly, this literature only considers three social aspects of processing conspecifics: social presence, social identity, and social impact. A fourth aspect, social support, is typically considered along with social stressors and studied in the context of anxiety. A final aspect, social content, lies strictly in the domain of language and is not studied in behavioral neuroscience. No such dedicated study of social presence exists in the human literature. Below I describe brain research in the rodent and nonhuman primate literatures that is relevant to social presence. I include terms and concepts like social recognition and social exploration but ignore studies of social facilitation (e.g., Lipina & Roder 2013), since this literature explores imitation. I also ignore the passive and active presence of others in single or joint tasks; though relevant to social cognition, such effects are akin to imitation, mimicry, synchronicity, and other social coordination phenomena beyond the scope of this review.

Researchers have long established that rodents can discriminate between the presence and absence of a conspecific (Husted & McKenna 1966, Thor & Holloway 1982). Most of this work has focused on social identity—that is, how rodents dissociate conspecifics from each other (Petrulis 2009). However, some research has explored social presence—that is, the presence versus absence of a conspecific—and this research implicates large elevations in firing rates in excitatory CA1 cells in the ventral hippocampus (Rao et al. 2019). Other research investigating social recognition—that is, responding to a familiar conspecific—demonstrates that agonists that block vasopressin expression in the lateral septum of the hippocampus affect social but not object recognition (Everts & Koolhaas 1997). Similarly, oxytocin agonists applied to the medial amygdala and lateral septum block social memory formation but not object memory formation (Lukas et al. 2013).

Knockout mice demonstrating depleted D-serine levels, an agonist for N-methyl-D-aspartate (NMDA) receptors, show reduced social exploration (Matveeva et al. 2019), but vagal deafferentation does not affect social exploration in rats (Klarer et al. 2019), suggesting the gut–brain axis may not be necessary for identifying conspecifics. Manipulating the excitatory/inhibitory balance in neurons in the forebrain of knockout mice also affects this balance in their amygdala, and such female mice prefer stranger chambers relative to empty chambers, whereas wild-type mice show no such preference (Powers et al. 2021). The research summarized above varies in the operationalization of social presence, incorporating recognition and exploration, but it broadly implicates the hippocampus and amygdala in responses to the presence of a conspecific.

Studies comparing conspecific faces to objects in chimpanzees using positron emission tomography (PET) show preferential processing for faces in the ventral temporal cortex, including the fusiform face gyrus (FFA) (Parr et al. 2009), akin to brain activity in humans for the same task (Kanwisher et al. 1997, McCarthy et al. 1997). Macaques also show evidence of face-selective neurons in the temporal lobe; functional magnetic resonance imaging (fMRI) data (Pinsk et al. 2005) converge with single-cell recordings to demonstrate face selectivity in the anterior and posterior superior temporal sulcus (STS) (Baylis et al. 1987, Bruce et al. 1981, Desimone et al. 1984, Gross et al. 1972, Perrett et al. 1982, Tanaka et al. 1991, Yamane et al. 1988). However, researchers have not explored brain responses of nonhuman primates to social presence beyond face perception. Instead, existing research documents social identity (e.g., Basile et al. 2009) and social impact responses (e.g., Keupp et al. 2019) (for a review of both social identity and social impact in rhesus macaques, see Monfardini et al. 2017).

HUMAN PRESENCE BRAIN NETWORK

Over the past two decades, a few different brain networks have been reliably mapped to psychological constructions (see Koban et al. 2021 for a review). While this does not completely solve the reverse inference problem that plagues brain imaging research (Poldrack 2011), it does provide potential hypotheses regarding the psychological processes involved during certain behaviors. Below I describe a few of these major networks relevant to human presence, with the aim to use them as guides when discussing the brain imaging results comparing responses to humans and to AIs (see **Figure 1*a***).

"Social" Brain Regions

Previous research has already taken similar approaches to exploring social versus nonsocial cognition, finding that public repositories of brain imaging research like Neurosynth host over 1,000 studies comparing social to nonsocial stimuli (Tso et al. 2018). Consistently, the social cognition brain network is more engaged by human than nonhuman stimuli, including the dorsal medial prefrontal cortex (DMPFC) and ventral medial prefrontal cortex (VMPFC), posterior cingulate (PCC), fusiform gyrus (FG), and temporal pole (TP), along with other brain regions occasionally engaged in social cognition tasks, including the amygdala, occipital-temporal junction (OCT), and inferior frontal gyrus (IFG) (see **Figure 1*a***).

Tso et al. (2018) and many others (see Amodio & Frith 2006, Mars et al. 2012, Van Overwalle 2009) come to the same conclusion: The mere presence of social information is enough to engage the social cognition brain network. However, the definition of social is up to the individual researcher, and the results conflate different definitions, including social presence, social identity, social impact (behavioral facilitation/inhibition), social support, and social content. Though it is informative to know that at the aggregate level the social brain network supports the social context, questions around the triggers for such processing remain.

Social Cognition

Perhaps the most reliable result in the social neuroscience literature is the involvement of the social cognition brain network when people complete tasks involving people. This network involves

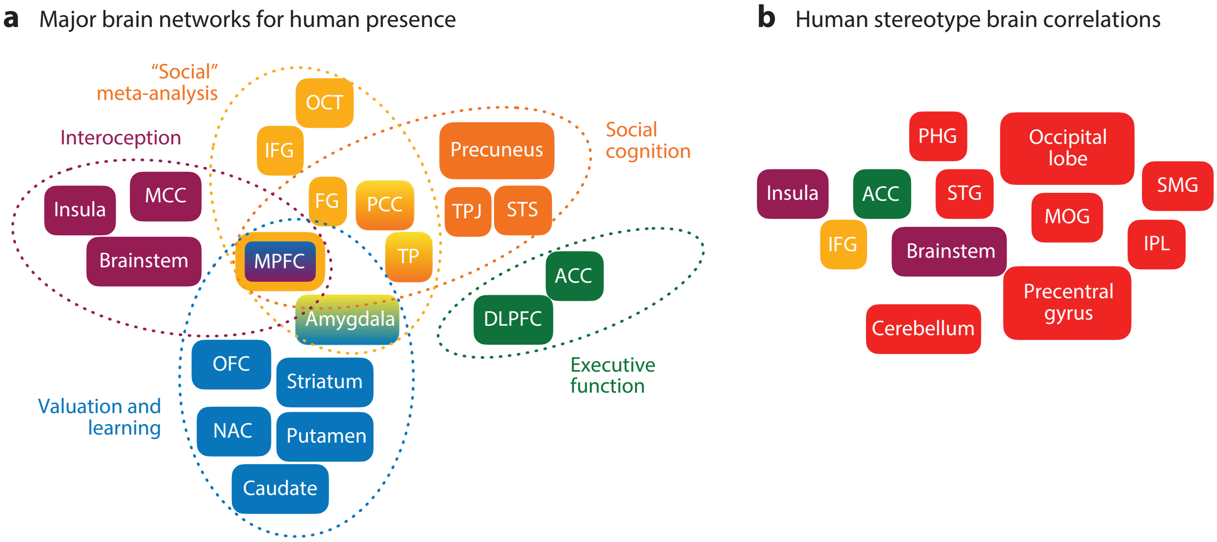

Figure 1

(*a*) Major brain networks for human presence. This map depicts brain regions involved in the four major psychological processes involved in human presence detection: the "social" brain regions more active in a Neurosynth meta-analysis (Tso et al. 2018) (*light orange*), the social cognition brain network (*dark orange*), the executive function brain network (*green*), the valuation and learning brain network (*blue*), and the interoception brain network (*purple*). The MPFC overlaps with three of the four networks (all but the executive function network). (*b*) Human stereotype brain correlates. This panel depicts brain regions that correlate with traits associated with the human stereotype (Harris & Fiske 2011). Brain regions represented in panel *a* retain their respective colors, and novel brain regions are depicted in red. Abbreviations: ACC, anterior cingulate; DLPFC, dorsolateral prefrontal cortex; FG, fusiform gyrus; IFG, inferior frontal gyrus; IPL, inferior parietal lobule; MCC, middle cingulate; MOG, middle occipital gyrus; MPFC, medial prefrontal cortex; NAC, nucleus accumbens; OCT, occipital-temporal junction; OFC, orbitofrontal cortex; PCC, posterior cingulate; PHG, parahippocampal gyrus; SMG, supramarginal gyrus; STG, superior temporal gyrus; STS, superior temporal sulcus; TP, temporal pole; TPJ, temporal-parietal junction.

the medial prefrontal cortex (MPFC) and areas of the temporal lobe, from the temporal-parietal junction (TPJ) along the superior temporal sulcus (STS) to the anterior temporal pole (ATP), including the FG, precuneus, and PCC (Mars et al. 2012; see **Figure 1*a***). Also termed the default mode network, this reliable network is engaged whenever people perceive, think about, or make decisions involving other people. Though early research focused on just parts of this network and its involvement in mentalizing (see Amodio & Frith 2006, Frith & Frith 2001, Gallagher & Frith 2003, Van Overwalle 2009), there is enough evidence in the more than two decades of social neuroscience research to suggest that the network extends beyond the MPFC and TPJ.

Valuation and Learning

The once disparate psychological constructs of subjective valuation and learning often localize to the same set of brain regions, suggesting they have more in common than simply guiding decision making. These traditional brain regions include the VMPFC; medial orbitofrontal cortex (MOFC); striatum, including caudate, putamen, and nucleus accumbens (NAC); and amygdala (Rangel et al. 2008; see **Figure 1*a***). More specifically, social decision making involves these traditional learning and valuation mechanisms as well as other brain regions associated with social cognition and emotion regulation (Rilling & Sanfey 2011).

Executive Function

There is also a reliable engagement of a brain network involved in executive function, cognitive control, conflict monitoring, and emotion regulation, in which a response must be inhibited to facilitate a different response. This network includes the anterior cingulate (ACC) and dorsolateral prefrontal cortex (DLPFC) (Botvinick et al. 2001; see **Figure 1*a***). The ACC is thought to change the weighting of the response options, while the DLPFC assigns the weights to the response options.

Interoception

Being aware of and monitoring internal bodily signals involve a reliable set of brain regions (Critchley & Garfinkel 2017). These brain regions center on the insula and extend to the brainstem and middle cingulate (MCC), along with the MPFC (see **Figure 1*a***). The insula is considered the key node of this network, allowing for awareness of internal signals (Craig 2009).

Human Stereotype

There is less brain imaging work exploring the human stereotype. A single study examined the brain correlates while participants rated 60 varied human targets on several previously rated dimensions consistent with the stereotype of what people consider human (including human typicality, warmth, competence, ease of inferring mental state, similarity, and familiarity; Harris & Fiske 2011). Brain correlates of these human perception dimensions include the anterior and posterior insula, ACC, superior temporal gyrus (STG), parahippocampal gyrus (PHG), occipital lobe, middle occipital gyrus (MOG), cerebellum, IFG, precentral gyrus, inferior parietal lobule (IPL), supramarginal gyrus (SMG), and brainstem (see **Figure 1*b***). The IFG is more associated with “social” as a search term in brain data repositories, and the STG sits next to the STS, while the ACC is associated with executive function and the insula with interoception. This set of results presents additional brain regions to consider when comparing responses to humans to responses to AIs.

COMPARING RESPONSES TO HUMANS AND TO AI

In this review of the literature, I focused on studies comparing responses to humans and to AI. I performed a literature search on PubMed and PsycInfo during the first week of October 2022 with the search terms (human AND (computer OR algorithm OR lottery) AND fMRI AND social). I limited the search to human participants and peer-reviewed English-language papers, excluding book chapters, review papers, and dissertations. I excluded other comparisons, such as those of humans and nonhuman animals. I also excluded nonadult and clinical populations. I only considered papers whose authors performed and reported contrasts in responses to human and AI stimuli. In most papers, this is not the main comparison of interest but is nonetheless reported. Cases where these types of stimuli were used, but the contrast was not performed or reported, are also excluded from this review. Moreover, I only considered cases with neurotypical populations. Studies where participants were administered drugs before completing experiment paradigms were also excluded. Finally, I only considered studies using fMRI, functional near infrared spectroscopy (fNIRS), or PET.

