

# External and Internal Attribution in Human-Agent Interaction: Insights From Neuroscience and Virtual Reality

Nina Lauharatanahirun<sup>1,2</sup> , Andrea Stevenson Won<sup>3</sup> , and Angel Hsing-Chi Hwang<sup>4</sup> 

1 Department of Biomedical Engineering, Pennsylvania State University, University Park, Pennsylvania, USA


2 Department of Biobehavioral Health, Pennsylvania State University, University Park, Pennsylvania, USA

3 Department of Communication, Cornell University, Ithaca, New York, USA

4 Ann S. Bowers College of Computing and Information Science, Cornell University, Ithaca, New York, USA

## Abstract

Agents are designed in the image of humans, both internally and externally. The internal systems of agents imitate the human brain, both at the levels of hardware (i.e., neuromorphic computing) and software (i.e., neural networks). Furthermore, the external appearance and behaviors of agents are designed by people and based on human data. Sometimes, these humanlike qualities of agents are purposely selected to increase their social influence over human users, and sometimes the human factors that influence perceptions of agents are hidden. Inspired by Blascovich's "threshold of social influence" (Blascovich et al., 2002), a model designed to explain the effects of different methods of anthropomorphizing embodied agents in virtual environments, we propose a novel framework for understanding how humans' attributions of human qualities to agents affects their social influence in human-agent interaction. The External and Internal Attributions model of social influence (EIA) builds on previous work on agent-avatars in immersive virtual reality and provides a framework to link previous social science theories to neuroscience. EIA connects external and internal attributions of agents to two brain networks related to social influence: the external perception system, and the mentalizing system. Focusing human-agent interaction research along each of the attributional

**CONTACT** Nina Lauharatanahirun  • [nina.lauhara@psu.edu](mailto:nina.lauhara@psu.edu) • Pennsylvania State University • 531 Chemical and Biomedical Engineering Building • University Park, PA 16802

ISSN 2638-602X (print)/ISSN 2638-6038 (online)  
[www.hmcjournal.com](http://www.hmcjournal.com)



Copyright 2024 Authors. Published under a Creative Commons Attribution 4.0 International (CC BY-NC-ND 4.0) license.

dimensions of the EIA model, or at the functional integration of the two, may lead to a better understanding of the thresholds of social influence necessary for optimal human-agent interaction.

**Keywords:** human-AI, human-agent, neuroscience, virtual reality, social influence

## Introduction

Communicating and interacting with nonhuman agents is becoming increasingly prevalent. In this paper, we define agents as computer programs designed to take actions and/or have specific goals. Such agents can range from virtual assistant technologies to fully autonomous robots. While the technological capability and sophistication of artificially intelligent systems continues to advance, our understanding of how humans process interactions with artificial agents is incomplete. Recently, it has even been suggested that the field of human-robot interaction is approaching a “social robotics winter,” referencing the mismatch between the promise of social robots and the outcome of failed human-robot interactions (Henschel et al., 2020). One source of this mismatch between unrealistic human expectations and social robotics reality comes from attempts to leverage human social reflexes to enhance trust and liking toward agents. However, such interactions can be problematic. Unrealistic expectations and incorrect grounding of human-agent interactions may set humans up for unsuccessful, disappointing, or disingenuous interactions with agents. In such cases, people may be reluctant to adopt or interact with these agents in the future. Thus, it becomes paramount to understand human expectations and perceptions of agent systems with the goal of managing such beliefs in pursuit of more authentic and realistic interactions with these technologies.

We integrate selected research from human-machine communication, human-computer interaction, human-robot interaction, psychology, virtual reality, and social cognitive neuroscience to inform a conceptual framework of humans’ perceptions of agents. We propose a novel adaptation of a key model for human-agent interactions in virtual reality—Blascovich’s Model of Social Influence (Blascovich et al., 2002). We build on this model to define two dimensions of agent characteristics as perceived by humans. Our proposed dimensions are (1) *external* attributions: the tendency to ascribe *anthropomorphic embodiment*, humanlike appearance and/or behavior, to nonhuman agents; and (2) *internal* attributions: the tendency to ascribe agentic humanlike internal states (e.g., mental states, motivations, intentions, and autonomy) to nonhuman agents. We explain how these two dimensions map onto two dissociable neural processing systems—the external perception system and the mentalizing system—that serve as the basis for social cognition and behavior. Finally, we review relevant human-agent and human-computer interaction theories and empirical support for these dimensions. Our aim is for this integrated framework to provide a useful scaffold for research on *understanding* and *predicting* human perceptions of agents, with the broader goal of facilitating transparent and authentic human-agent interactions.

Below, we will first discuss Blascovich’s Model of Social Influence by agent-avatars in virtual reality (Blascovich et al., 2002). In a selective review of the neuroscience literature,

we will relate human-computer interaction theory broadly and Blascovich's model specifically to these two pathways through which our brains process social information. We aim to contribute a better understanding of the neural basis of these social perceptions. We hope that by using neuroscience as a basis for understanding the pathways by which human users become socially influenced by nonhuman agents, will lead to more authentic and more useful social interactions with agents in the future.

## Blascovich's Model of Social Influence

In 2002, Blascovich and colleagues (2002) published a key paper on the experimental potential of agent-avatars (virtual representations that could look and behave like people but were controlled by a computer system; Fox et al., 2015). Specifically, they described how agent-avatars in immersive virtual environments could be provided with human-like appearances and behaviors with the goal of using such agent-avatars for experiments in social psychology. This paper introduced two intersecting dimensions: (1) *behavioral realism* (humanoid appearance and behavior) and (2) *social presence* (whether an entity is believed to be controlled by another person, or by a computer program) as part of a framework that explains under what circumstances such embodied agents (human-appearing social actors controlled by a computer) would be socially influential.

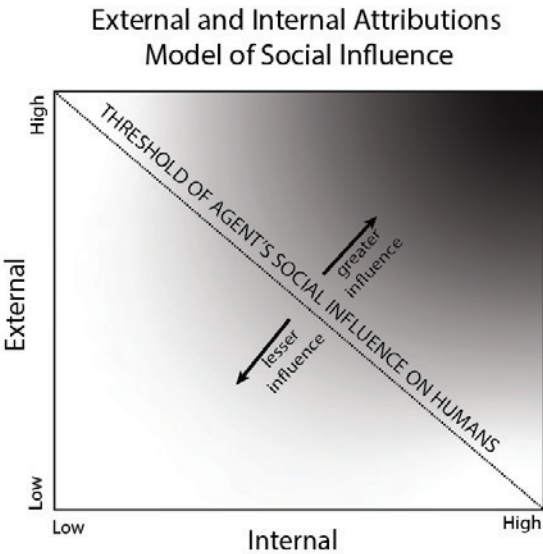
While this framework was proposed as a way to justify immersive virtual reality as a tool for social psychology, the authors made several propositions relevant to artificial agents more broadly considered. First, they proposed that the level of "behavioral realism" exhibited by a virtual human, which included both avatar appearance and behavior (speech, gestures, etc.) could influence human users socially *even if they were aware that the virtual human was an agent* (i.e., controlled by a computer rather than human). Second, they recognized that the influence of human agency was still important—that an agent that a participant believed was controlled by a human (rather than a computer) would be socially influential even if its level of behavioral realism was low. While this model was specific to the field of embodied agents in virtual reality, it can be usefully applied to a much broader context of social agents. The conceptual framework of the intersection between behavioral realism (which we will characterize as *external attributions*) and social presence (which we will expand to *internal attributions*) can be adapted to guide experimental work on identifying the features that create more authentic human-agent interactions. This model allows us to conceptually understand not only group-level effects, but also how individuals may differ in how they experience social influence. Figure 1 shows these relationships, below.

