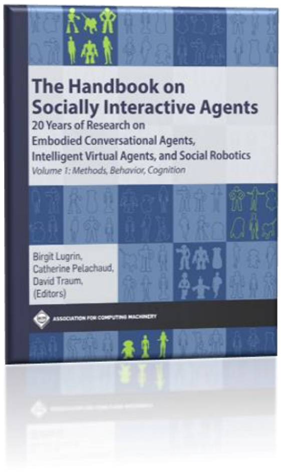
Theory of Mind and Joint Attention

Jairo Perez-Osorio, Eva Wiese and Agnieszka Wykowska



The Handbook on Socially Interactive Agents

20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics

Volume 1: Methods, Behavior, Cognition

Birgit Lugrin, Catherine Pelachaud, David Traum, (Editors)

ASSOCIATION FOR COMPUTING MACHINERY

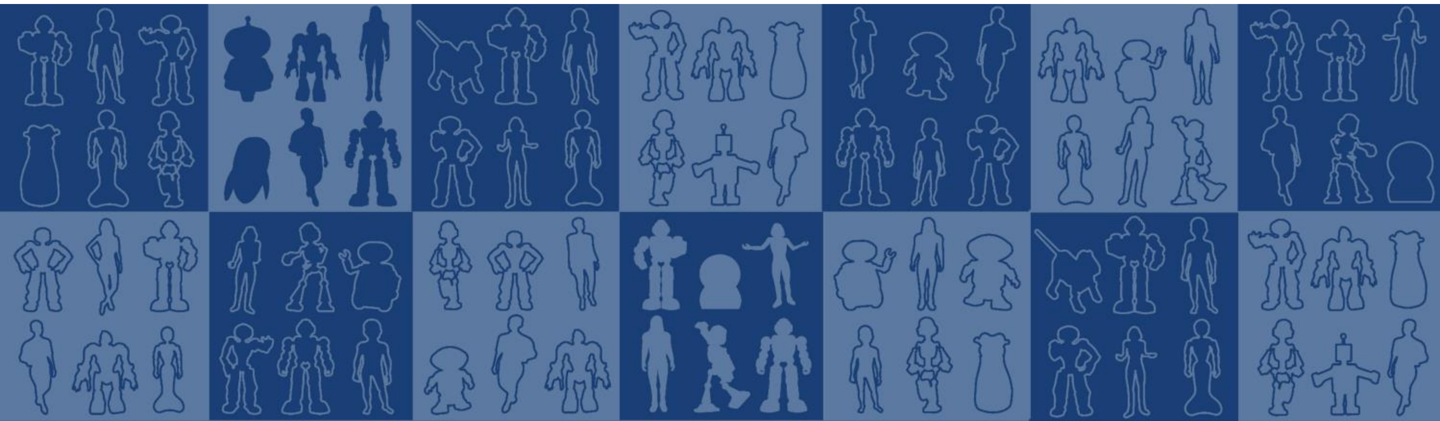
Citation information:

Perez-Osorio, J., Wiese, E., and Wykowska, A. (2021). Theory of Mind and Joint Attention. In B. Lugrin, C. Pelachaud, D. Traum (Eds.), *Handbook on Socially Interactive Agents – 20 Years of Research on Embodied Conversational Agents, Intelligent Virtual Agents, and Social Robotics*, Volume 1: Methods, Behavior, Cognition (pp. 313-347). ACM.