Two literature reviews capture comparisons of responses to human and to AI targets, though restricted by psychological domain. The first, published a decade ago, focused on decision-making paradigms and primarily compared responses to humans and disembodied AIs (Lee & Harris 2014), concluding that human partners drove more activity in the social cognition, valuation, and learning brain networks relative to AIs, but AIs did not drive more brain activity compared to humans. The second review, published more recently, focused on the uncanny valley, and as such it explored mainly the social perception paradigms related to humans and to embodied AIs (robots; Vaitonyte et al. 2023). This review found more engagement of TPJ in response to humans, and more engagement of the precuneus and VMPFC in response to AIs. These two reviews highlight an interesting finding: Embodied AI drove more engagement of some brain regions relative to humans, while disembodied AI did not, with humans driving more brain activity than both embodied and disembodied AI. I included papers from both reviews that met the criteria highlighted above.

I map the activation patterns found in the studies reviewed below to the brain networks described above (see **Table 1**). This provides broader reach than simply considering the social cognition brain network, allowing hypotheses about the involvement of certain types of cognitive processes in human and AI presence.

Experimental Task Design

There are several factors that limit the inferences one can make when comparing across the brain imaging studies described below because of how such experiments are designed and executed. Besides the fact that most of the studies were not interested in directly comparing responses to humans and to AIs, and therefore were not designed to reflect this comparison, there is huge variability in how the human and the AI were operationalized. In most cases, the human was operationalized as an unknown other person who largely remained anonymous, but in some studies participants would meet their human partner before scanning. Sometimes, the human was a friend or known person, and occasionally the human was the experimenter. In a few cases, the human was gender matched to the participant.

Of course, it is not possible to match the level of familiarity across humans and AIs, so AIs were often operationalized in a very different manner. Disembodied AI was described as a computer algorithm that either behaved randomly or attempted to mimic human behavior by behaving intentionally or by using human behavior to determine how to behave. Sometimes, disembodied AI was described as a lottery. Embodied AI often took the form of robots that were either humanoid (including androids) or not, including digital AIs such as avatars. Machines were sometimes used as embodied AIs, as were human faces.

In addition, the nature of the task occasionally changed when participants played against humans and against AIs. For instance, in many decision-making paradigms, because AIs were often described as behaving randomly or as lotteries, the task became nonstrategic and based on luck rather than being a strategic task where another agent's motives and intentions had to be considered. Yet in other paradigms, the AI was described as intentional, even mimicking human behavior, or represented with a human face.

Researchers are also not consistent in the type of brain imaging analysis performed in their studies. Because of the assumption that humans should engage the social cognition brain network more than AI, some studies localized this brain network before performing region of interest (ROI) analyses, ignoring other brain regions that may dissociate the two types of agents. More recent studies tend to perform psychophysiological interaction (PPI) analyses, using a single brain region or multiple brain regions as seeds to reveal brain regions that display a similar activation profile. These studies often moderate these analyses with a variable of interest relevant to their research program (e.g., age). However, the majority of studies rely on whole brain contrasts that reveal brain regions more engaged by one type of agent relative to the other.

Finally, though I excluded studies that examined human–AI comparisons with electroencephalograms, I included studies that used PET or fNIRS, since these studies could localize brain activity to specific regions and structures. In addition, the pipeline of brain imaging analysis differs across labs as well as the software used to analyze the brain data, the magnets used to collect

Table 1 Brain imaging studies comparing responses to humans and to AI

DECISION MAKING						
Authors	**Phenomenon**	**Psychological variable**	**Human agent**	**Nonhuman agent**	**Brain regions more active in response to human**	**Brain regions more active in response to AI**
Schindler et al. (2019)	Personality feedback	Social feedback (like/dislike)	Human (unknown)	Random computer algorithm	(Bi)SFG, (Bi)MPFC, (Bi)OFC, (L)precuneus, (L)PCC	None
Kätsyri et al. (2013)	First-person shooter game	Social feedback (win/loss)	Human (unknown)	Computer	(Bi)precuneus, (Bi)MPFC[a]	None
Van den Bos et al. (2007)	Time estimation task	Social feedback (correct/incorrect)	Human (experimenter)	Computer	(Bi)TP	None
Delgado et al. (2008)*	Auction	Social feedback (win/loss)	Human (unknown)	Lottery controlled by computer	(R)precuneus, (L)IPL	(R)PCL, (Bi)cuneus, (Bi)IFG, (Bi)MTG, (R)MOG, (Bi)LG, (R)FG, (R)STG, (L)MPFC, (R)hippocampus, (L)PHG
Coricelli & Nagel (2009)*	Beauty contest game	Decision making	Human (unknown)	Computer	(Bi)VMPFC, (Bi)DMPFC, (L)RCC, (L)STS, (L)PCC, (Bi)TPJ, (L)DLPFC, (Bi)LOFC	None
McDonald et al. (2020)	Penalty shot task	Decision making	Human (unknown)	Computer algorithm mimicking human behavior	(R)TPJ, (R)DMPFC, (Bi)IFC	None
Wen & Hsieh (2015)	Joint Simon task	Executive function	Human (unknown)	Computer	(Bi)ACC/MPFC/SFG	None
Rilling et al. (2004b)*	Prisoner's dilemma game	Player offer (compete/cooperate)	Human (unknown)	Computer	(R)DLPFC, (R)STG, (Bi)FG, (R)precentral gyrus, (R)STS, (R)IFG, (R)SFG, (R)PCC, (R)FP, (R)caudate, (L)cerebellum, (L)MFG	(Bi)IPL, (R)MFG, (R)postcentral gyrus, (R)precentral gyrus
Rilling et al. (2002)*	Prisoner's dilemma game	Player offer (compete/cooperate)	Human (unknown)	Computer	(Bi)PCL, (R)caudate, (L)postcentral gyrus, (R)MPFC, (L)STG	(L)insula, (Bi)OFC, (L)anterior insula, (L)FP[b]
Rilling et al. (2004a)*	Prisoner's dilemma game	Player offer (compete/cooperate)	Human (unknown)	Computer	(Bi)VMPFC	None

(Continued)

Table 1 *(Continued)*

DECISION MAKING						
Authors	**Phenomenon**	**Psychological variable**	**Human agent**	**Nonhuman agent**	**Brain regions more active in response to human**	**Brain regions more active in response to AI**
Chen et al. (2016)	Prisoner's dilemma game	Player offer (compete/cooperate)	Same sex confederate (unknown)	Computer game (pre-programmed algorithm)	(Bi)amygdala, (Bi)anterior insula[c]	None
Krach et al. (2009)	Prisoner's dilemma game	Entire decision-making process	Human (unknown)	Computer	(Bi)SFG, (L)cerebellum, (R)MTG, (R)ACC, (Bi)thalamus, (R)olfactory cortex, (R)ITG, (R)cerebellum, (R)SMA[d], (R)AG, (R)PCC[e]	None
Krach et al. (2008)*	Prisoner's dilemma game	Entire decision-making process	Human (unknown)	Anthropomorphized robot, functional robot, computer	(R)AG, (R)TPJ, (R)FEF	None
Koban et al. (2014)	Conflict game	Decision making	Human (unknown confederate) met before scanning	Computer (pre-programmed)	(Bi)DMPFC[f]	None
Assaf et al. (2009)	Domino game	Decision making	Human (unknown)	Computer	(Bi)TPJ, (Bi)TP, (R)FG, (L)MPFC[g]	None
Singer et al. (2004)*	Prisoner's dilemma game	Faces of partners	Human (unknown)	Computer based on population statistics (via human face)	(R)FG, (L)MTG, (L)IFG, (R)PSTS, (L)ASTS, (Bi)anterior insula, (L)posterior insula, (R)OFC, (R)putamen, (Bi)NAC, (Bi)LOFC, (Bi)POF, (L)LG, (Bi)RSC, (R)ITG, (R)AMTG, (R)precentral sulcus, (L)cuneus	None
Sanfey et al. (2003)*	Ultimatum game	Unfair offers	Human (unknown)	Computer	(Bi)insular cortex	None
Lee & Harris (2014)	Investment task	Decision and anticipation of outcome	Human (unknown)	Computer	(Bi)IFG, (R)ACC	None
McCabe et al. (2001)*	Trust game	Decision to trust	Human (unknown)	Computer	(Bi)MPFC[h]	None
Phan et al. (2010)*	Trust game	Outcomes of decisions to trust	Human (unknown)	Computer	None	None

(Continued)

Table 1 (*Continued*)

DECISION MAKING						
Authors	**Phenomenon**	**Psychological variable**	**Human agent**	**Nonhuman agent**	**Brain regions more active in response to human**	**Brain regions more active in response to AI**
Sun et al. (2015)	Modified trust game	Honesty detection	Human (unknown)	Computer	None	None
Fareri et al. (2022)	Trust game	Decision making	Self-selected friend of same sex and age group, strangers	Computer (reciprocate 50% of time)	(R)SMA, (R)TPJ, (L)OP[i]	None
Chaminade et al. (2015)	Rock-paper-scissors	Decision making	Human (unknown)	Intentional robot	(R)TPJ, (Bi)precuneus, (R)STS, (Bi)MPFC, (R)SFG	None
Chaminade et al. (2012)	Rock-paper-scissors	Decision making	Human (unknown)	Intentional robot Random computer program	(R)TPJ (L)SFG, (Bi)precuneus, (L)AIPS, (L)precentral gyrus, (L)MFG, (L)PIPS, (R)MPFC, (R)TPJ, (R)aThalamus	None
Gallagher et al. (2002)[j]*	Rock-paper-scissors	Decision making	Human (experimenter)	Computer	(Bi)apCC	None
Piva et al. (2017)	Poker game	Honesty detection	75% strangers, 25% acquainted	Computer (invariant fixed optimal strategy)	(L)DLPFC, (L)FPA, (Bi)DLPFC, (L)ptBroca, (L)PSSC[k]	None
Carter et al. (2012)*	Poker game	Honesty detection	Human (unknown)	Computer	TPJ	None
De Quervain et al. (2004)*	Third-party punishment	Altruistic punishment	Human (unknown)	Random device (via human opponent)	(R)caudate	None
Xiong et al. (2020)	Altruistic (pain relief) task	Altruism	Human (unknown)	Random computer	(L)FG, (Bi)MOG, (R)cerebellum, (R)aINS/lOFC, (R)DMPFC, (R)dACC, (Bi)ITG, thalamus[l]	None
Zhang et al. (2016)	Modified dictator game	Decision making	Human (unknown)	Computer	(Bi)LOFC, (Bi)VMPFC[m], (Bi)DMPFC, (Bi)aINS, (Bi)precuneus, (Bi)AG, (L)SFG, (L)putamen, (R)STS[n]	None

(*Continued*)

Table 1 (*Continued*)