Blascovich et al.'s (2002) model of social influence has been highly influential in work on embodied agents. Considerable work has examined the effects of anthropomorphic external cues, generally finding that greater anthropomorphism leads to greater trust (De Visser et al., 2016) although a meta-analysis from 2007 found that overall effects of anthropomorphism from embodied agents were small (Yee et al., 2007) and more recent analyses have found mixed effects of different aspects of anthropomorphism (for example, appearance versus behavior) on measures of social presence (Oh et al., 2018). The attribution of agency has also commonly been manipulated. A meta-analysis by Fox et al. (2015, p. 1) found support for the importance of internal attributions of agency specifically, identifying an interaction effect such that "studies conducted on a desktop that used objective

measures showed a stronger effect for agency than those that were conducted on a desktop but used subjective measures.” However, a more recent meta-analysis (Felnhofer et al., 2023, p. 1) found that “while deliberate social responses like social presence and evaluation depend on perceived agency, automatic behaviors do not.”

Here, we make an important distinction. In this paper, we are not discussing agent-mediated interactions between humans in which an intelligent agent mediates or otherwise serves as an assistant to human communication (Hancock et al., 2020; Hohenstein & Jung, 2018). Instead, we are examining “stand-alone” agents which present as entities with which individual humans can engage 1:1 (Nass & Moon, 2000; Nass et al., 1994).

**FIGURE 1** An adapted version of the threshold model of social influence in virtual environments, as applied to agents. *External* attribution replaces “behavioral realism” to indicate that the agent is behaving and/or appearing in a human fashion. *Internal* attribution replaces “social presence” to indicate the extent to which the human user attributes internal states, especially intentional agency, to the agent’s actions.



The tendency to anthropomorphize objects and nonhuman agents has been reliably demonstrated in contexts ranging from geometric shapes (Heider & Simmel, 1944) to computer-animated blobs (Guthrie, 1995; Morewedge et al., 2007). Anthropomorphism is broadly defined as the “tendency to imbue the real or imagined behavior of nonhuman agents with humanlike characteristics, motivation, intentions, or emotions” (Epley et al., 2007, p. 864). According to psychological research, humans are motivated to engage in anthropomorphic behavior based on two primary factors (Epley et al., 2007). First, humans are driven to effectively manage uncertainty and need to predict and understand their interaction partners for effective communication and interaction. Second, humans are driven

to form social connections with other humans, a desire that may extend to nonhuman artificial agents. These motivations may lead humans to seek relevant cues in human-agent interaction.

In the context of human-agent interaction, Gambino and colleagues (2020) suggest that people may be consciously assessing the “humanness” of agents and then behaving accordingly. This aligns with work by Sundar (1998), proposing that human users’ information processing styles interact with agent characteristics following Petty and Cacioppo’s Elaboration Likelihood Model (ELM; Petty et al., 1986). The ELM suggests individuals can take either a central or peripheral route for decision-making. If people take the central path, they adopt logical, systematic approaches to processing information; with the peripheral route, people make “fast and frugal” decisions based on heuristic cues. For example, Sundar (1998) posited that highly engaged users would take the central route and evaluate computer-mediated content more systematically, considering the source of online news and the credibility of the author(s) who wrote the news article. On the other hand, casual viewers who take the peripheral route would be more affected by tangential factors, such as visual layout and design of the content. Following on this, Sundar and Nass (2001) proposed that varying the perceived source of presented news (visible, technological, audience, or self) also changed ratings of the story itself. More recently, Sundar and colleagues (2015; 2020) adopted a dual-path framework to conceptualize users’ perception of machine agency. In this work, Sundar uses the Theory of Interactive Media Effects (TIME; Sundar et al., 2015) framework. The TIME model suggests that users evaluate applications of emerging technology either through an action route or a cue route: Through the action route, users determine how to interact with an application based on its actual functions, such as system performance, technical capability, and interacting behaviors demonstrated on an user interface; through the cue route, users evaluate novel applications based on peripheral features (e.g., appearances and content presentation) that are not necessarily related to their technical performance and capabilities per se. Based on the TIME framework, Sundar and Kim propose that the affordances of a given system can lead users to deploy different cognitive heuristics (Sundar & Kim, 2019; Lee, 2018). These include machine heuristics and social heuristics. The former refer to users’ common expectations and even stereotypical impressions for mechanical/computational systems, such as they could perform complex computation tasks accurately and efficiently. By contrast, social heuristics point to humans’ tendencies to treat nonhuman subjects as social entities, such as interacting with them through natural language and verbal communication. This assertion implies that there may be multiple pathways to influence how users make attributions about agents. If users rely on cues and are prompted to use a more social heuristic rather than a “machine heuristic,” for example, through external, anthropomorphic embodiment cues, then this could affect which associated brain networks become active. Alternatively, the “action route” could lead users to actively assess an agent’s source attribution and internal states during interaction, which could also lead to less “mindless” assessments of machine agency. However, while these external and internal factors can be manipulated independently, their effects on attribution are likely intertwined; for example, a person interacting with a very humanlike agent may not be able to avoid attributing internal states to that agent.

## The External-Internal Attribution Model and Neuroscience

Paralleling the dimensions of external and internal attributions in our proposed EIA model, research from social cognitive neuroscience has identified brain networks that are involved in the representation and processing of *external* and *internal* information of others during social interactions. Multiple brain networks are involved in processing external features, such as appearance and movement, of human or nonhuman others. While the social robotics and human neuroscience literatures use slightly different terminology, the networks are analogous. For instance, the action-observation network described in social robotics research (Henschel et al., 2020) is analogous to what is called the mirror neuron network in cognitive neuroscience (Sperduti et al., 2014; Spunt & Lieberman, 2014; Spunt et al., 2015). Similarly, the person-perception brain network (Henschel et al., 2020) from social robotics is equivalent to the face-body perception network (Downing et al., 2001; Kanwisher et al., 1997) from neuroscience. We will refer to brain networks that support the human brain's processing of embodiment cues such as perceptions of movement and appearance as the *external perception system*. Another brain system that is equally important in guiding social influence during social interactions is the *mentalizing system* which is also referred to in the literature as theory of mind. The mentalizing system is involved in processing the *internal* states of another (Alcalá-López et al., 2019; Frith & Frith, 2006; Sperduti et al., 2014; Spunt & Lieberman, 2014; Spunt et al., 2015). Social neuroscience has primarily been focused on understanding the brain systems that support social information processing between humans, but we propose that this line of research may complement existing research in the human-agent interaction field. Below, we operationalize *external* attributions as anthropomorphic embodiment, and *internal* attributions as focusing on intentional agency, where agency refers to an agent's ability to have internal states guiding decision-making and potentially autonomous actions. We integrate both behavioral and neuroscience findings and discuss how our proposed dimensions relate to these brain networks as they are currently understood.

### External Attributions of Anthropomorphic Embodiment

External attribution cues have been much leveraged by designers of human-agent interactions, and these methods of anthropomorphically embodying agents are closely related to Blascovich et al.'s (2002) concept of "behavioral realism," in which an entity's physical form appears and/or behaves like a human being. Embodiment in agents is most clearly illustrated by robots, as the robots necessarily are physically embodied (Breazeal, 2003; Duffy, 2003). However, embodiment can also be a component of "embodied agents"; for example, virtual representations of humans that exist only digitally, such as in virtual or augmented reality applications, or even in AI assistants such as Siri and Alexa which can evoke anthropomorphic embodiment concepts such as gender or age (i.e., the voices used by these devices imply the source of an adult female).