DOI of the final chapter: 10.1145/3477322.3477332

DOI of volume 1 of the handbook: 10.1145/3477322

Correspondence concerning this chapter should be addressed to Agnieszka Wykowska (agnieszka.wykowska@iit.it)



Chapter 9

Theory of Mind and Joint Attention

Jairo Perez-Osorio[a], Eva Wiese[b], and Agnieszka Wykowska[a]

[a]Istituto Italiano di Tecnologia, Genoa, Italy
[b]George Mason University, Fairfax, VA, USA

1. Social Cognitive Neuroscience and SIA

Increasing technological progress in the last decades has facilitated the development of artificial agents. Social robots, virtual agents, and smart assistants have been introduced slowly but firmly into our daily lives. From entertainment to education, from healthcare to the conquest of the solar system, artificial agents are becoming increasingly essential for the human social landscape. However, before they can become fully integrated into our lives, it is important to consider how to measure the impact interactions with these agents might have on human cognition, and how to evaluate whether the behavior of artificial agents has desired effects on everyday life.

In order to attain an overarching comprehension of the dynamics of social interactions with artificial agents, research in human–agent interaction (HAI) would immensely benefit from the methods and approaches used in social cognitive neuroscience (SCN). This discipline focuses on studying the intricate interplay between social and neurophysiological aspects when the brain is engaged in social-cognitive processing during interactions with others, using objective behavioral and neuro-physiological measures in carefully designed and controlled experiments. SCN is characterized by hypothesis-driven experiments that manipulate experimental variables targeting specific cognitive processes, and aims at interpreting behavioral responses and their neural correlates in the context of theoretical models of social cognition. Adopting methods of SCN for the study of social interactions with artificial agents offers multiple advantages. First, in addition to other methods commonly used in HAI studies, SCN methods allow examining cognitive mechanisms that are not necessarily explicit or available to introspection. Specifically, although HAI methods, such as subjective ratings, surveys, and questionnaires allow the assessment of attitudes, experiences, or perceptions during interactions with artificial agents, using only those methods does not cover the more implicit processing of information in social interactions (e.g., gaze, posture, voice pitch, turn-taking). Furthermore, the normative interpretation and understanding of these signals is typically learned implicitly through experience, which can make it difficult for participants to describe it verbally when explicit assessment techniques are used. For instance, could you tell how long it takes you to experience mutual gaze or the duration of a handshake with a stranger as uncomfortable? You probably have not thought about this before, or measured it empirically, which makes it difficult to give a precise answer. Thus, although self-reported measures are easily obtained and suitable for the assessment of some aspects of social cognition, they may be insufficient to evaluate all the different layers of social interactions with natural and artificial agents.

Another challenge in the assessment of social cognition is to implement paradigms that are capable of capturing the dynamic and proactive nature of social interactions, involving predictive processes that rely on inferences regarding others' intentions and mental states. SCN has revealed that when interacting with the world, human brains constantly select, process, and compare sensory inputs to previous representations of knowledge and experiences to build accurate representations of the world in a dynamic cycle that updates priors and adjusts predictions of future events [Friston 2005]. Given that these processes also unfold when interacting with the social world, any paradigm or method that does not appropriately elicit or allow for the dynamics of this process to unfold may not accurately assess the underlying social-cognitive mechanisms [Schilbach et al. 2013]. For instance, mutual gaze during conversations varies depending on the context and the interlocutor and people are usually unaware of those variations: a cognitively demanding conversation elicits less mutual eye contact than chitchatting, and people prefer to look longer in the eyes of familiar than unfamiliar people [Beattie and Ellis 2017]. Coming back to the original question of how long mutual gaze can last before starting to feel uncomfortable, this example shows that it is impossible to define an ideal time window as such behavior is dynamic and linked to social context and personal experience. However, we can examine how humans engage in mutual gaze using well-designed experimental protocols that use objective behavioral and neurophysiological measures and that do not restrict the natural dynamics of social interactions.