SOCIAL PERCEPTION						
Authors	**Phenomenon**	**Psychological variable**	**Human agent**	**Nonhuman agent**	**Brain regions more active in response to human**	**Brain regions more active in response to AI**
Kätsyri et al. (2020)	Emotion inference	Person perception	Human (unknown)	Computer-generated faces matched to human faces	(L)FG, (L)SFG, (Bi)ITG/MTG	(L)GR, (L)NAC
Ikeda et al. (2017)[§]	Emotion inference	Person perception	Human (unknown)	Android	None	(R)subthalamic nucleus, (R)calcarine sulcus
Gobbini et al. (2011)	Passive viewing of emotional expressions	Person perception	Human faces	Robot faces	(R)mPFC, (L)ATC, (R)TPJ, (R)cuneus, (R)amygdala	(R)MOG, (Bi)FG, (R)PHG, (Bi)MTG, (Bi)IPS, (L)PrcG, (Bi)MFG, (Bi)IFG, (Bi)SMA, (L)NAC, (R)INS, (L)thalamus, (R)caudate, (Bi)cerebellum
Özdem et al. (2017)	Eye gaze cueing	Attention	Programmer (stranger)	Robot (pre-programmed)	(Bi)TPJ, (R)cuneus, (R)SPL, (R)postcentral gyrus	None
Desmet et al. (2014)	Observing human–machine interactions	Person perception	Human (unknown)	Everyday machine	(Bi)precentral gyrus, (R)SPL, (Bi)MTG	None
Wang & Quadflieg (2015)[§]	Helping judgment from dyad	Person perception	Human (unknown)	Robot	(L)TPJ	(Bi)MOG, (Bi)ITG, (R)precuneus, (R)DMPFC, (L)VMPFC
Carter et al. (2011)	Passive viewing	Person perception	Human (unknown)	Humanoid robot, box-like machine	(Bi)MePC	None
Saygin et al. (2012)[§]	Passive viewing	Person perception	Human (unknown)	Android, robot	(L)LTC	(R)FG, (Bi)AIPS
Chaminade et al. (2007)	Alternative forced-choice task	Person perception	Human (unknown)	Robot	None	None
Cheetham et al. (2011)[§]	Target monitoring task	Person perception	Human (unknown)	Avatar	(R)cerebellum, (Bi)FG, (L)precuneus, (R)MCC, (L)insula	(R)FG, (R)STG, (R)MFG
Ganesh et al. (2012)	Referential processing task	Person perception	Self and close and distant familiar others	Avatar representing the self and cartoon	None	(L)AG, (Bi)RCC
Rosenthal-von der Pütten et al. (2019)[§]	Rating task	Person perception	Humans (unknown) with and without physical disabilities	Android, humanoid, mechanistic robots, artificial (synthetic humans)	None	(R)DMPFC, (R)VMPFC, (R)TPJ, (L)PCC[o], (R)TPJ[p]
Anders et al. (2015)	Text message communication task	Social communication	Friend	Computer	(R)precuneus, (L)TPJ, (R) DMPFC, (Bi)TP	None

(*Continued*)

Table 1 (*Continued*)

SOCIAL PERCEPTION						
Authors	**Phenomenon**	**Psychological variable**	**Human agent**	**Nonhuman agent**	**Brain regions more active in response to human**	**Brain regions more active in response to AI**
Hogenhuis & Hortensius (2022)	Listening and speaking during interaction	Person perception	Gender-matched confederates	Robot, semi-human form, face	(Bi)cerebellum, (L)SPL, (Bi)IPL, (L)MPFC, (L)MTG, (Bi)TP	(Bi)Heschl's gyrus, (Bi)LOC, (R)CS, (Bi)INS, (L)IFG, (R)cerebellum, (L)OFG
Di Cesare et al. (2016)	Listening	Person perception	Human (unknown)	Robot (vocal synthesizer)	(Bi)MTG, (Bi)PMFG, (L)precentral gyrus, (L)postcentral gyrus, (L)MCC, (L)putamen, (Bi)SMG, (L)anterior insula	None

Abbreviations: ACC, anterior cingulate; AG, angular gyrus; aINS, anterior insula; AIPS, anterior intraparietal sulcus; AMTG, anterior middle temporal gyrus; apCC, anterior paracingulate cortex; ASTS, anterior superior temporal sulcus; ATC, anterior temporal cortex; aThalamus, anterior thalamus; (Bi), bilateral; CS, calcarine sulcus; DLPFC, dorsolateral prefrontal cortex; DMPFC, dorsomedial prefrontal cortex; FEF, frontal eye fields; FG, fusiform gyrus; FP, frontal pole; FPA, frontopolar area; GR, gyrus rectus; IFC, inferior frontal cortex; IFG, inferior frontal gyrus; INS, insula; IOFC, inferior orbitofrontal cortex; IPL, inferior parietal lobule; IPS, intraparietal sulcus; ITG, inferior temporal gyrus; (L), left; LG, lingual gyrus; LOC, lateral occipital cortex; LOFC, lateral orbitofrontal cortex; LTC, lateral temporal cortex; MCC, middle cingulate; MePC, medial posterior cortex; MFG, middle frontal gyrus; MOG, middle occipital gyrus; MPFC, medial prefrontal cortex; MTG, middle temporal gyrus; NAC, nucleus accumbens; OFC, orbitofrontal cortex; OFG, occipital fusiform gyrus; OP, occipital pole; PCC, posterior cingulate; PCL, paracentral lobule; PHG, parahippocampal gyrus; PIPS, posterior intraparietal sulcus; PMFG, posterior middle frontal gyrus; POF, parieto-occipital fissure; PrcG, precentral gyrus; PSSC, primary somatosensory cortex; PSTS, posterior superior temporal sulcus; ptBroca, pars triangularis Broca's area; (R), right; RCC, rostral cingulate; RSC, retrospenial cortex; SFG, superior frontal gyrus; SMA, supplementary motor area; SMG, supramarginal gyrus; SPL, superior parietal lobule; STG, superior temporal gyrus; STS, superior temporal sulcus; TP, temporal pole; TPJ, temporal-parietal junction; VMPFC, ventral medial prefrontal cortex.

[*] Reviewed in Lee & Harris (2014).

[§] Reviewed in Vaitonyte et al. (2023).

[a] Activation results from interaction contrast where win–lose was factored into the regression model.

[b] These activations result from comparing the outcome when both players cooperate with the other three possible outcomes in the prisoner's dilemma game independently for humans and computers.

[c] Human conditions modulate activity in these brain regions when statistically controlling for sex and drug effects.

[d] Only in male participants.

[e] Only in female participants.

[f] Activation results from interaction contrast where conflict–no conflict was factored into the regression model.

[g] Brain regions result from a region of interest analysis rather than a whole brain analysis.

[h] Only for the six highest trusting participants, not the six lowest.

[i] Default mode network used as a seed in psychophysiological interaction (PPI) analysis with age as a moderator for all brain regions.

[j] Positron emission topography study, not functional magnetic resonance imaging.

[k] AG used as a seed in a PPI analysis with choice history as a moderator just for final brain regions.

[l] Brain regions modulate by social behavior, not as a result of a direct contrast.

[m] VMPFC used as a seed in a PPI analysis with waive/retain punishment as a moderator. VMPFC defined from activity in Li et al. (2009).

[n] LOFC used as a seed in a PPI analysis with waive/retain punishment as a moderator. LOFC defined from activity in Spitzer et al. (2007).

[o] DMPFC used as a seed in a PPI analysis with nonhuman/human contrast.

[p] FG used as a seed in a PPI analysis with nonhuman/human contrast.

the data, and the decisions taken regarding smoothing, high and low pass filtering, and the brain template [Talairach or Montreal Neurological Institute (MNI)] to which all data were fitted. This is more variability than one might find with behavioral paradigms in which, despite differences between labs, the paradigm and data collection mechanisms are still relatively homogeneous. Therefore, any of these variables could account for different brain results, making consistency, when found, more remarkable and more likely to be indicative of underlying brain mechanisms.

Decision Making

The most popular psychological phenomenon in the literature comparing brain activations in response to humans and AI is decision making. This vast literature includes financial and social decision making, social feedback, and moral questions regarding fairness, trust, and reciprocity. I organize the discussion of this literature by the paradigms commonly used, including the prisoner's dilemma game, trust game, ultimatum game, rock-paper-scissors game, and others (for a review of social decision-making paradigms, see Van Dijk & De Dreu 2021). Most studies involve disembodied AI, and few report more engagement in response to AIs than in response to humans.

Prisoner's dilemma game. This game involves coordinating behavior with a partner to optimize one's outcomes. Partners and players must make binary decisions to cooperate or to defect. If both players choose the same option, payouts are usually lower than if a player chooses to defect while the partner chooses to cooperate. Therefore, maximizing outcomes requires inferring the partner's choice and choosing accordingly. This game relies on an inference about the loyalty of the partner, a moral domain, as well as reputation building and impression formation as forms of social learning over repeated trials with the same partner.

Researchers have investigated brain responses to a partner's decision to cooperate and to defect (Chen et al. 2016; Rilling et al. 2002, 2004a,b). In these studies, researchers tend to use exclusively computers as the AI. There was no difference in choices to cooperate with human or AI partners, though men were more likely to cooperate with computer rather than human partners following trials where they had cooperated but their partner defected (Chen et al. 2016). Cooperation by the player (Rilling et al. 2004a,b) and mutual cooperation between both players (Rilling et al. 2002) were also reported as more unlikely in the AI relative to the human condition.

These studies find increased brain activity in response to humans compared to AIs, including in areas of the striatum, DLPFC, STG and STS, FG, IFG, superior frontal gyrus (SFG), middle frontal gyrus (MiFG) and MPFC,[1] PCC, frontal pole (FP), pre- and postcentral gyrus, cerebellum, VMPFC, amygdala, insula, and paracentral lobule (PCL). This result includes brain regions associated with social cognition, the human stereotype, executive function, valuation and learning, and interoception. There were also other areas not included in any of the networks described thus far.

Interestingly, about half of these studies tend to find more engagement of brain regions in response to AI than in response to humans, suggesting an AI presence in the brain. These regions included the IPL, MiFG, pre- and postcentral gyrus, insula, orbitofrontal cortex (OFC), and FP, which include areas implicated in the human stereotype, interoception, and valuation and learning.

Two studies have explored the entire decision-making process during the prisoner's dilemma game (Krach et al. 2008, 2009). One of the studies has explored embodied AI in the form of anthropomorphized robots (looking like a small child) and functional robots (not looking like humans)

---

[1] Some researchers refer to this area as medial frontal gyrus; however, the term medial frontal gyrus is troublesome since different researchers use it to refer to a variety of brain regions. Throughout this review, I use the brain region that is most consistent with the reported brain map coordinates (here, MPFC).

in addition to AI represented by computers. Participants reported having more fun interacting with and attributed more intelligence to more humanlike partners, with the most fun with the human partners, the least with the disembodied AI, and the robots in the middle (Krach et al. 2008). There were, however, no differences in the self-reported experience of winning and losing against human and AI partners. The second study reported no behavioral differences in interactions with humans and with AI (Krach et al. 2009).

These studies did not find more brain activity in response to AI than to human partners. However, they did find increased brain activity in response to humans relative to both forms of AI, including in MPFC and TPJ when compared to embodied AI and in SFG, cerebellum, and middle temporal gyrus (MTG) when compared to disembodied AI. The different activation patterns in response to embodied and disembodied AIs are such that the former contrast shows activity consistent with social cognition, while the latter shows activity not consistent with any brain networks described thus far but partially consistent with the prisoner's dilemma results described above. They also found the AG, TPJ, and frontal eye fields[2] (FEF) to be correlated with the humanness of the agents.

Moreover, one study explored gender differences in response to disembodied AI partners, and it found that men engaged ACC, thalamus, olfactory cortex, inferior temporal gyrus (ITG), cerebellum, and supplementary motor area (SMA) more in response to humans than to AI, while women engaged the angular gyrus (AG) and PCC more in response to humans than to AI.

Yet another study that utilized the prisoner's dilemma game focused on person perception by analyzing brain responses when the face of the partner was displayed (including human faces used to represent AI partners) (Singer et al. 2004). Participants reported being more emotionally involved with and upset by the action of human relative to AI partners. They also considered the experience to be more real when playing with human relative to AI partners. Human defectors were liked less than AI defectors, and memory for cooperator faces was better for humans than for AI.