In our framework, we operationalize these external attribution cues as a continuum in which human-like characteristics or cues (e.g., speech, cadence, tone of voice, physical appearance, movement, or other behaviors or features) are applied to nonhuman agents. For example, providing an agent with a female voice, giving it the body of an older adult,



or having it raise “eyebrows” as a means of nonverbal expression are all ways to embody nonhuman agents by leveraging human appearance or human behavioral cues. This definition is in line with current research showing that altering artificial agents to appear more human-like in terms of their appearance and behavior can lead to smoother human-agent communication and enhanced engagement (Waytz et al., 2010).

Embodiment features can trigger and enhance anthropomorphism providing more channels for communication (Deng et al., 2019) leading to enhanced human-agent communication and performance (Wainer et al., 2007). Previous research that examined the effect of the physical appearance and behavior of agents on users’ perception and behaviors (von der Pütten et al., 2010; De Visser et al., 2016) supports the effectiveness of anthropomorphic embodiment on evoking social responses in humans. For instance, it is well documented that the fusiform face area/fusiform gyrus (FFA/FFG) responds selectively to faces (Kanwisher et al., 1997) and that the extrastriate body area (EBA) responds selectively to bodies and body parts (Downing et al., 2001), which are key to the fundamental detection and recognition of other people. This recruitment of the FFA represents a fundamental low-level process that is often integrated with higher order cognitive and emotional attributions/appraisals. In the social robotics literature, activation of such brain areas as the FFA/FFG is referred to as the person perception network (PPN; Henschel et al., 2020). Importantly, evidence from brain imaging studies indicates that humans activate the PPN when observing robots express humanlike emotions (Hortensius & Cross, 2018) and when observing other humans interact with robots (Wang & Quadflieg, 2015), although this is moderated by what Blascovich’s model would identify as the factors leading to social influence. Specifically, the right FFA and bilateral posterior superior temporal sulcus showed higher levels of activation in response to human-human interactions relative to human-robot interaction (Wang & Quadflieg, 2015). Moreover, another study found that FFA/FFG activity corresponded with subjective ratings of human likeness ratings, where decreasing activity was observed for artificial agents (Rosenthal-von der Pütten et al., 2019).

Evidence from social robotics research has shown that changing robot appearance (e.g., giving robots faces and human shapes) and robot motor behavior (e.g., hand gestures when communicating) can activate similar brain areas typically recruited during human-human social interactions (Chaminade et al., 2010; Cross et al., 2012), brain regions known as the *mirror neuron network* or the *action-observation network*. While the promise and broad application of the mirror neuron network to higher levels of social cognitive function may have been overstated, its involvement in linking perceptions and actions of others has been replicated in many empirical studies (for reviews see Bonini et al., 2022; Heyes & Catmur, 2022). Perception of agents is an active and automatic process that involves identifying and extracting features of an interaction partner (e.g., speech, appearance, gestures) from the influx of sensory information to help the human observer understand *what* the agent is and *what its function* might be (often indicated through motor movements). This process of identification activates the mirror neuron network in the brain, which includes but is not restricted to the dorsal and ventral regions of the premotor cortex, anterior inferior parietal lobule, anterior temporal cortex, and the temporal parietal junction/superior temporal sulcus (Rizzolatti & Craighero, 2004; Spunt & Lieberman, 2014). When we observe others, this network of brain areas communicates sensory information about another’s motor actions into a representation of a goal-directed action (Iacoboni et al., 2005;

Zacks et al., 2001). For instance, the superior temporal sulcus has been linked to the perception of faces (Haxby et al., 2000; Puce et al., 1998), biological motion (Grossman et al., 2000; Herrington et al., 2011), understanding other's actions (Vander Wyk et al., 2009), and voice perception (Deen et al., 2015). Thus, the human brain synthesizes incoming sensory information regarding the anthropomorphically embodied features of an agent, which in turn can lead to the formation of perceptions that guide our attributional inferences.

Mirror neurons are brain cells distributed across motor, sensory, and motivational brain areas that have been proposed to play a role in social cognition, supporting social interaction (Bonini et al., 2022). Mirror neurons were first discovered in the ventral premotor region F5 of the macaque (di Pellegrino et al., 1992; Gallese et al., 1996; Rizzolatti et al., 1996) and have been identified in a number of species, including humans (Molenberghs et al., 2012; Mukamel et al., 2010). Activation of mirror neurons occurs both when a person performs an action and when a similar action is performed by another individual, thus providing a neural basis for linking perceptions (observations) with motor movements. This key *mirroring* feature of neurons is thought to subserve people's ability to learn new behaviors through imitation and understand the actions of others (for review, see Bonini et al., 2022; Heyes & Catmur, 2022). In nonhuman animal studies, mirror neurons were thought to exist primarily in the ventral premotor cortex and inferior parietal lobule (e.g., di Pellegrino et al., 1992; Rizzolatti et al., 1996); however, human experimental studies have shown that this mirroring feature allowing the mapping of other's actions onto self-related brain regions is not limited to these two brain structures alone. Perhaps one of the most influential studies regarding mirror neurons within the human brain comes from Mukamel and colleagues (2010) who recorded electrophysiological signals from neurons in the medial frontal and temporal cortices while human participants both executed and observed grasping motor movements. The results from their study provide evidence that human neurons in the medial frontal lobe (supplementary motor area), hippocampus, parahippocampal gyrus, and entorhinal cortex fired in response to both performing and observing grasping motor actions. These results not only provide direct evidence of the existence of mirror neurons in the human brain, but indicate that the mirror neuron property exists in brain structures beyond what was previously observed in animal studies.

With regard to social interactions, being able to recognize and perceive the actions of others is key for planning or predicting how we should behave in future situations. While this is a core social cognitive function and the initial starting point for better understanding how humans perceive agents during social interactions, we acknowledge that social interactions are complex and involve the simultaneous processing of multisensory information in response to another's expressions, behaviors, movements, and intentions. In the last decade, social robotics researchers have leveraged neuroimaging technologies to advance our understanding of the neurocognitive mechanisms subserving social behavior during human-robot interactions (for reviews see Cross et al., 2019; Henschel et al., 2020; Hortensius & Cross, 2018). In these studies, the mirror neuron network is referred to as the action-observation network (AON) which includes areas of the parietal, premotor, and middle temporal cortices. Research studies show that the action-observation network is active not only when humans observed other humans, but also when robots grasp and handle objects (Cross et al., 2012; Cross et al., 2019; Henschel et al., 2020). For instance, one study found that AON activation was stronger when human participants were observing



unfamiliar robotic movements (regardless of whether humans or robots performed the action; Cross et al., 2012). This result suggests that humans may engage the mirror neuron network during uncertain social interactions. As previous studies investigating the mirror neuron network suggest (Molenberghs et al., 2012; Mukamel et al., 2010), engagement of this system helps humans learn about their interaction partners. In sum, the action observation/mirror neuron network plays a reflexive and automatic role in understanding the actions of others (humans or artificial agents), and suggests that this system permits the connection of self to other through the simulation of other's actions at the motor level. Together with the face-body/person-perception network, anthropomorphic embodiment cues are processed by an *external perception system* in the brain that ultimately shapes the extent of social influence an agent can have based on whether the human observer's mind determines whether actors exhibit social or nonsocial features.