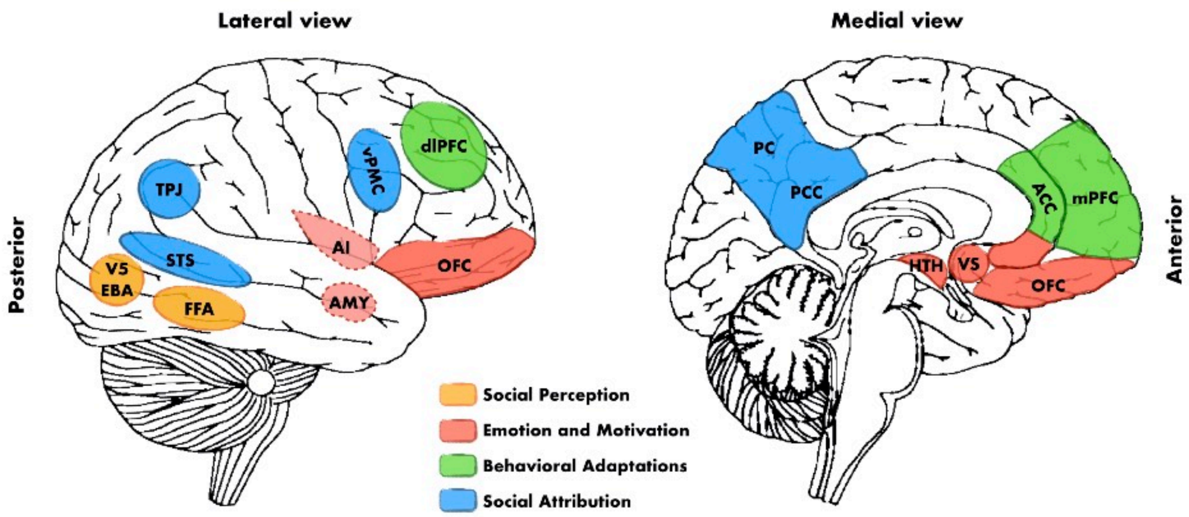


Figure 1. Areas of the brain associated with processing social information. Classification of the brain areas linked with social processing and divided in four cognitive processes: (1) The perception of basic social stimuli, such as biological motions (V5), part of the body (extra-striate body area, EBA), and faces (fusiform face area, FFA); (2) emotional and motivational appraisal, where the amygdala (AMY), the anterior insula (AI), the subgenual and perigenual anterior cingulate cortex (ACC), together with the orbitofrontal cortex (OFC) are closely linked with subcortical structures as the ventral striatum (VS) and the hypothalamus (HTH); (3) Emotional and motivational appraisal areas work closely with regions such as the dorsolateral and the medial prefrontal cortex (dlPFC, mPFC) and the ACC in goal-directed, adaptive behaviors, and the categorization processes. Finally, (4) areas associated with social attribution, like the ventral premotor cortex (vPMC), the superior temporal sulcus (STS), the AI, the posterior cingulate cortex (PCC), and the precuneus (PC) participate in more automatic, bottom-up inferences of other people's mental states; whereas structures like the mPFC and the temporoparietal junction (TPJ) are involved in more cognitive theory of mind skills. Adapted from Billeke and Aboitiz [2013].

Another way in which SCN informs HAI is by elucidating the brain structures involved in social cognition using diverse methods such as electroencephalography, functional magnetic resonance imaging, and functional near-infrared spectroscopy. The areas and networks involved in the processing of social stimuli have been collectively termed "the social brain," which includes structures like the medial prefrontal cortex, the temporoparietal junction, the superior temporal sulcus, and the fusiform area, inferior temporal sulcus, among others (see Figure 1).