These researchers found no increased brain activity in response to AI relative to humans but did find increased brain activity for humans compared to AI in the FG, anterior middle temporal gyrus (AMTG) and ITG, IFG, anterior superior temporal sulcus (ASTS) and posterior superior temporal sulcus (PSTS), anterior and posterior insula, lateral orbitofrontal cortex (LOFC), striatum, lingual gyrus (LG), parieto-occipital fissure (POF), precentral sulcus, retrosplenial cortex (RSC), and cuneus. Some of these brain regions overlap with social cognition, interoception, and valuation, but others are not part of previously reported results or previously described brain networks.

Finally, researchers have used paradigms whose structure is similar to the prisoner's dilemma game. In one such conflict game (Koban et al. 2014), players' choices could be the same as, or different from, those of their partners. Conflict or not depended on the alignment of one's choice with the partner's choice. Players were rewarded on no conflict trials, but on conflict trials, they had to decide whether to share the reward with their partner or keep all of it, as in the dictator game. Therefore, this game is similar to the prisoner's dilemma game because players have to coordinate their behavior with their partners' to maximize outcomes. The AI was a computer, but participants got to meet the previously unfamiliar human partners before scanning.

Participants shared more with human partners, though reaction times did not differ by type of partner. Researchers modeled the entire trial procedure, including the decision and outcome phases. Researchers found no brain activity elicited more by AI than by humans but did find more DMPFC activity in response to humans compared to AI.

---

[2] The researchers describe this area as superior medial frontal gyrus but frontal eye field is preferred here.

In another similar interpersonal competitive game, the domino game (Assaf et al. 2009), players again had to coordinate their behavior with a partner's choices. Specifically, the player chose to either match or not one of the numbers on their domino tile with the number on the partner's tile. The partner remained unaware of the player's choice and could either reveal their choice or not. If the choice was revealed and there was a mismatch the player lost, whereas they were rewarded for a match. Failure to reveal the choice would forfeit these rewards and punishments. The AI was a disembodied computer.

Participants wanted to win the game to the same extent against human and AI partners and did not differ in their tile selection strategy for either partner. They also considered the partner's move to the same extent regardless of partner type, and players' risk-taking behavior in the scanner did not differ between interactions with human and with AI partners. Humans elicited more brain activity than the AI in TPJ, TP, FG, and MPFC. This activity is consistent with the social cognition brain network because the researchers explored these brain regions as ROIs rather than obtaining the comparison from a whole brain contrast.

Therefore, the prisoner's dilemma game engages all four major brain networks, as well as a host of other brain regions, more in response to humans than to AI (see **Figure 2*a***). This suggests that human presence requires brain involvement including, but also beyond, the four major brain networks in the prisoner's dilemma game. On the other hand, AIs engage fewer brain regions when compared to humans, but they involve brain regions associated with interoception, valuation and learning, and the human stereotype as well as brain regions not associated with the four major brain networks (see **Figure 2*b***). This suggests that AI presence does not involve the social cognition brain network in the prisoner's dilemma game.

Trust game. The trust game is one of the most popular behavioral economic paradigms, yet there are not many brain imaging studies that use it to compare responses to humans and to AIs. The game requires the player to decide whether to trust a partner or not with an investment. If the partner is trusted, the investment is multiplied, and the partner can decide how much to return to the player. Therefore, this game requires trustworthiness inferences of a partner, a fundamental person perception dimension and a rudimentary form of mentalizing. This game relies on reputation building and impression formation as forms of social learning over repeated trials with the same partner.

Only two studies used the trust game when comparing brain activations in response to humans and to AI. Both papers used a computer as the AI, and both failed to find increased brain activity in response to AI relative to humans. One paper did not report behavioral differences (McCabe et al. 2001), while the other reported no difference on trustworthiness decisions and in ratings of AI and human partners known to reciprocate 50% of the time.

Both studies also struggled to find increased activity in response to humans compared to AI. In one paper that focused on the deliberation before deciding to trust, only the half of the sample that engaged in more trusting behavior showed increased activity in MPFC when interacting with humans relative to AI (McCabe et al. 2001). In the second paper that examined reactions to the partner's decision to reciprocate trust, no brain regions differentiated between responses to humans and to AI (Phan et al. 2010).

Another study had participants play the role of trustees and receive the investments from their partners (Sun et al. 2015). Participants were also told how their partners wanted to split the profits. Participants then decided whether to go along with the proposed split or not. Finally, participants' decisions to go along with the proposal or not were revealed on half the trials and not revealed on the other half. If the decision was revealed, the participants received nothing, but if it was not revealed, then the proposed split went through. However, there were no behavioral or brain differences in interactions with human and AI partners reported in this study.

(Caption appears on following page)

Figure 2 (*Figure appears on preceding page*)

Brain maps depicting activation patterns in response to human and to AI in decision-making paradigms. Colors represent the major brain networks: the "social" brain regions more active in a Neurosynth meta-analysis (Tso et al. 2018) (*light orange*), the social cognition brain network (*dark orange*), the executive function brain network (*green*), the valuation and learning brain network (*blue*), the interoception brain network (*purple*), and the human stereotype (*red*). Brain regions that do not belong to any network are depicted in black. Abbreviations: ACC, anterior cingulate; AG, angular gyrus; AIPS, anterior intraparietal sulcus; DACC, dorsal anterior cingulate; DLPFC, dorsolateral prefrontal cortex; FG, fusiform gyrus; FP, frontal pole; FPA, frontopolar area; IFC, inferior frontal cortex; IFG, inferior frontal gyrus; IOFC, inferior orbitofrontal cortex; IPL, inferior parietal lobule; ITG, inferior temporal gyrus; LG, lingual gyrus; MiFG, middle frontal gyrus; MOG, middle occipital gyrus; MPFC, medial prefrontal cortex; MTG, middle temporal gyrus; OFC, orbitofrontal cortex; OP, occipital pole; PCC, posterior cingulate; PCL, paracentral lobule; PHG, parahippocampal gyrus; POF, parieto-occipital fissure; PSSC, primary somatosensory cortex; RCC, rostral cingulate; RSC, retrosplenial cortex; SFG, superior frontal gyrus; SMA, supplementary motor area; STG, superior temporal gyrus; STS, superior temporal sulcus; TP, temporal pole; TPJ, temporal-parietal junction.

Another study explored trust behavior when players had friends, strangers, or a disembodied AI (computer) as partners in the trust game (Fareri et al. 2022). In this game, researchers found that players trusted friends more than strangers, and older participants showed longer reaction times after violations of trust by computer partners. Brain imaging analyses used the social cognition brain network as a seed in a PPI analysis, moderated by age, and found correlated activity in the SMA, TPJ, and occipital pole (OP) (Fareri et al. 2022).

A final study used an investment task similar to the trust game (Lee & Harris 2014). Participants had to choose one of three partners with whom to invest, before getting a return on the investment framed as due to the partner's generosity or ability to turn a profit. Here, disembodied algorithms served as the AI and were represented by named geometric shapes.

Participants showed no difference in trait attributions to humans and to AI. However, trait warmth information about humans, not AI, hampered learning about best investment strategy, while trait competence information about both types of partners did not. A preference for interacting with human rather than AI partners was present before and after scanning. The researchers focused on the decision, anticipation, and outcome phases and found no increased brain activity in response to AI compared to humans but increased activity in response to humans compared to AI in the IFG during the decision and anticipation phases and in the ACC during the anticipation phase.

These results suggest that the trust game elicits engagement of social cognition and executive function brain networks more for humans than for AI, along with other brain regions beyond the four major brain networks (see **Figure 2*c***). They also highlight the difference between a trust game context and the prisoner's dilemma game, showing how human presence changes in different social contexts, even when engaged in the same psychological function (decision making). There is no increased activity to report in response to AI relative to humans.

Ultimatum game. The game involves the moral domain of fairness, since players can willingly forgo money to deny a partner money if they feel they have been treated unfairly by that partner. The game begins with the partner being given a sum of money that they must split equally between themselves and the player. If the player does not agree with the proposed split, neither the player nor the partner receives anything. Rational economic theory suggests that players always accept offers greater than zero, since some money is better than no money, but behavioral data show that people are willing to forgo up to 40% of the split to punish a partner who is unfair. This game therefore involves responding to perceived unfairness by deciding whether to punish a partner at a personal cost. This game tends not to be played repeatedly against the same partner but is instead played against different partners.

Inhibitory transcranial magnetic stimulation of the DLPFC during the ultimatum game shows that players are more willing to accept unfair offers from human partners (Baumgartner et al. 2011, Van't Wout et al. 2005). However, inactivating the DLPFC has no such effect on offers from AI partners, suggesting that the DLPFC has a causal role in fairness judgments involving humans.

Researchers have conducted a single brain imaging study exploring differences between human and AI partners in the ultimatum game (Sanfey et al. 2003). Participants were more likely to reject unfair offers from human than from AI partners. Moreover, the insular cortex was more engaged in response to human than to AI partners, with no brain activity more engaged by AI than human partners.

These results suggest the ultimatum game engages interoception more for human than for AI partners (see **Figure 2*d***), and a node in the executive function network is necessary for responses to unfair human offers. This further illustrates the differences between contexts and highlights the involvement of the four brain networks primarily in instantiating a human presence in the ultimatum game context. There is no increased brain engagement in response to AI relative to humans.

Rock-paper-scissors game. This is another competitive game that researchers have used to explore brain responses to humans and to AI. In this game, players choose one of three responses (rock, paper, scissors). Each response has a win and a loss versus each of the remaining two, such that no response is dominant over the set of three. Players must choose a response they think will win against their partner's response. This game therefore involves inferences of the partner's mind, though no moral domain is being interrogated. The game relies on reputation building and impression formation as forms of social learning over repeated trials with the same partner.

In one such paper, researchers reported that participants experienced the human and AI conditions differently in self-reports, but there were no performance differences between the two types of partners (Gallagher et al. 2002). Participants were told that they were playing against the experimenter and a computer. Interestingly, participants varied greatly in how they described the AI, from something whose strategy they correctly understood to something "baffling." In another study, participants did not perceive differences in their success in the game based on whether they were playing against human or AI partners (Chaminade et al. 2012). In this later game, the AI was either embodied (robot) or disembodied (random computer).

The researchers analyzed the entire epoch of a game trial, including player choice and outcome, across all studies. They found increased brain response to humans relative to AI in the anterior paracingulate cortex (also known as MPFC), and no increased brain activity in response to AI relative to humans (Gallagher et al. 2002). The other study also found more activity in the MPFC for humans relative to disembodied AIs as well as more activity in the SFG and MiFG, precuneus, anterior intraparietal sulcus (AIPS), precentral gyrus, TPJ, and thalamus (Chaminade et al. 2012). When looking at robot partners, this study reported more engagement of TPJ with human partners. The final study also found more brain activity in response to human than to AI partners in the TPJ, precuneus, STS, MPFC, and SFG (Chaminade et al. 2015). These results include brain regions implicated in social cognition, among other tasks, and are consistent with other brain activation patterns described by other decision-making studies above.

The results above suggest the rock-paper-scissors game engages brain regions associated with social cognition and the human stereotype, as well as regions not associated with the four major brain networks, in interactions with humans compared to AI (see **Figure 2*e***). There is no increased brain engagement in response to AI relative to humans.

Poker games. Modeled after the popular game in casinos worldwide, poker games require participants to bet or fold based on the value of the cards they think their partner holds. Partners then make the same decision before outcomes are revealed. This game therefore requires considering what a partner knows and behaving to maximize personal outcomes in a competitive environment. However, this game does require detection of honesty of the partner and therefore inferences about their moral character on each round. This game relies on reputation building and impression formation as forms of social learning over repeated trials with the same partner. Behavior on previous trials influenced bluffing behavior on the subsequent trial more with human than with AI partners (Carter et al. 2012, Piva et al. 2017). Participants did not differ across partner type on reaction time, self-reported task difficulty, or overall performance (points). The consistency of behavior across the pair of studies is more surprising, given that 25% of the human opponents in one study were known.