Neuroscience findings can also help address aspects of embodiment that may be problematic. According to the uncanny valley hypothesis (Mori, 1970; Mori et al., 2012), human perceptions of artificial agents are nonlinear such that likability increases with anthropomorphized agents but precipitously decreases if these agents are perceived to be too humanlike. This has been partially addressed in the neuroscience literature, in that previous work has aimed to uncover the neurocognitive mechanisms associated with human responses to unknown artificial agents. For example, one study identified that nonlinear responses in the ventromedial prefrontal cortex (vmPFC) similarly aligned with the subjective likability and human likeness ratings of artificial agents (Rosenthal-von der Pütten et al., 2019). Responses of the vmPFC scaled with human ratings such that higher ratings of likability and human likeness were associated with greater vmPFC activity, and this association decreased for highly humanlike agents (Rosenthal-von der Pütten et al., 2019). The study also found that amygdala responses predicted when human participants would reject gifts from artificial agents, which is in line with other reports implicating the amygdala's involvement in the processing of social information (Phelps & LeDoux, 2005) such as face processing (Adolphs, 2009) and anthropomorphism (Heberlein & Adolphs, 2004). The role of the amygdala in anthropomorphic perceptions and behavior is not new: Researchers examining patients with basolateral amygdala lesions found that they exhibited decreased anthropomorphic behavior for inanimate stimuli relative to healthy controls (Waytz et al., 2019).

These findings elucidate the neural infrastructure that enables anthropomorphic behavior in guiding humans to process signals and information as social or nonsocial.

## Internal Attributions and Intentional Agency

Early communication theories have suggested that when humans interacted with agents, including text-based interactions with a computer, people were unable to avoid applying human-human social scripts to their interactions (Reeves & Nass, 1996). Conversational agents such as chatbots, virtual agents, and social robots were designed based on the influential "computers-as-social-actors" or CASA theory, which states that humans interact with computers as if they are human (Nass & Moon, 2000; Nass et al., 1994). In these studies, even though human users were consciously aware that computers were not sentient agents, they attributed intentional agency to the devices rather than, for example, to the human programmers of the devices (Nass & Moon, 2000; Nass et al., 1994). However, more recent

work suggests that as people gain experience with computers and incorporate agents into other aspects of their life, they may no longer attribute agency in the same way (Gambino et al., 2020; Heyselaar, 2023).

We operationalize *internal attribution cues* as a continuum in which attributions of humanlike mental states, motivations, intentions, and autonomy are applied to nonhuman agents. One example is the extent to which an artificial agent is perceived to have internal states indicating that it has internal agency; that it is “alive” and “in control” of its own expressions and behaviors. When humans interact with others (humans or artificial agents), we attempt to understand who we are interacting with and will often make attributional inferences about another’s internal states (e.g., beliefs, values) to both explain and predict another’s actions (Frith & Frith, 2006). Even though machines, robots, and artificial agents lack a mind per se, they are programmed with existing policies for actions, movements, and expressions, and thus these internal attributions remain useful and relevant.

The internal attribution dimension in the proposed model maps onto an inferential social cognitive process that involves attributing mental states, intentions, and internal states known as “*mentalizing*” (Frith & Frith, 2006, p. 531). It has been argued that humans and primates alike have evolved to develop larger brain volumes (Dunbar, 1998) as well as specialized brain networks that support social cognition (Adolphs, 2009; Fareri & Delgado, 2014; Kliemann & Adolphs, 2018; Lockwood et al., 2020; Spunt et al., 2015). Being able to engage in social interactions involves a diverse suite of social cognitive abilities that range from low-level sensory processes such as recognizing faces (discussed above as a component of external attributions) to high-level cognitive functions such as making inferences about the intentions of others.

Neuroimaging evidence over the last decade suggests that a network of brain areas is recruited and reliably activated to support higher-level social cognitive processes such as mentalizing. The mentalizing brain network includes key brain regions such as the superior temporal sulcus (STS), temporal parietal junction (rTPJ, lTPJ), posterior cingulate cortex (PCC), and the ventromedial prefrontal cortex (vmPFC). Perhaps one of the most consistently reported brain areas subserving social cognition is the superior temporal sulcus (Deen et al., 2015; Pelphrey et al., 2004; Saxe et al., 2004; Yamada et al., 2022; Zilbovicius et al., 2006). The medial prefrontal cortex has been suggested to play a general role in representing social or emotionally relevant information about oneself (Frith & Frith, 2006; Northoff & Bermpohl, 2004) or another person (Saxe & Powell, 2006). Finally, the brain area that is most notably associated with theory of mind or mentalizing is the temporal parietal junction (TPJ). The TPJ is theorized to play a role in synthesizing lower-level processing streams into higher-order social-cognitive functions. Research has demonstrated that the anterior TPJ is recruited for regulating attentional processes and mentalizing in social situations (Krall et al., 2015; Saxe, 2006; Saxe & Powell, 2006; Van Overwalle, 2009). These neurobiological correlates are important for linking human brain processes with the human mind and, thus, behavior during social interactions.

Recently, likely due to the advancement of technology, researchers have started to examine whether social cognition and mentalizing of humans recruits similar neural circuitry when compared to nonhuman artificial agents. For instance, one study found that social cognitive brain areas such as the TPJ and mPFC selectively responded to humans only relative to humanoid robots (Chaminade et al., 2012). This finding suggests that while

there may be some similarity in how humans perceive appearance and motor features of humans and nonhuman agents, humans still distinguish between intentional agents and entities that may have humanlike internal states (e.g., desires, beliefs) guiding their behavior. Another line of evidence from social neuroscience research has used economic games to understand the neural bases of social interactions (Chang et al., 2023; Fareri et al., 2012; McCabe et al., 2001; Rilling et al., 2004). In these studies, humans engage in social exchange games with humans and computers. Behaviorally, studies have shown that humans entrust resources similarly to humans and agent partners (Schniter et al., 2020). However, future studies examining how the human brain processes these exchanges with agents relative to other people are needed to better understand the neural mechanisms that give rise to social cognition and perception within social interactions. Research in this area would increase our understanding of under what circumstances AI and other nonhuman entities may be perceived as intentional social beings with internal states.

One aspect of internal attribution that has been less explored, at least in quantitative social science, is the fact that most agents are designed and created by an organization (e.g., technology companies) or groups of people and, therefore, their creation cannot be attributed to a single person (Luria, 2020). For example, Apple's Siri voice agent has the modified voice of a human woman, but Siri's design is indebted to hundreds or perhaps thousands of researchers and designers, and the data that built it and refines its output arises from millions of individual human users (Hwang & Won, 2022).

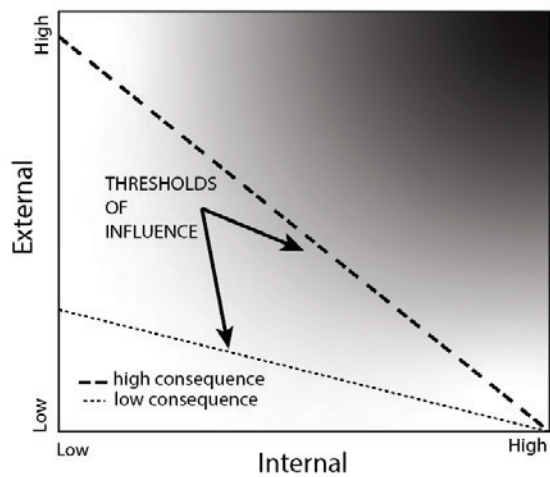
Addressing this gap, one school of researchers proposed that the perceived agency of an agent can be attributed to a single "source" (e.g., Apple), which can then be distributed to various entities through embodiment in different devices (e.g., Siri on your phone, on your tablet, etc.; Luria et al., 2019). This allows a single source of agency (Apple) to be deployed and become omnipresent across different Apple devices, and even re-embodied when a physical artifact is renewed or replaced (e.g., when one gets a new iPhone, and hears Siri's voice coming out of the new speaker). This again suggests that the users can conceive of agency as distinct from embodiment, and hints at a more accurate view of attribution, since most agents are the product of many, many human minds contributing to the overall goals of a business or other entity. This more complex view of the relationship between external attribution (embodiment in a given device) with an internal attribution (a central source of agency such as a company) hints to how users can associate internal states such as intentions with a corporate entity rather than an individual device.