These networks show characteristic patterns of activation during the processing of social signals commonly used to communicate interest, highlight the relevance of events or objects in the environment, or coordinate interactions, such as biological motion, facial expressions, and eye and head movements. Furthermore, those signals are tightly linked to higher cognitive processes like recognizing others' feelings and internal states, identifying others' intentions, or deciding whether they are friend or foe. Understanding the neurobiological basis of behavior is crucial for cognitive, developmental, clinical, comparative, and social psychology, as well as philosophy and evolutionary anthropology [Singer 2012]. In this context, examining the engagement of social brain areas in interactions with artificial agents seems like a natural step to follow for HAI as well." Using a combination of subjective/explicit measures like self-report and questionnaires, and objective measures like metrics related to performance, behavior, psychophysiology, and neuroimaging, HAI is able to have a more comprehensive view of behavioral and brain mechanisms involved in social interactions with human and non-human agents. The present chapter provides an overview from the perspective of SCN regarding theory of mind (ToM) and joint attention (JA) as crucial mechanisms of social cognition and discusses how these mechanisms have been investigated in social interaction with artificial agents. In the final sections, the chapter reviews computational models of ToM and JA in social robots (SRs) and intelligent virtual agents (IVAs) and discusses the current challenges and future directions.

2. Theory of Mind and Joint Attention—Crucial Mechanisms of Social Cognition

2.1 Theory of Mind

Imagine that as you are walking on the street someone is coming from the opposite direction toward you. At some point, and before passing each other, the person all of a sudden puts the palm of her/his hand on their forehead, stops walking, turns around, and goes back to where she/he came from. How would you explain this behavior? Probably, you would guess that she/he might have forgotten something and decided to go back. And you might be right. Importantly, most of the explanations you would choose to explain the observed behavior would refer to mental states, such as thoughts, preferences, intentions, or emotions. This is based on the ability to perceive and understand that others have beliefs, desires, goals, and knowledge different from your own and that others' behavior is driven by their internal representations of the world. Social interaction heavily requires awareness of our counterpart's knowledge of the world [Frith et al. 1991, Baron-Cohen 1995]. The ability of referring to others' mental states in explaining their behavior has been termed mentalizing [Frith and Frith 2006] or using a ToM [Baron-Cohen 1997]. ToM is the basis for a wide range of social processes such as competitive and cooperative joint actions, language, action execution, imagination, and even humor. The strategy of referring to mental states to predict others' behavior has also been called adopting the intentional stance [Dennett 1971, 1987]. Please note, however, that we distinguish the concept of "intentional stance" from the concept of "theory of mind." The first is related to the general strategy that one adopts when explaining the behavior of another agent, based on the assumptions regarding the agents' rationality and capacity of having mental states. The latter, on the other hand, is the active process of inferring a particular mental state in a particular context. One can infer a wrong mental state based on observation of the other's behavior (and thus, fail the ToM test, see below), but still adopt the intentional stance to the agent in general.

Now, imagine that you are asking a humanoid robot for directions at the counter of a train station. You are interested in knowing about restaurants nearby, and the robot manages to answer your questions. After an effortless dialogue, the robot says: "I will give you a map," turns around, and looks for something inside a shelf. All of a sudden, the robot stops every movement, turns back toward you, and says: "Good morning, how can I help you?" How would you explain this behavior to someone else? You would probably describe it in terms of the robot's presumed internal states and say something like: "The robot forgot that I was there waiting for the map." Most people facing a situation like this would also use mentalistic terms to describe the behavior of the robot. In fact, it is very intuitive for humans to attribute human intentions, preferences, capacities, and emotions to non-human agents and interact as if those agents would actually have a mind. However, it might be that attributions or more cognitive states, such as thoughts or intentions, are more likely than attributions of more phenomenal or affective states, such as pain or happiness [Huebner 2010]. The tendency to attribute mental states to non-human agents is a part of the process called anthropomorphism. When an entity is anthropomorphized, its behavior, as well as inferences drawn from it, is interpreted in human-centric terms [Epley 2008]. If paper constantly jams in a printer, for instance, an anthropomorphic interpretation would say that “the printer refuses to work,” or even say “that stubborn printer.” A technical or more detailed explanation of the same behavior could feel artificial and be more complex, preventing effective communication. Humans have extensive experience understanding other humans’ minds, which is probably why mentalistic explanations for the behavior of non-human entities often seem intuitive. Note, however, that typically people do not believe that cars or other complex non-human systems have internal mental states, but that they often explain behavior and natural phenomena within their “mentalistic comfort zone.” In summary, humans have the tendency to interact with non-human agents as if they had mental states. This ability facilitates prediction, understanding, and interaction with social counterparts.