These paradigms identify more engagement of the TPJ (Carter et al. 2012), DLPFC, and frontopolar area (FPA) (Piva et al. 2017) in response to human relative to AI partners. This activity partially overlaps with social cognition and executive function. Moreover, several brain regions correlated in a PPI analysis with previous opponent behavior—the sole behavior in which human and AI partners differed—including Broca's area and primary somatosensory cortex (PSSC) (Piva et al. 2017).

The results above highlight a single node in the social cognition and executive function brain networks that is more engaged by humans than AI in poker games (see **Figure 2*f***). There is significant involvement of brain regions beyond the four major brain networks, highlighting a different pattern of brain activation associated with human presence in poker games compared to the other decision-making games. There is no increased brain engagement in response to AI relative to humans.

Altruism and punishment games. Researchers have used modifications of classic economic games or developed novel paradigms to study altruistic behavior. In one such study, participants observed a dictator game—an allocation from a dictator to a receiver—before deciding how much to punish the dictator (De Quervain et al. 2004). Punishment was symbolic, free, or costly, with the latter category considered a case of altruistic punishment. The AI was described as a device that made random decisions, a disembodied AI. Participants perceived the AI as less unfair and expressed less of a desire to punish the AI relative to the human partners. Participants engaged the caudate more when punishing the human than when punishing the AI. These brain regions overlap with valuation and learning brain networks.

In another altruism paradigm, partners decided whether to incur a painful stimulus to reduce the severity of the painful stimulus the participant experienced. The partner was either an unknown human or a disembodied AI (a computer). Brain activity focused on the outcome phase, where the partner's decision to help or not modulated several brain regions associated with social behavior, including FG, MOG, cerebellum, anterior insula/inferior orbitofrontal cortex (IOFC), DMPFC, dorsal anterior cingulate (DACC), ITG, and thalamus. This brain activity partially overlaps with human stereotype, interoception, and executive function brain regions.

Another study explored punishment decisions using a modified dictator game (Zhang et al. 2016). Participants played as the dictator, and their unknown human or disembodied AI partner had the ability to retain or waive the ability to punish the participants based on their proposed split. This decision was made before the participant decided to split the resources. Participants allocated more to human than to AI partners when partners decided to waive the ability to punish, and less to human than to AI partners when partners decided to retain the ability to punish. Brain analyses using VMPFC and LOFC seeds defined from activity in similar studies, modulated by waive/retain punishment decisions, showed correlations with DMPFC, anterior insula, precuneus, AG, SFG, putamen, and STS. These activations partially overlap with activations related to valuation and learning, social cognition, interoception, and other decision-making tasks observed in the studies above.

The results above highlight contributions from all four major brain networks, as well as brain regions beyond these networks, when responding to a human presence in altruism and punishment games (see **Figure 2*g***). This broad swath of brain activity is similar to the one observed in prisoner's dilemma and ultimatum games, and it suggests that these games in which morality is heavily involved require similar psychological processes, and many of them involve the instantiation of a human presence. There is no increased brain engagement in response to AI relative to humans.

Social feedback games. Several studies have provided participants with feedback from either a human or an AI partner. These studies explore participants' responses to the feedback, which affects their reputation and makes impression management concerns salient. These studies often demonstrate differences between receiving feedback from humans and from AI, highlighting an important difference between humans and AI—a lack of impression management concerns for the latter.

In one such study, participants completed a structured interview on camera before receiving feedback on their personality from either an unknown human stranger or a disembodied AI, described as a randomly operating computer algorithm (Schindler et al. 2019). This feedback took the form of previously rated positive, negative, or neutral traits that the partner believed best described the participant. While no behavioral data were reported, researchers found more activity in response to human than AI partners in the SFG, PCC, paracingulate gyrus (MPFC), OFC, and precuneus. These activation patterns are consistent with the social cognition, valuation, and decision-making paradigms reported above.

In another study, participants played a first-person shooter video game against either an unknown human or a disembodied AI (computer) partner (Kätsyri et al. 2013). In the game, participants had to hunt and destroy their partner's tank without having their own tank destroyed. Win events occurred when the partner's tank was successfully destroyed, and loss events occurred when the partner successfully destroyed the participant's tank. Participants reported more social presence for human relative to AI partners, as measured by the Social Presence in Gaming Questionnaire (De Kort et al. 2007). However, there were no differences in affective tone, pleasantness, or game scores between interactions with human and AI partners. The researchers found more engagement of precuneus and MPFC, areas consistent with social cognition, during win/loss feedback when playing against human relative to AI partners.

Another study exploring wins and losses used an auction game to instantiate the social context in which participants played against human or AI partners (Delgado et al. 2008). This study was interested in overbidding behavior, a phenomenon whereby people bid more than the value of an item during an auction, presumably because of a desire to win the auction against other bidders rather than a desire to win the item. Here, the disembodied AI was described as a lottery controlled by a computer, while participants briefly met their human partners before scanning. While no difference in behavior was reported, the researchers did find more engagement of the precuneus and IPL in response to the human relative to AI partner. They also found more engagement of the PCL, cuneus, IFG, MTG, MOG, LG, FG, STG, and MPFC in response to AI than in response to human partners. These brain regions are consistent with the human stereotype and are beyond the major brain networks.

A final social feedback study provided participants feedback about their performance on a time estimation task (Van den Bos et al. 2007). Participants received either juice (reward) or quinine (punishment) in response to their time estimation performance. Administration of the liquids was determined either by the experimenter or by a disembodied AI (a computer). Participants changed their time estimates more following punishment from the human partner than from the AI, despite no differences in accuracy across the two partner types. Brain imaging revealed that the TP, a node in the social cognition brain network, was more engaged when receiving feedback from the human relative to the AI partner.

The results above suggest social feedback primarily requires social cognition and valuation and learning, and it possibly involves the human stereotype and brain processing regions beyond the major four networks (see **Figure 2*h***). The focus on valuation and learning and social cognition seems appropriate for a social feedback context. One study (Delgado et al. 2008) reported more brain engagement in response to AI than to humans, and it identified a distributed network beyond the major brain networks, along with social cognition and human stereotype activations.

Other decision-making games. A couple of studies have explored decision making in other scenarios beyond those reviewed above. One such study employed the beauty contest game to probe higher-level thinking in response to human and AI partners (Coricelli & Nagel 2009). This game involves estimating what the average opinion expects the average opinion to be. Stated differently, participants must decide what the average response of a group of people is, and then they adjust that estimate given that they know all other players will estimate the average and make a response themselves. This meta-cognitive task therefore explores more complex mentalizing. In this study, participants played with unknown human or disembodied AI (computer) partners and took longer to make an estimate when playing against the human than when playing against the AI. Brain activity showed more engagement of MPFC, rostral cingulate (RCC), ACC, PCC, STS, and TPJ in response to human relative to AI partners. These activations are consistent with social cognition and partial overlap with executive function.

Another study used a penalty shot game, modeled after the sporting equivalent found in football and hockey (McDonald et al. 2020). Participants had to move a puck across a goal line before their unknown human or disembodied AI (computer algorithm mimicking human behavior) partner could block it, and they were rewarded for trials where they achieved this goal. While no behavior differences were reported between partnering with humans and partnering with AI, participants engaged DMPFC more when informed that their partner would be a human rather than an AI, and they engaged the TPJ and inferior frontal cortex (IFC) more when playing against human than against AI partners. The results for the other decision-making games just discussed enlist the social cognition brain network primarily, with a node involved in executive function, and brain regions beyond the major networks (see **Figure 2*j***). There is no increased brain engagement in response to AI relative to humans.

Executive function. Executive function mechanisms are present in many of the decision-making tasks reviewed above. However, there is only one study that has explored executive function using a traditional paradigm from that literature. This study asked participants to complete a joint Simon task with either a human partner they met before scanning or a disembodied AI (computer) (Wen & Hsieh 2015). Participants were instructed to respond to one color stimulus with a keypress, while their partner responded to a different color stimulus. Thus, the task is a go–no go (response inhibition) task completed with a partner. There were no behavioral differences between playing with human and playing with AI partners. Brain imaging revealed more engagement of ACC, extending into MPFC[3] and SFG, for human relative to AI partners. This result suggests executive function is the additional brain engagement when interacting with humans compared to AI, along

---

[3] The researchers described the medial frontal gyrus, but given the location of the coordinates MPFC is preferred here.

with social cognition and other brain regions beyond the four networks (see **Figure 2*k***). There is no increased brain engagement in response to AI relative to humans.

Social Perception

Researchers have constructed paradigms to explore differences in social perception of humans and of AI. Most of these studies have used robots—an embodied AI—as the AI stimulus, while others have used avatars, machines, and occasionally disembodied AI like computers. These studies involve passive viewing paradigms, social interactions, and eye-gaze and other attentional cueing tasks. These studies tend to find more brain engagement when perceiving AI compared to humans.

Passive viewing of human–robot interactions. A few studies have explored brain differences when participants observe humans and an embodied AI interacting. In one such study (Desmet et al. 2014), participants observed either humans or machines make errors (e.g., a printer malfunction). While no behavioral differences were reported, the researchers found increased activity in the precentral gyrus, superior parietal lobule (SPL), and MTG when participants observed human relative to AI errors.

In another study, participants observed human–robot or human–human dyads interacting and had to indicate whether one of the agents was helping the other (Wang & Quadflieg 2015). Humanoid robots therefore served as the embodied AI. Participants perceived human–AI interactions as involving more helping than human–human interactions. Human–AI interactions were also perceived as more eerie and less believable than human–human interactions, and robots were perceived as less capable of emotions and less intelligent than humans.

Brain imaging revealed increased brain activity in the TPJ, a node in the social cognition brain network, when participants observed human–human compared to human–AI interactions (Wang & Quadflieg 2015). However, brain imaging also revealed increased brain activity in MOG, ITG, precuneus, DMPFC, and VMPFC when participants observed human–AI compared to human–human interactions. These brain activations partially overlap with human stereotype and social cognition brain activations.

These results suggest that nodes in the social cognition and human stereotype brain networks, as well as brain regions beyond the major networks, are involved in instantiating human and AI presence when people passively view social interactions (see **Figure 3*a***). Interestingly, different nodes are involved when observing humans and AI (see **Figure 3*b***).

Passive viewing of behavior. Some studies required participants to passively view humans or embodied AIs engaging in behaviors while researchers collected brain imaging data. In one such study (Saygin et al. 2012), participants viewed humans and embodied AIs (robots and androids) performing behaviors (e.g., drinking from a cup) and occasionally assessed the truth value of statements about the behavior (e.g., I saw her drinking from a cup). Eye movement and behavioral data revealed no difference in responses to the human and AI targets. Brain imaging data revealed more engagement of the lateral temporal cortex (LTC) for human compared to AI targets, and more engagement of the FG and AIPS for AI compared to human targets.

In another study, participants observed a human, embodied AIs (a humanoid robot and a box-like machine), and a claw performing reaching and grasping actions toward an object (Carter et al. 2011). Participants did not perform a specific task; they simply watched the videos depicting the movements. Brain imaging results revealed increased engagement of MPFC when observing human relative to AI targets.