## Perceptions During Social Interactions and the Importance of Social Context

Social interactions are complex. People have to represent their own intentions, beliefs, and values, but also must engage in perspective-taking to understand others' motives, beliefs, and values. Moreover, social interactions require human brains to integrate low-level sensory information that relies on external attributions (e.g., visual, auditory, somatosensory) with higher-level social cognitive processes requiring internal attribution, such as mental state reasoning. Understanding how the human brain integrates both low-level sensory features and higher-level social information for understanding others is not only an interesting area in its own right, but it is also an area ripe for interdisciplinary insights.

Blascovich and colleagues (2002) proposed two additional factors that could moderate the threshold of social influence and that are relevant to current communication theories. A reflexive response could be evoked by any agent, but a socially significant situation (for example, taking romantic advice from an agent) would have a higher bar of social influence. In addition, the value or meaning of the interaction to the human user was important. For trivial tasks, Blascovich and colleagues proposed that behavioral realism was *less* likely to be influential, while consequential tasks would retain the higher threshold of social influence. Figure 2 shows these proposed dual thresholds of social influence—a testable proposition that contrasts interestingly with other predictions that different cues will have different weight depending on context and on the importance of the situation to the human interactant.

**FIGURE 2** Is anthropomorphism more important to social influence in high-consequence situations? The original Theory of Social Influence proposed the answer was “yes,” as shown above, but other communication theories predict that low-consequence situations might lead people to rely even more on cues such as anthropomorphism.



Integrating research across internal and external attributions suggests that on the one hand, there may be similarity in how humans perceive appearance and motor features across humans and artificial agents (Chaminade et al., 2012; Frith, 2008; Johnson, 2003; Scholl & Tremoulet, 2000; Thompson et al., 2011). On the other hand, the results also indicate that humans distinguish between intentional agents that have internal states and agents that do not. We argue that, in addition to considering “surface” aspects of human characteristics (i.e., appearance, behavior), designers of artificial agents should also consider how humans perceive the “deeper” social goals and intentions of artificial agents. We believe that studying how people perceive agents within *social contexts* provides an ideal testbed to identify the intersection of external and internal attribution features that reliably recruit

brain systems involved in social cognitive processes (i.e., external perception system and mentalizing system).

Our existing methods for studying human agent interaction are often unidirectional and static to enable controlled testing of experimental manipulations. However, our perceptions are dynamic and continuously updated as we process and integrate incoming information during interactions with agents. Equally important is the fact that these human-agent interactions do not occur within a vacuum. We often interact with agents when other humans are present, and our perceptions may be moderated by how other humans perceive and respond to the agents involved in the interaction. We suggest that we can complement existing behavioral paradigms with neuroimaging and physiological measures to objectively measure how the human brain and mind responds to agents, how humans perform tasks with agents, and how they develop mutual understanding and social engagement over time.

## Next Steps

Further, we ask how understanding the roles that humans play in creating artificial agents might enhance the perception of *intentional agency attributed to humans* who design, build, and provide data to create artificial agents. Such an improved understanding will have at least two potentially useful effects. First, it will make more transparent the influence of the groups of people whose data, opinions, or technical skills inform the creation of AI agents. This will make discussions of bias in AI more intelligible and more salient. For example, many people are still not aware that conversational agents are built using specific datasets that over-represent some humans (people publishing in academic journals, people posting on the programming site Stack Overflow) and under-represent others (people without access to the internet; people who are not literate). While this will not necessarily increase trust in agents, it will allow people to calibrate their trust in these agents based on their real social knowledge of other humans' abilities and biases. We note again that the CASA paradigm described above found that people did not naturally make attributions to, for example, the programmer behind the computer agent. However, we are now living in different times. For instance, a recent replication of the original CASA study found that participants do not treat desktop computers as social actors (Heyselaar, 2023) highlighting the need to conduct new research studies with emergent technologies. Given people's increased experience with agents and the different cultural context in which human-agent interactions occur, it is now time to ask again whether providing more information about the humans and human organizations behind the agents can lead people to make such attributions. Below, we list some research questions that can shed light on whether such conscious reflection on the human element can predict, and improve, the outcomes of human-agent interaction.

## Suggested Research Questions

**RQ1.** When humans are interacting with a group of humans or a group of artificial agents, is intentional agency ascribed to the group as a singular unit? Are similar social cognitive brain networks recruited during interactions with a group of humans versus a group of artificial agents?



**RQ2a.** Does the combination of agency and embodiment mutually enhance activation of the social brain? or:

**RQ2b.** Do the multiple sources of human agency that contribute to artificial agents conflict with anthropomorphic cues, which are necessarily single?

**RQ3a.** Does the *type* of task (consequential and/or social, following Blascovich's proposed moderators of the threshold of social influence) moderate the degree to which mentalization is linked to social influence and/or task success?

**RQ3b.** Does the *type* of task (consequential and/or social, following Blascovich's proposed moderators of the threshold of social influence) moderate the degree to which anthropomorphic cues are linked to social influence and/or task success?

## Conclusion

The modernization and technological advancement occurring within our society necessitates a deeper understanding of how humans perceive agents during human-agent interactions, which may benefit from interdisciplinary perspectives. The broad goal of our proposed framework is to integrate research across disciplines to support the mechanistic understanding of human social cognition during social interactions. Specifically, the intersection of external and internal attributions as described in the EIA model may provide an accessible framework for understanding the social influence agents may have on humans. The framework also provides researchers across disciplines a guide to experimentally test which features activate human social cognitive processing (at the level of the brain or mind) when interacting with artificial agents. It may also help researchers gain insights regarding the conditions under which human perceptions may lead to unrealistic expectations and inaccurate predictions of an agent's actions. Considering social influence as a product of both external and internal attribution cues can also provide a framework for better understanding how neuroscience can be used to enhance our understanding of human-agent interaction and integrate it into more recent work from communication examining AI-mediated communication (Hancock et al., 2020). In turn, we believe this lens can lead to design recommendations for AI that are both more effective and truer to the actual AI ecosystem.

## Author Biographies

**Dr. Nina Lauharatanahirun** (PhD, Virginia Tech) is an Assistant Professor of Biomedical Engineering and Biobehavioral Health at Pennsylvania State University, and the director of the Decision Neuroscience Laboratory. The lab's work is focused on understanding the neurobehavioral mechanisms of social decision-making with the goal of leveraging theoretically grounded neurobehavioral signals for the design of algorithmic solutions that improves human-human and human-agent team decisions.

 <https://orcid.org/0000-0001-8229-1099>

**Dr. Andrea Stevenson Won** (PhD, Stanford University) is an Associate Professor of Communication at Cornell University, and the director of the Virtual Embodiment Lab. The lab's work examines tracking and transforming aspects of embodiment, including appearance and behavior, with a focus on virtual reality's clinical, collaborative, and educational capabilities.