These results suggest that different nodes of the social cognition brain network, as well as activity beyond the major brain networks, are involved in human and AI presence when people

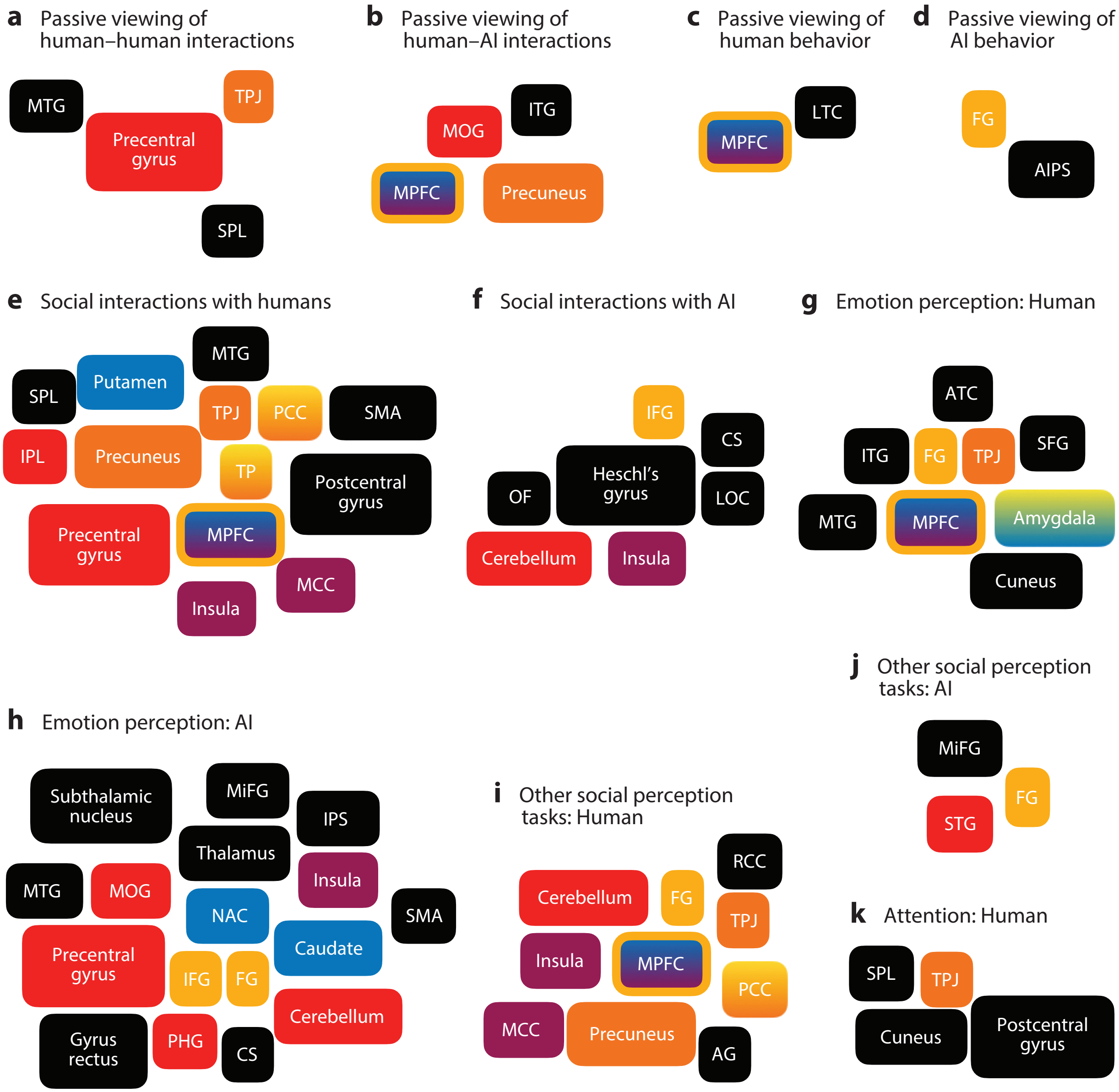

Figure 3

Brain maps depicting activation patterns of response to human and AI in social perception paradigms. Colors represent the major brain networks: the "social" brain regions more active in a Neurosynth meta-analysis (Tso et al. 2018) (*light orange*), the social cognition brain network (*dark orange*), the valuation and learning brain network (*blue*), the interoception brain network (*purple*), and the human stereotype network (*red*). Brain regions that do not belong to any network are depicted in black. Abbreviations: AG, angular gyrus; AIPS, anterior intraparietal sulcus; ATC, anterior temporal cortex; CS, calcarine sulcus; FG, fusiform gyrus; IFG, inferior frontal gyrus; IPL, inferior parietal lobule; IPS, intraparietal sulcus; ITG, inferior temporal gyrus; LOC, lateral occipital cortex; LTC, lateral temporal cortex; MCC, middle cingulate; MiFG, middle frontal gyrus; MOG, middle occipital gyrus; MPFC, medial prefrontal cortex; MTG, middle temporal gyrus; NAC, nucleus accumbens; OF, occipital fissure; PCC, posterior cingulate; PHG, parahippocampal gyrus; RCC, rostral cingulate; SFG, superior frontal gyrus; SMA, supplementary motor area; STG, superior temporal gyrus; TP, temporal pole; TPJ, temporal-parietal junction.

passively observe behavior (see **Figure 3*c***). Notably, there is not a relatively lower number of brain regions involved in observing AI (see **Figure 3*d***).

Social interactions. Researchers have also investigated how people interact with AI and the corresponding brain correlates. Human social interactions can occur in person or via technology, and researchers have explored both types of interactions with AI. In one study (Anders et al. 2015), participants were told that either their friend (who accompanied them to the experiment) or a disembodied AI (computer algorithm) would evaluate the affective state of photographed persons. This task was meant to simulate text messages. Participants would see these messages while in the scanner and had to indicate who sent the message. Behavioral results showed no difference in reaction time to the text messages from humans or AI. Brain imaging results revealed more engagement of the precuneus, TPJ, DMPFC, and TP when receiving a message from a human compared to an AI, an activity consistent with the social cognition brain network.

Another study explored speaking and listening while participants interacted with either a gender-matched human or an embodied AI (a robot with a semihuman form, face, glasses, and wig) (Hogenhuis & Hortensius 2022). Participants discussed a marketing campaign to promote fruits and vegetables with the agents. While no behavioral data were reported, the researchers reported more engagement of the cerebellum, SPL, IPL, PCC, MTG, and TP when people interacted with humans compared to AI, an activity consistent with the human stereotype and social cognition networks. They also found more engagement of Heschl's gyrus, lateral occipital cortex, calcarine sulcus (CS), insular cortex, IFG, cerebellum, and occipital fusiform gyrus (OFG) when people interacted with AI compared to humans, partially consistent with interoception, but they also found largely novel brain region engagements.

A final study considered social interaction from the perspective of the listener; participants were asked to listen to verbs spoken by either human or AI (robotic) voices devoid of vitality forms—the emotional content of speech (Di Cesare et al. 2016). The researchers did not collect behavioral information from the participants. They did report increased engagement of MTG, SMA,[4] pre- and postcentral gyri, MCC, putamen, MPFC,[5] and anterior insula when participants listened to human compared to AI voices. These activation patterns partially overlap with those involved in valuation and learning and with the results reported in the studies above.

These results suggest that all major brain networks, except for executive function, are involved when people engage in social interactions with human beings compared to AI (see **Figure 3*e***). In the case of interactions with AI compared to humans, nodes in these networks are also engaged, along with brain regions beyond these networks (see **Figure 3*f***). Moreover, there is a smaller number of brain regions involved in AI presence than in human presence.

Emotion perception. Very few studies have explored emotion inferences from faces in humans and AI. In one such study (Ikeda et al. 2017), researchers asked participants to infer the emotional expression from embodied AI (an android) and human models of the android from videos. Participants perceived the motion of the AI as less natural than that of the human. They also perceived the positive facial expressions of the human to be more positive than the positive facial expressions of the AI. The researchers reported no brain regions that were more engaged in response to humans than to AI, but they found more engagement of the subthalamic nucleus and the CS when participants observed the AI relative to humans.

Another study explored whether participants would rate the emotional expressions of faces differently for humans and embodied AIs (computer-generated faces) (Kätsyri et al. 2020).

---

[4] The researchers report left (L) posterior medial frontal gyrus but the coordinates suggest (L)SMA.

[5] The researchers report (L) superior medial gyrus but the coordinates suggest (L)MPFC.

Participants reported that angry and fearful expressions were more unpleasant than neutral expressions, a difference that was significantly larger for AI compared to human faces. Brain imaging results revealed more engagement of the FG, SFG, and ITG extending to MTG in response to humans compared to AIs, activity partially consistent with the social cognition and human stereotype networks. The researchers also reported more engagement of the gyrus rectus and nucleus accumbens in response to AI compared to humans, which partial overlap with the valuation and learning brain regions.

Researchers have also explored whether passively viewing facial expressions from humans and AI results in differential brain activity. In one such study, participants passively viewed human and embodied AI (robot) faces displaying emotional expressions (Gobbini et al. 2011). Before scanning, participants rated the faces on the intensity of the emotional expressions and identified the expressions, judging human expressions as more accurate depictions of the emotion and as more intense than AI expressions. Brain imaging results revealed more engagement of the MPFC, anterior temporal cortex, TPJ, cuneus, and amygdala in response to human relative to AI faces, showing higher activation primarily in social cognition brain regions. The researchers also reported more engagement of MOG, FG, PHG, MTG, IPS, precentral gyrus, IFG, MiFG, SMA,[6] NACC, insula cortex, thalamus, caudate, and cerebellum in response to AI compared to human faces. This activity is consistent with social cognition, valuation and learning, human stereotype, and interoception activation patterns as well as with the results reported above.

Similar to engagement in social interactions, emotion perception involves all major brain networks except for executive function, including brain regions beyond these networks, when comparing humans to AI (see **Figure 3*g***). The human stereotype brain regions are also absent for human presence but involved in AI presence (see **Figure 3*h***), as are the brain regions associated with social cognition, valuation and learning, and interoception and brain regions beyond the major networks. Again, human and AI presence differ on the nodes of the brain networks, and again executive function is absent for both.

Other social perception tasks. Some studies asked participants to rate images of humans and embodied AIs while researchers collected brain data. In one such study (Rosenthal-Von der Pütten et al. 2019), participants rated humans with and without physical disabilities and embodied AIs (androids, humanoid and mechanistic robots, and artificial/synthetic humans). Participants found the human targets more likeable, familiar, and humanlike than the AIs. Brain imaging revealed more engagement of the DMPFC, VMPFC, TPJ, and PCC in response to AI relative to human targets when the DMPFC and fusiform gyrus were used as seeds in a PPI analysis. This activity is consistent with social cognition brain activity.

A single study utilized a referential processing task to examine brain differences between perceiving humans and AI (Ganesh et al. 2012). This task requires participants to indicate the extent to which a trait describes themselves or a target (Rogers et al. 1977). In the brain imaging study, the researchers asked participants to evaluate traits pertaining to humans (the self, a close other, a distant other) and to an embodied AI (an avatar) and to recognize the traits used in the experiment post-scanning. While no behavioral data were reported, the researchers did report more engagement of the AG and RCC in response to AI relative to human targets.

A single study employed a target monitoring task to determine brain differences when perceiving human and AI targets (Cheetham et al. 2011). In this task, participants were presented with human or embodied AI (avatar) face pairs and asked to detect the presence of specific upside-down faces. While no behavioral data were reported, the researchers found increased engagement of the

---

[6] The researchers report (L) medial frontal gyrus but the coordinates suggest (L)SMA.

cerebellum, FG, precuneus, MCC, and insular cortex in response to human relative to AI targets. This activity shows partial overlap with social cognition and interoception and is consistent with activation in other paradigms reported above. They also found increased engagement of the FG, STG, and MiFG in response to AI relative to human targets, consistent with the results reported above.

A single study used an alternative forced-choice paradigm to investigate brain differences between perceptions of humans and AIs (Chaminade et al. 2007). This paradigm involved asking participants to indicate whether the observed movement was made by a biological or an artificial agent. Researchers employed human and embodied AI (robot) targets among others, including point-light displays, ellipses, clowns, and aliens. Participants reported that the human target was more biological than the AI. However, researchers found no brain activation differences in responses to human and AI targets.

Just as engagement in a social interaction and emotion perception, the other social perception tasks involve all major brain networks except for executive function, including brain regions beyond these networks, when comparing humans to AI (see **Figure 3*i***). The human stereotype brain regions are also absent for both human and AI presence, and there is increased engagement in parts of the social cognition brain network, along with brain regions beyond the major networks, in response to AI (see **Figure 3*j***). Human and AI presence differ in terms of the brain networks involved, and again executive function is absent for both.

Attention. A single study has explored the brain correlates of attentional cueing using eye gaze from embodied AI (Özdem et al. 2017). In this study, a robot or human face cued participants via eye movements to the location of a probe on the screen. There were no behavioral differences between humans and AI serving as cues. There was, however, more engagement of the TPJ, cuneus, SPL, and postcentral gyrus when receiving cues from humans relative to AI, showing partial overlap with social cognition activation patterns and with the results reported above. These results highlight more engagement of a node in the social cognition brain network and brain activity beyond the major networks when people receive attentional cues from humans relative to AI (see **Figure 3*k***).

CONCLUSION

The literature reviewed above highlights differences between the brain processing of humans and AI. First, human presence involves brain systems beyond the social cognition brain network when controlling for agency, intentionality, the physical form, and the ability to engage in a social interaction. Second, AIs are not homogenous; the distinction between embodied and disembodied AI reveals a general pattern of increased brain activity relative to humans for the former and not the latter. Third, there is no involvement of executive function in social perception, whereas it is involved in decision making. Fourth, AI presence tends to rely on the same psychological processes as human presence but utilizes different nodes in the brain networks. Further research is needed in psychological domains beyond social perception and decision making to further elucidate the brain correlates of human and AI presence.

Limitations

Several limitations plague the analysis above beyond the traditional concerns surrounding brain imaging techniques. In many studies in the decision-making domain, AI provided both human and AI responses, yet participants believed that they were interacting with humans. This ability of AI to mimic human behavior was highlighted at the start of this review and demonstrates the potential of AI presence to serve as human presence, and it may explain why disembodied AI did not drive brain activation relative to humans. Technologies like ChatGPT can produce exam scripts that pass exams at prestigious universities (Terwiesch 2023), suggesting this power to mimic human presence makes AI a unique agent that the brain cannot differentiate from humans. This may occur because of the lack of a physical presence in disembodied AI or the adoption of the human form in embodied AI. Perhaps the brain can eventually learn to better discriminate AI and humans as AI becomes more integrated into human society.

There is another limitation specific to AI presence: the problem of always studying it in comparison to human presence. There is scant brain imaging research that considers how people perceive and interact with AI beyond comparisons with humans. Given the richness of the representation of humans in the brain, comparisons with AI currently limit our understanding of the brain correlates of AI presence.

With a few exceptions, all the studies reviewed above were conducted in Western, educated, industrialized, rich, and democratic (WEIRD) nonglobal samples (Henrich et al. 2010). This severely limits the generalizability of their findings, since the brains of other human beings in other parts of the world may process AI and humans differently. Brain research beyond WEIRD contexts is required to better understand both human and AI presence.

The research reviewed spans two decades. During that time, AI has become more embedded in human society. Moreover, AI has changed drastically in that time, with more sophisticated algorithms and more human-like robots and avatars. Perhaps brain responses to AI 20 years ago were very different from responses today. In addition, brain imaging practices, including data preprocessing and analyses and sample sizes, to name a few, have dramatically changed over that period of time as well. Yet the analysis above collapsed over this significant time period.

Impact on Real World

Although AIs are perceived as agents, they are not yet processed as human, accounting for behavioral differences in the treatment of the two categories of agents and moral differences unique to humans. Therefore, given that the perception of human beings is differentiated from the perception of AI, there remain dimensions of humanity that AI must achieve to be perceived as fully human. While this distinction continues to exist, social science and health fields must grapple with the challenge presented by AI as an excellent mimic of human behavior while not being human. How does one determine culpability for harm caused by AI, such as when a self-driving car kills its drivers or other road users? Is it ethical to use chatbots to create text used for films, journalism, or leisure reading? Answers to such questions require further research exploring human and AI presence.

FUTURE ISSUES

- Is the brain response to AIs changing as human beings gain more experience with them and they become more integrated into human life?
- What is the developmental impact of interacting with AI from early in life? Will the brain continue to distinguish humans and AI in the same way?
- Over time, will the distinction between humans and AI fade in the brain as humans continue to evolve?

- How does the distinction in brain processing between humans and AI influence moral and legal support for the treatment and use of AI?
- What would brain imaging data reveal if researchers considered AI as the topic of interest in human–AI comparisons, treating the human response as the baseline instead?

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Amodio DM, Frith CD. 2006. Meeting of minds: the medial frontal cortex and social cognition. *Nat. Rev. Neurosci.* 4:268–77

Anders S, Heussen Y, Sprenger A, Haynes JD, Ethofer T. 2015. Social gating of sensory information during ongoing communication. *NeuroImage* 104:189–98

Assaf M, Kahn I, Pearlson GD, Johnson MR, Yeshurun Y, Calhoun VD, Hendler T. 2009. Brain activity dissociates mentalization from motivation during an interpersonal competitive game. *Brain Imaging Behav.* 3:24–37

Basile M, Lemasson A, Blois-Heulin C. 2009. Social and emotional values of sounds influence human (*Homo sapiens*) and non-human primate (*Cercopithecus campbelli*) auditory laterality. *PLOS ONE* 4:e6295

Baumgartner T, Knoch D, Hotz P, Eisenegger C, Fehr E. 2011. Dorsolateral and ventromedial prefrontal cortex orchestrate normative choice. *Nat. Neurosci.* 14:1468–74

Baylis GC, Rolls ET, Leonard CM. 1987. Functional subdivisions of the temporal lobe neocortex. *J. Neurosci.* 7:330–42

Bender EM, Gebru T, McMillan-Major A, Shmitchell S. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 610–23. New York: ACM

Botvinick MM, Braver TS, Barch DM, Carter CS, Cohen JD. 2001. Conflict monitoring and cognitive control. *Psychol. Rev.* 108:624–52

Bruce C, Desimone R, Gross CG. 1981. Visual properties of neurons in a polysensory area in superior temporal sulcus of the macaque. *J. Neurophys.* 46:369–84

Carter EJ, Hodgins JK, Rakison DH. 2011. Exploring the neural correlates of goal-directed action and intention understanding. *NeuroImage* 54:1634–42

Carter RM, Bowling DL, Reeck C, Huettel SA. 2012. A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science* 337:109–11

Chaminade T, Da Fonseca D, Rosset D, Cheng G, Deruelle C. 2015. Atypical modulation of hypothalamic activity by social context in ASD. *Res. Autism Spectr. Dis.* 10:41–50

Chaminade T, Hodgins J, Kawato M. 2007. Anthropomorphism influences perception of computer-animated characters' actions. *Soc. Cogn. Affect. Neurosci.* 2:206–16

Chaminade T, Rosset D, Da Fonseca D, Nazarian B, Lutcher E, et al. 2012. How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Front. Hum. Neurosci.* 6:103

Cheetham M, Suter P, Jäncke L. 2011. The human likeness dimension of the "uncanny valley hypothesis": behavioral and functional MRI findings. *Front. Hum. Neurosci.* 5:126

Chen X, Hackett PD, DeMarco AC, Feng C, Stair S, Haroon E. et al. 2016. Effects of oxytocin and vasopressin on the neural response to unreciprocated cooperation within brain regions involved in stress and anxiety in men and women. *Brain Imaging Behav.* 10:581–93

Coricelli G, Nagel R. 2009. Neural correlates of depth of strategic reasoning in medial prefrontal cortex. *PNAS* 106:9163–68

Craig AD. 2009. How do you feel—now? The anterior insula and human awareness. *Nat. Rev. Neurosci.* 10:59–70

Critchley HD, Garfinkel SN. 2017. Interoception and emotion. *Curr. Opin. Psychol.* 17:7–14

De Kort YA, Ijsselsteijn WA, Poels K. 2007. *Digital games as social presence technology: development of the Social Presence in Gaming Questionnaire (SPGQ)*. Paper presented at PRESENCE 2007, 10th Annual International Workshop on Presence, Barcelona, Oct. 25–27

De Quervain DJF, Fischbacher U, Treyer V, Schellhammer M, Schnyder U, Buck A, et al. 2004. The neural basis of altruistic punishment. *Science* 305:1254–58

Delgado MR, Schotter A, Ozbay EY, Phelps EA. 2008. Understanding overbidding: using the neural circuitry of reward to design economic auctions. *Science* 321:1849–52

Descartes R. 1637. *Discours de la methode pour bien conduire sa raison, & chercher la verité dans les sciences: plus la dioptrique, les eurons, et la geometrie, qui sont des essais de cete methode*. Leyde, Neth.: De l'imprimerie de I. Marie

Desimone R, Albright TD, Gross CG, Bruce C. 1984. Stimulus-selective properties of inferior temporal neurons in the macaque. *J. Neurosci.* 4:2051–62

Desmet C, Deschrijver E, Brass M. 2014. How social is error observation? The neural mechanisms underlying the observation of human and machine errors. *Soc. Cogn. Affect. Neurosci.* 9:427–35

Di Cesare G, Fasano F, Errante A, Marchi M, Rizzolatti G. 2016. Understanding the internal states of others by listening to action verbs. *Neuropsychologia* 89:172–79

Everts HG, Koolhaas JM. 1997. Lateral septal vasopressin in rats: role in social and object recognition? *Brain Res.* 760:1–7

Fareri DS, Hackett K, Tepfer LJ, Kelly V, Henninger N, Reeck C, et al. 2022. Age-related differences in ventral striatal and default mode network function during reciprocated trust. *NeuroImage* 256:119267

Fiske ST, Taylor SE. 2013. *Social Cognition: From Brains to Culture*. London: SAGE

Frith U, Frith C. 2001. The biological basis of social interaction. *Curr. Dir. Psychol. Sci.* 10:151–55

Gallagher HL, Frith CD. 2003. Functional imaging of "theory of mind." *Trends Cogn. Sci.* 7:77–83

Gallagher HL, Jack AI, Roepstorff A, Frith CD. 2002. Imaging the intentional stance in a competitive game. *NeuroImage* 16:814–21

Ganesh S, Van Schie HT, De Lange FP, Thompson E, Wigboldus DH. 2012. How the human brain goes virtual: Distinct cortical regions of the person-processing network are involved in self-identification with virtual agents. *Cereb. Cortex* 22:1577–85

Gobbini MI, Gentili C, Ricciardi E, Bellucci C, Salvini P, et al. 2011. Distinct neural systems involved in agency and animacy detection. *J. Cogn. Neurosci.* 23:1911–20

Goffman E. 1959. *The Presentation of Self in Everyday Life*. Garden City, NY: Doubleday

Gross CG, Rocha-Miranda CD, Bender DB. 1972. Visual properties of neurons in inferotemporal cortex of the Macaque. *J. Neurophys.* 35:96–111

Hare B, Tomasello M. 2005. Human-like social skills in dogs? *Trends Cogn. Sci.* 9:439–44

Harris LT. 2017. *Invisible Mind: Flexible Social Cognition and Dehumanization*. Cambridge, MA: MIT Press

Harris LT, Fiske ST. 2011. Dehumanized perception: a psychological means to facilitate atrocities, torture, and genocide? *Z. Psychol./J. Psychol.* 219:175–81

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293:2425–30

Henrich J, Heine S, Norenzayan A. 2010. The weirdest people in the world? *Behav. Brain Sci.* 33:61–83

Hogenhuis A, Hortensius R. 2022. Domain-specific and domain-general neural network engagement during human–robot interactions. *Eur. J. Neurosci.* 56:5902–16

Husted JR, McKenna FS. 1966. The use of rats as discriminative stimuli. *J. Exp. Anal. Behav.* 9:677–79

Ikeda T, Hirata M, Kasaki M, Alimardani M, Matsushit K, et al. 2017. Subthalamic nucleus detects unnatural android movement. *Sci. Rep.* 7:17851

James W. 1890. *The Principles of Psychology*. New York: H. Holt & Co.