 <https://orcid.org/0000-0001-5240-6166>

**Angel Hsing-Chi Hwang** (PhD, Cornell University) is a Post-Doctoral Associate at the Ann S. Bowers College of Computing and Information Science at Cornell University, whose work focuses on researching and designing human-AI interaction at large scales in various applied settings (e.g., Future of Work, mental health care ecosystem, and policy sandbox and prototyping).

 <https://orcid.org/0000-0002-0951-7845>

## References

- Adolphs, R. (2009). The social brain: Neural basis of social knowledge. *Annual Review of Psychology*, 60, 693–716. <https://doi.org/10.1146/annurev.psych.60.110707.163514>
- Alcalá-López, D., Vogeley, K., Binkofski, F., & Bzdok, D. (2019). Building blocks of social cognition: Mirror, mentalize, share?. *Cortex*, 118, 4–18. <https://doi.org/10.1016/j.cortex.2018.05.006>
- Blascovich, J., Loomis, J., Beall, A. C., Swinth, K. R., Hoyt, C. L., & Bailenson, J. N. (2002). Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, 13(2), 103–124. [https://doi.org/10.1207/S15327965PLI1302\\_01](https://doi.org/10.1207/S15327965PLI1302_01)
- Bonini, L., Rotunno, C., Arcuri, E., & Gallese, V. (2022). Mirror neurons 30 years later: Implications and applications. *Trends in cognitive sciences*, 26(9), 767–781.
- Breazeal, C. (2003). Toward sociable robots. *Robotics and Autonomous Systems*, 42(3–4), 167–175. [https://doi.org/10.1016/S0921-8890\(02\)00373-1](https://doi.org/10.1016/S0921-8890(02)00373-1)
- Chaminade, T., Rosset, D., Da Fonseca, D., Nazarian, B., Lutchter, E., Cheng, G., & Deruelle, C. (2012). How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. *Frontiers in Human Neuroscience*, 6, 103. <https://doi.org/10.3389/fnhum.2012.00103>
- Chaminade, T., Zecca, M., Blakemore, S. J., Takanishi, A., Frith, C. D., Micera, S., Dario, P., Rizzolatti, G., Gallese, V., & Umiltà, M. A. (2010). Brain response to a humanoid robot in areas implicated in the perception of human emotional gestures. *PLoS ONE*, 5(7), e11577. <https://doi.org/10.1371/journal.pone.0011577>
- Chang, L. A., Armaos, K., Warns, L., Ma de Sousa, A. Q., Paauwe, F., Scholz, C., & Engelmann, J. B. (2023). Mentalizing in an economic games context is associated with enhanced activation and connectivity in the left temporoparietal junction. *Social Cognitive and Affective Neuroscience*, 18(1), nsad023. <https://doi.org/10.1093/scan/nsad023>
- Cross, E. S., Hortensius, R., & Wykowska, A. (2019). From social brains to social robots: Applying neurocognitive insights to human–robot interaction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 374(1771), 20180024. <https://doi.org/10.1098/rstb.2018.0024>

- Cross, E. S., Liepelt, R., de C. Hamilton, A. F., Parkinson, J., Ramsey, R., Stadler, W., & Prinz, W. (2012). Robotic movement preferentially engages the action observation network. *Human Brain Mapping*, 33(9), 2238–2254. <https://doi.org/10.1002/hbm.21361>
- Deen, B., Koldewyn, K., Kanwisher, N., & Saxe, R. (2015). Functional organization of social perception and cognition in the superior temporal sulcus. *Cerebral Cortex*, 25(11), 4596–4609. <https://doi.org/10.1093/cercor/bhv111>
- Deng, E., Mutlu, B., & Mataric, M. J. (2019). Embodiment in socially interactive robots. *Foundations and Trends in Robotics*, 7(4), 251–356. <https://doi.org/10.1561/23000000056>
- De Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, 22(3), 331. <https://doi.org/10.1037/xap0000092>
- di Pellegrino, G., Fadiga, L., Fogassi, L., Gallese, V., & Rizzolatti, G. (1992). Understanding motor events: A neurophysiological study. *Experimental Brain Research*, 91(1), 176–180. <https://doi.org/10.1007/BF00230027>
- Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, 293(5539), 2470–2473. <https://doi.org/10.1126/science.1063414>
- Duffy, B. R. (2003). Anthropomorphism and the social robot. *Robotics and Autonomous Systems*, 42(3–4), 177–190. [https://doi.org/10.1016/S0921-8890\(02\)00374-3](https://doi.org/10.1016/S0921-8890(02)00374-3)
- Dunbar, R. I. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews*, 6(5), 178–190. [https://doi.org/10.1002/\(SICI\)1520-6505\(1998\)6:5%3C178::AID-EVAN5%3E3.0.CO;2-8](https://doi.org/10.1002/(SICI)1520-6505(1998)6:5%3C178::AID-EVAN5%3E3.0.CO;2-8)
- Epley, N., Waytz, A., & Cacioppo, J. T. (2007). On seeing human: A three-factor theory of anthropomorphism. *Psychological Review*, 114(4), 864–886. <https://doi.org/10.1037/0033-295X.114.4.864>
- Fareri, D. S., Chang, L. J., & Delgado, M. R. (2012). Effects of direct social experience on trust decisions and neural reward circuitry. *Frontiers in Neuroscience*, 6, 148. <https://doi.org/10.3389/fnins.2012.00148>
- Fareri, D. S., & Delgado, M. R. (2014). Social rewards and social networks in the human brain. *The Neuroscientist*, 20(4), 387–402. <https://doi.org/10.1177/1073858414521869>
- Felnhofer, A., Knaust, T., Weiss, L., Goinska, K., Mayer, A., & Kothgassner, O. D. (2023). A virtual character's agency affects social responses in immersive virtual reality: A systematic review and meta-analysis. *International Journal of Human-Computer Interaction*, 1–16. <https://doi.org/10.1080/10447318.2023.2209979>
- Fox, J., Ahn, S. J., Janssen, J. H., Yeykelis, L., Segovia, K. Y., & Bailenson, J. N. (2015). Avatars versus agents: A meta-analysis quantifying the effect of agency on social influence. *Human-Computer Interaction*, 30(5), 401–432. <https://doi.org/10.1080/07370024.2014.921494>
- Frith, C. D. (2008). Social cognition. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1499), 2033–2039. <https://doi.org/10.1098/rstb.2008.0005>
- Frith, C. D., & Frith, U. (2006). The neural basis of mentalizing. *Neuron*, 50(4), 531–534. <https://doi.org/10.1016/j.neuron.2006.05.001>
- Gallese, V., Fadiga, L., Fogassi, L., & Rizzolatti, G. (1996). Action recognition in the premotor cortex. *Brain*, 119(2), 593–609. <https://doi.org/10.1093/brain/119.2.593>