Kanwisher N, McDermott J, Chun MM. 1997. The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17:4302–11

Kätsyri J, De Gelder B, De Borst AW. 2020. Amygdala responds to direct gaze in real but not in computer-generated faces. *NeuroImage* 204:116216

Kätsyri J, Hari R, Ravaja N, Nummenmaa L. 2013. The opponent matters: elevated fMRI reward responses to winning against a human versus a computer opponent during interactive video game playing. *Cereb. Cortex* 23:2829–39

Keupp S, Titchener R, Bugnyar T, Mussweiler T, Fischer J. 2019. Competition is crucial for social comparison processes in long-tailed macaques. *Biol. Lett.* 15:20180784

Klarer M, Weber-Stadlbauer U, Arnold M, Langhans W, Meyer U. 2019. Abdominal vagal deafferentation alters affective behaviors in rats. *J. Affect. Disord.* 252:404–12

Koban L, Gianaros PJ, Kober H, Wager TD. 2021. The self in context: brain systems linking mental and physical health. *Nat. Rev. Neurosci.* 22:309–22

Koban L, Pichon S, Vuilleumier P. 2014. Responses of medial and ventrolateral prefrontal cortex to interpersonal conflict for resources. *Soc. Cogn. Affect. Neurosci.* 9:561–69

Krach S, Blümel I, Marjoram D, Lataster T, Krabbendam L, et al. 2009. Are women better mindreaders? Sex differences in neural correlates of mentalizing detected with functional MRI. *BMC Neurosci*. 10:9

Krach S, Hegel F, Wrede B, Sagerer G, Binkofski F, Kircher T. 2008. Can machines think? Interaction and perspective taking with robots investigated via fMRI. *PLOS ONE* 3:e2597

Lee VK, Harris LT. 2014. Sticking with the nice guy: Trait warmth information impairs learning and modulates person perception brain network activity. *Cogn. Affect. Behav. Neurosci.* 14:1420–37

Li J, Xiao E, Houser D, Montague PR. 2009. Neural responses to sanction threats in two-party economic exchange. *PNAS* 106:16835–40

Lipina TV, Roder JC. 2013. Co-learning facilitates memory in mice: a new avenue in social neuroscience. *Neuropharmacology* 64:283–93

Lukas M, Toth I, Veenema AH, Neumann ID. 2013. Oxytocin mediates rodent social memory within the lateral septum and the medial amygdala depending on the relevance of the social stimulus: male juvenile versus female adult conspecifics. *Psychoneuroendocrinology* 38:916–26

Markus HR, Kitayama S. 1991. Culture and the self: implications for cognition, emotion, and motivation. *Psychol. Rev.* 98:224–53

Mars RB, Neubert FX, Noonan MP, Sallet J, Toni I, Rushworth MF. 2012. On the relationship between the "default mode network" and the "social brain." *Front. Hum. Neurosci.* 6:189

Matveeva TM, Pisansky MT, Young A, Miller RF, Gewirtz JC. 2019. Sociality deficits in serine racemase knockout mice. *Brain Behav*. 9:e01383

McCabe K, Houser D, Ryan L, Smith V, Trouard T. 2001. A functional imaging study of cooperation in two-person reciprocal exchange. *PNAS* 98:11832–35

McCarthy G, Puce A, Gore JC, Allison T. 1997. Face-specific processing in the human fusiform gyrus. *J. Cogn. Neurosci.* 9:605–10

McDonald KR, Pearson JM, Huettel SA. 2020. Dorsolateral and dorsomedial prefrontal cortex track distinct properties of dynamic social behavior. *Soc. Cogn. Affect. Neurosci.* 15:383–93

McNeill WH. 1997. *Keeping Together in Time*. Cambridge, MA: Harvard Univ. Press

Monfardini E, Reynaud AJ, Prado J, Meunier M. 2017. Social modulation of cognition: lessons from rhesus macaques relevant to education. *Neurosci. Biobehav. Rev.* 82:45–57

Özdem C, Wiese E, Wykowska A, Müller H, Brass M, Van Overwalle F. 2017. Believing androids—fMRI activation in the right temporo-parietal junction is modulated by ascribing intentions to non-human agents. *Soc. Neurosci.* 12:582–93

Parr LA, Hecht E, Barks SK, Preuss TM, Votaw JR. 2009. Face processing in the chimpanzee brain. *Curr. Biol.* 19:50–53

Perrett DI, Rolls ET, Caan W. 1982. Visual neurones responsive to faces in the monkey temporal cortex. *Exp. Brain Res*. 47:329–42

Petrulis A. 2009. Neural mechanisms of individual and sexual recognition in Syrian hamsters (*Mesocricetus auratus*). *Behav. Brain Res.* 200:260–67

Phan KL, Sripada CS, Angstadt M, McCabe K. 2010. Reputation for reciprocity engages the brain reward center. *PNAS* 107:13099–104

Pinsk MA, DeSimone K, Moore T, Gross CG, Kastner S. 2005. Representations of faces and body parts in macaque temporal cortex: a functional MRI study. *PNAS* 102:6996–7001

Piva M, Zhang X, Noah JA, Chang SW, Hirsch J. 2017. Distributed neural activity patterns during human-to-human competition. *Front. Hum. Neurosci.* 11:571

Poldrack RA. 2011. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron* 72:692–97

Powers KG, Ma X-M, Eipper BA, Mains RE. 2021. Cell-type specific knockout of peptidylglycine α-amidating monooxygenase reveals specific behavioral roles in excitatory forebrain neurons and cardiomyocytes. *Genes Brain Behav*. 20:e12699

Rangel A, Camerer C, Montague PR. 2008. A framework for studying the neurobiology of value-based decision making. *Nat. Rev. Neurosci.* 9:545–56

Rao RP, Von Heimendahl M, Bahr V, Brecht M. 2019. Neuronal responses to conspecifics in the ventral CA1. *Cell Rep*. 27:3460–72

Rilling JK, Gutman DA, Zeh TR, Pagnoni G, Berns GS, Kilts CD. 2002. A neural basis for social cooperation. *Neuron* 35:395–405

Rilling JK, Sanfey AG. 2011. The neuroscience of social decision-making. *Annu. Rev. Psychol.* 62:23–48

Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, Cohen JD. 2004a. Opposing BOLD responses to reciprocated and unreciprocated altruism in putative reward pathways. *NeuroReport* 15:2539–43

Rilling JK, Sanfey AG, Aronson JA, Nystrom LE, Cohen JD. 2004b. The neural correlates of theory of mind within interpersonal interactions. *NeuroImage* 22:1694–703

Rogers TB, Kuiper NA, Kirker WS. 1977. Self-reference and the encoding of personal information. *J. Pers. Soc. Psychol.* 35:677–88

Rosenthal-von der Pütten AM, Krämer NC, Maderwald S, Brand M, Grabenhorst F. 2019. Neural mechanisms for accepting and rejecting artificial social partners in the uncanny valley. *J. Neurosci.* 39:6555–70

Sanfey AG, Rilling JK, Aronson JA, Nystrom LE, Cohen JD. 2003. The neural basis of economic decision-making in the ultimatum game. *Science* 300:1755–58

Saygin AP, Chaminade T, Ishiguro H, Driver J, Frith C. 2012. The thing that should not be: predictive coding and the uncanny valley in perceiving human and humanoid robot actions. *Soc. Cogn. Affect. Neurosci.* 7:413–22

Schindler S, Kruse O, Stark R, Kissler J. 2019. Attributed social context and emotional content recruit frontal and limbic brain regions during virtual feedback processing. *Cogn. Affect. Behav. Neurosci.* 19:239–52

Schwarz N, Clore GL. 1983. Mood, misattribution, and judgments of well-being: informative and directive functions of affective states. *J. Pers. Soc. Psychol.* 45:513–23

Singer T, Kiebel SJ, Winston JS, Dolan RJ, Frith CD. 2004. Brain responses to the acquired moral status of faces. *Neuron* 41:653–62

Spitzer M, Fischbacher U, Herrnberger B, Grön G, Fehr E. 2007. The neural signature of social norm compliance. *Neuron* 56:185–96

Sun D, Chan CC, Hu Y, Wang Z, Lee TM. 2015. Neural correlates of outcome processing post dishonest choice: an fMRI and ERP study. *Neuropsychologia* 68:148–57

Tanaka K, Saito HA, Fukada Y, Moriya M. 1991. Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *J. Neurophys.* 66:170–89

Tang S, Harris L. 2015. Construing a transgression as a moral or a value violation impacts other versus self-dehumanisation. *Rev. Int. Psychol. Soc.* 28:95–123

Terwiesch C. 2023. *Would Chat GPT3 get a Wharton MBA? A prediction based on its performance in the operations management course*. Work. Pap., Mack Inst. Innov. Manag., Wharton Sch., Univ. Pa., Philadelphia

Thor DH, Holloway WR. 1982. Social memory of the male laboratory rat. *J. Comp. Phys. Psychol.* 96:1000–6

Tso IF, Rutherford S, Fang Y, Angstadt M, Taylor SF. 2018. The "social brain" is highly sensitive to the mere presence of social information: an automated meta-analysis and an independent study. *PLOS ONE* 13:e0196503

Vaitonyte J, Alimardani M, Louwerse MM. 2023. Scoping review of the neural evidence on the uncanny valley. *Comput. Hum. Behav. Rep.* 9:100263

Van den Bos W, McClure SM, Harris LT, Fiske ST, Cohen JD. 2007. Dissociating affective evaluation and social cognitive processes in the ventral medial prefrontal cortex. *Cogn. Affect. Behav. Neurosci.* 7:337–46

Van Dijk E, De Dreu CK. 2021. Experimental games and social decision making. *Annu. Rev. Psychol.* 72:415–38

Van Overwalle F. 2009. Social cognition and the brain: a meta-analysis. *Hum. Brain Mapp*. 30:829–58

Van't Wout M, Kahn RS, Sanfey AG, Aleman A. 2005. Repetitive transcranial magnetic stimulation over the right dorsolateral prefrontal cortex affects strategic decision-making. *NeuroReport* 16:1849–52

Wang Y, Quadflieg S. 2015. In our own image? Emotional and neural processing differences when observing human–human versus human–robot interactions. *Soc. Cogn. Affect. Neurosci.* 10:515–24

Wen T, Hsieh S. 2015. Neuroimaging of the joint Simon effect with believed biological and non-biological co-actors. *Front. Hum. Neurosci.* 9:483

Wittgenstein L. 1993. *Philosophical Occasions 1912–1951*, ed. JC Klagge, A Nordmann. Indianapolis, IN: Hackett Publ. Co.

Xiong W, Gao X, He Z, Yu H, Liu H, Zhou X. 2020. Affective evaluation of others' altruistic decisions under risk and ambiguity. *NeuroImage* 218:116996

Yamane S, Kaji S, Kawano K. 1988. What facial features activate face neurons in the inferotemporal cortex of the monkey? *Exp. Brain Res.* 73:209–14

Zhang Y, Yu H, Yin Y, Zhou X. 2016. Intention modulates the effect of punishment threat in norm enforcement via the lateral orbitofrontal cortex. *J. Neurosci.* 36:9217–26