- Gambino, A., Fox, J., & Ratan, R. A. (2020). Building a stronger CASA: Extending the computers are social actors paradigm. *Human-Machine Communication*, 1, 71–85. <https://doi.org/10.30658/hmc.1.5>
- Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., & Blake, R. (2000). Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience*, 12(5), 711–720. <https://doi.org/10.1162/089892900562417>
- Guthrie, S. E. (1995). *Faces in the clouds: A new theory of religion*. Oxford University Press.
- Hancock, J. T., Naaman, M., & Levy, K. (2020). AI-mediated communication: Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1), 89–100. <https://doi.org/10.1093/jcmc/zmz022>
- Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, 4(6), 223–233. [https://doi.org/10.1016/s1364-6613\(00\)01482-0](https://doi.org/10.1016/s1364-6613(00)01482-0)
- Heberlein, A. S., & Adolphs, R. (2004). Impaired spontaneous anthropomorphizing despite intact perception and social knowledge. *Proceedings of the National Academy of Sciences*, 101(19), 7487–7491. <https://doi.org/10.1073/pnas.0308220101>
- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, 57(2), 243. <https://doi.org/10.2307/1416950>
- Henschel, A., Hortensius, R., & Cross, E. S. (2020). Social cognition in the age of human–robot interaction. *Trends in Neurosciences*, 43(6), 373–384. <https://doi.org/10.1016/j.tins.2020.03.013>
- Herrington, J. D., Nymberg, C., & Schultz, R. T. (2011). Biological motion task performance predicts superior temporal sulcus activity. *Brain and Cognition*, 77(3), 372–381. <https://doi.org/10.1016/j.bandc.2011.09.001>
- Heyes, C., & Catmur, C. (2022). What happened to mirror neurons? *Perspectives on Psychological Science*, 17(1), 153–168.
- Heyselaar, E. (2023). The CASA theory no longer applies to desktop computers. *Scientific Reports*, 13(1), 19693. <https://doi.org/10.1038/s41598-023-46527-9>
- Hohenstein, J., & Jung, M. (2018, April). AI-supported messaging: An investigation of human-human text conversation with AI support. In *Extended abstracts of the 2018 CHI conference on human factors in computing systems* (pp. 1–6). <https://doi.org/10.1145/3170427.3188487>
- Hortensius, R., & Cross, E. S. (2018). From automata to animate beings: The scope and limits of attributing socialness to artificial agents. *Annals of the New York Academy of Sciences*, 1426(1), 93–110. <https://doi.org/10.1111/nyas.13727>
- Hwang, A. H. C., & Won, A. S. (2022, April). AI in your mind: Counterbalancing perceived agency and experience in human-AI interaction. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1–10). <https://doi.org/10.1145/3491101.3519833>
- Iacoboni, M., Molnar-Szakacs, I., Gallese, V., Buccino, G., Mazziotta, J. C., & Rizzolatti, G. (2005). Grasping the intentions of others with one's own mirror neuron system. *PLoS Biology*, 3(3), e79. <https://doi.org/10.1371/journal.pbio.0030079>
- Johnson, S. C. (2003). Detecting agents. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1431), 549–559. <https://doi.org/10.1098/rstb.2002.1237>

- Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience*, 17(11), 4302–4311. <https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997>
- Kliemann, D., & Adolphs, R. (2018). The social neuroscience of mentalizing: Challenges and recommendations. *Current Opinion in Psychology*, 24, 1–6. <https://doi.org/10.1016/j.copsyc.2018.02.015>
- Krall, S. C., Rottschy, C., Oberwelland, E., Bzdok, D., Fox, P. T., Eickhoff, S. B., Fink, G. R., & Konrad, K. (2015). The role of the right temporoparietal junction in attention and social interaction as revealed by ALE meta-analysis. *Brain Structure and Function*, 220, 587–604. <https://doi.org/10.1007/s00429-014-0803-z>
- Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1), 205395171875668. <https://doi.org/10.1177/2053951718756684>
- Lockwood, P. L., Apps, M. A., & Chang, S. W. (2020). Is there a ‘social’ brain? Implementations and algorithms. *Trends in Cognitive Sciences*, 24(10), 802–813. <https://doi.org/10.1016/j.tics.2020.06.011>
- Luria, M. (2020). Mine, yours or Amazon’s?: Designing agent ownership and affiliation. *Companion Publication of the 2020 ACM Designing Interactive Systems Conference*, 537–542. <https://doi.org/10.1145/3393914.3395830>
- Luria, M., Reig, S., Tan, X. Z., Steinfeld, A., Forlizzi, J., & Zimmerman, J. (2019, June). Re-embodiment and co-embodiment: Exploration of social presence for robots and conversational agents. In *Proceedings of the 2019 on Designing Interactive Systems Conference* (pp. 633–644). <https://doi.org/10.1145/3322276.3322340>
- McCabe, K., Houser, D., Ryan, L., Smith, V., & Trouard, T. (2001). A functional imaging study of cooperation in two-person reciprocal exchange. *Proceedings of the National Academy of Sciences*, 98(20), 11832–11835. <https://doi.org/10.1073/pnas.211415698>
- Molenberghs, P., Cunnington, R., & Mattingley, J. B. (2012). Brain regions with mirror properties: A meta-analysis of 125 human fMRI studies. *Neuroscience & Biobehavioral Reviews*, 36(1), 341–349. <https://doi.org/10.1016/j.neubiorev.2011.07.004>
- Morewedge, C. K., Preston, J., & Wegner, D. M. (2007). Timescale bias in the attribution of mind. *Journal of Personality and Social Psychology*, 93(1), 1–11. <https://doi.org/10.1037/0022-3514.93.1.1>
- Mori, M. (1970) The uncanny valley. *Energy*, 7(4), 33–35.
- Mori, M., MacDorman, K., & Kageki, N. (2012). The uncanny valley [from the field]. *IEEE Robotics & Automation Magazine*, 19(2), 98–100. <https://doi.org/10.1109/MRA.2012.2192811>
- Mukamel, R., Ekstrom, A. D., Kaplan, J., Iacoboni, M., & Fried, I. (2010). Single-neuron responses in humans during execution and observation of actions. *Current Biology*, 20(8), 750–756. <https://doi.org/10.1016/j.cub.2010.02.045>
- Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. <https://doi.org/10.1111/0022-4537.00153>
- Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. *Conference Companion on Human Factors in Computing Systems*, 204. <https://doi.org/10.1145/259963.260288>
- Northoff, G., & Bermpohl, F. (2004). Cortical midline structures and the self. *Trends in Cognitive Sciences*, 8(3), 102–107. <https://doi.org/10.1016/j.tics.2004.01.004>



- Oh, C. S., Bailenson, J. N., & Welch, G. F. (2018). A systematic review of social presence: Definition, antecedents, and implications. *Frontiers in Robotics and AI*, 5, 409295.
- Pelphrey, K. A., Morris, J. P., & McCarthy, G. (2004). Grasping the intentions of others: The perceived intentionality of an action influences activity in the superior temporal sulcus during social perception. *Journal of Cognitive Neuroscience*, 16(10), 1706–1716. <https://doi.org/10.1162/0898929042947900>
- Petty, R. E., Cacioppo, J. T., Petty, R. E., & Cacioppo, J. T. (1986). *The elaboration likelihood model of persuasion* (pp. 1–24). Springer New York. [https://doi.org/10.1016/S0065-2601\(08\)60214-2](https://doi.org/10.1016/S0065-2601(08)60214-2)
- Phelps, E. A., & LeDoux, J. E. (2005). Contributions of the amygdala to emotion processing: From animal models to human behavior. *Neuron*, 48(2), 175–187. <https://doi.org/10.1016/j.neuron.2005.09.025>
- Puce, A., Allison, T., Bentin, S., Gore, J. C., & McCarthy, G. (1998). Temporal cortex activation in humans viewing eye and mouth movements. *Journal of Neuroscience*, 18(6), 2188–2199. <https://doi.org/10.1523/JNEUROSCI.18-06-02188.1998>
- Reeves, B., & Nass, C. I. (1996). *The media equation: How people treat computers, television, and new media like real people and places*. CSLI Publications; Cambridge University Press.
- Rilling, J. K., Sanfey, A. G., Aronson, J. A., Nystrom, L. E., & Cohen, J. D. (2004). The neural correlates of theory of mind within interpersonal interactions. *Neuroimage*, 22(4), 1694–1703. <https://doi.org/10.1016/j.neuroimage.2004.04.015>
- Rizzolatti, G., & Craighero, L. (2004). The mirror-neuron system. *Annual Review of Neuroscience*, 27, 169–192. <https://doi.org/10.1146/annurev.neuro.27.070203.144230>
- Rizzolatti, G., Fadiga, L., Gallese, V., & Fogassi, L. (1996). Premotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3(2), 131–141. [https://doi.org/10.1016/0926-6410\(95\)00038-0](https://doi.org/10.1016/0926-6410(95)00038-0)
- Rosenthal-von der Pütten, A. M., Krämer, N. C., Maderwald, S., Brand, M., & Grabenhorst, F. (2019). Neural mechanisms for accepting and rejecting artificial social partners in the uncanny valley. *The Journal of Neuroscience*, 39(33), 6555–6570. <https://doi.org/10.1523/JNEUROSCI.2956-18.2019>
- Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*, 16(2), 235–239. <https://doi.org/10.1016/j.conb.2006.03.001>
- Saxe, R., & Powell, L. J. (2006). It's the thought that counts: Specific brain regions for one component of theory of mind. *Psychological Science*, 17(8), 692–699. <https://doi.org/10.1111/j.1467-9280.2006.01768.x>
- Saxe, R., Xiao, D. K., Kovacs, G., Perrett, D. I., & Kanwisher, N. (2004). A region of right posterior superior temporal sulcus responds to observed intentional actions. *Neuropsychologia*, 42(11), 1435–1446. <https://doi.org/10.1016/j.neuropsychologia.2004.04.015>
- Schniter, E., Shields, T. W., & Sznycer, D. (2020). Trust in humans and robots: Economically similar but emotionally different. *Journal of Economic Psychology*, 78, 102253. <https://doi.org/10.1016/j.joep.2020.102253>
- Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, 4(8), 299–309. [https://doi.org/10.1016/S1364-6613\(00\)01506](https://doi.org/10.1016/S1364-6613(00)01506)

- Sperduti, M., Guionnet, S., Fossati, P., & Nadel, J. (2014). Mirror neuron system and mentalizing system connect during online social interaction. *Cognitive Processing*, 15(3), 307–316. <https://doi.org/10.1007/s10339-014-0600-x>
- Spunt, R. P., & Lieberman, M. D. (2014). Automaticity, control, and the social brain. In J. W. Sherman, B. Gawronski, & Y. Trope (Eds.), *Dual-process theories of the social mind* (pp. 279–296). The Guilford Press.
- Spunt, R. P., Meyer, M. L., & Lieberman, M. D. (2015). The default mode of human brain function primes the intentional stance. *Journal of Cognitive Neuroscience*, 27(6), 1116–1124. [https://doi.org/10.1162/jocn\\_a\\_00785](https://doi.org/10.1162/jocn_a_00785)
- Sundar, S. S. (1998). Effect of source attribution on perception of online news stories. *Journalism & Mass Communication Quarterly*, 75(1), 55–68. <https://doi.org/10.1177/107769909807500108>
- Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human–AI interaction (HAII). *Journal of Computer-Mediated Communication*, 25(1), 74–88. <https://doi.org/10.1093/jcmc/zmz026>
- Sundar, S. S., Jia, H., Waddell, T. F., & Huang, Y. (2015). Toward a theory of interactive media effects (TIME): Four models for explaining how interface features affect user psychology. In S. S. Sundar (Ed.), *The handbook of the psychology of communication technology* (1st ed., pp. 47–86). Wiley. <https://doi.org/10.1002/9781118426456.ch3>
- Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–9. <https://doi.org/10.1145/3290605.3300768>
- Sundar, S. S., & Nass, C. (2001). Conceptualizing sources in online news. *Journal of Communication* 51, 1 (2001), 52–72. <https://doi.org/10.1111/j.1460-2466.2001.tb02872.x>
- Thompson, J. C., Trafton, J. G., & McKnight, P. (2011). The perception of humanness from the movements of synthetic agents. *Perception*, 40(6), 695–704. <https://doi.org/10.1068/p6900>
- Vander Wyk, B. C., Hudac, C. M., Carter, E. J., Sobel, D. M., & Pelphrey, K. A. (2009). Action understanding in the superior temporal sulcus region. *Psychological Science*, 20(6), 771–777. <https://doi.org/10.1111/j.1467-9280.2009.02359.x>
- Van Overwalle, F. (2009). Social cognition and the brain: A meta-analysis. *Human Brain Mapping*, 30(3), 829–858. <https://doi.org/10.1002/hbm.20547>
- von der Pütten, A. M., Krämer, N. C., Gratch, J., & Kang, S. H. (2010). “It doesn’t matter what you are!” explaining social effects of agents and avatars. *Computers in Human Behavior*, 26(6), 1641–1650. <https://doi.org/10.1016/j.chb.2010.06.012>
- Wainer, J., Feil-Seifer, D. J., Shell, D. A., & Mataric, M. J. (2007). Embodiment and human-robot interaction: A task-based perspective. *RO-MAN 2007 The 16th IEEE International Symposium on Robot and Human Interactive Communication*, 872–877. <https://doi.org/10.1109/ROMAN.2007.4415207>
- Wang, Y., & Quadflieg, S. (2015). In our own image? Emotional and neural processing differences when observing human–human vs human–robot interactions. *Social Cognitive and Affective Neuroscience*, 10(11), 1515–1524. <https://doi.org/10.1093/scan/nsv043>
- Waytz, A., Cacioppo, J. T., Hurlmann, R., Castelli, F., Adolphs, R., & Paul, L. K. (2019). Anthropomorphizing without social cues requires the basolateral amygdala. *Journal of cognitive neuroscience*, 31(4), 482–496. [https://doi.org/10.1162/jocn\\_a\\_01365](https://doi.org/10.1162/jocn_a_01365)

- Waytz, A., Gray, K., Epley, N., & Wegner, D. M. (2010). Causes and consequences of mind perception. *Trends in Cognitive Sciences*, 14(8), 383–388. <https://doi.org/10.1016/j.tics.2010.05.006>
- Yamada, Y., Sueyoshi, K., Yokoi, Y., Inagawa, T., Hirabayashi, N., Oi, H., Shirama, A., & Sumiyoshi, T. (2022). Transcranial direct current stimulation on the left superior temporal sulcus improves social cognition in schizophrenia: An open-label study. *Frontiers in Psychiatry*, 13. <https://doi.org/10.3389/fpsyt.2022.862814>
- Yee, N., Bailenson, J. N., & Rickertsen, K. (2007, April). A meta-analysis of the impact of the inclusion and realism of human-like faces on user experiences in interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 1–10). <https://doi.org/10.1145/1240624.1240626>
- Zacks, J. M., Braver, T. S., Sheridan, M. A., Donaldson, D. I., Snyder, A. Z., Ollinger, J. M., Buckner, R. L., & Raichle, M. E. (2001). Human brain activity time-locked to perceptual event boundaries. *Nature Neuroscience*, 4(6), 651–655. <https://doi.org/10.1038/88486>
- Zilbovicius, M., Meresse, I., Chabane, N., Brunelle, F., Samson, Y., & Boddaert, N. (2006). Autism, the superior temporal sulcus and social perception. *Trends in Neurosciences*, 29(7), 359–366. <https://doi.org/10.1016/j.tins.2006.06.004>