2.1.1 Developing a Theory of Mind

From very early stages of development, we learn to read others’ minds. This capability is critical for cognitive development: it provides foundations for language acquisition, allows differentiating between self and others, and is crucial for social interactions. Interestingly, representing the mental states of others occurs effortlessly, automatically, and unconsciously, which explains why others’ behavior is often intuitively explained in mentalistic terms although it is questionable (as in the case of mindless artificial systems, such as a printer or car). In the past century, several disciplines have extensively studied the effects and origins of human ToM. Initial cognitive approaches by Heider [1958], for instance, suggested that people have a general understanding of others’ ideas and actions in particular situations—a “commonsense or folk psychology” that helps them deal effectively with social situations. This ability is firmly anchored on the assumption that beliefs and intentions play an active role in others’ behavior, together with subjective experiences and perceptions of the environment. Such inferences transcend the observed behavior and prove to be useful in predicting and understanding others’ actions. Subsequently, Premack and Woodruff [1978] defined ToM as the ability to reason about other’s behavior and mental states, based on observations during experiments with chimpanzees.

2.1.2 Tasks Used to Assess Theory of Mind Capabilities during Development

Various tasks have been traditionally used to measure the ToM capabilities in development, focusing on different aspects of the process: some tasks have been created to evaluate the capabilities of the participants to infer others' mental states (i.e., mentalizing). In such tasks, participants are usually exposed to descriptions of social situations and are asked to predict the mental states of the characters involved. Examples for this category are the false belief task [Wimmer and Perner 1983] or the strange stories task [Happé 1994]. The level of difficulty of these tasks depends on the target population, that is: more complex versions of these tasks exist for adults versus children; cognitive demands, culture, education, and language skills may also affect the test results. For instance, on the Sally–Anne false-belief task [Wimmer and Perner 1983], one of the most widely used task to evaluate ToM, kids are told a story in which a character, Sally, puts an object inside a basket before leaving the room. Once Sally is out of the room, another character, Anne, changes the location of the object to a box. At this point, the children are asked where Sally would look for the object upon her return. Initial findings revealed that four-year-olds can compute the perspective or state of (false) beliefs of others, revealing that ToM abilities have developed.

Another group of tests evaluates the capability to detect and interpret social signals, as they are pivotal in mindreading. Gaze following [Bayliss et al. 2007, Frischen et al. 2007] or identification of emotions from visual [e.g., Baron-Cohen et al. 1995, De Sonneville et al. 2002, Bayliss and Tipper 2006] or auditory stimuli [Nowicki and Carton 1993, Scherer and Scherer 2011] are examples for this category. Similarly, tests like Reading the Mind in Eyes [Baron-Cohen et al. 1997] or Reading the Mind in Film [Golan et al. 2006] measure whether participants can accurately identify internal states from social signals. While these traditional tasks offer highly controlled measures of participants' ToM abilities, they lack ecological validity. One major challenge with these paradigms is that the evaluation of ToM is based on experimental protocols with spectatorial perspective, meaning that participants are placed in the role of a passive observer rather than partaking in a dynamic social interaction. Furthermore, the evaluation of ToM in those paradigms takes place in non-social contexts, potentially delivering a biased assessment of social cognition compared to real-life situations. More recent approaches, such as the second-person neuroscience framework [Schilbach et al. 2013], stress the importance of allowing natural, reciprocal interactions in experimental paradigms when trying to understand the mechanisms of social cognition. Evidence from clinical studies supports the postulates of second-person neuroscience framework showing that although children with an autism spectrum condition are capable of passing the traditional lab-based false-belief tasks at a mental age of six years [Happé and Frith 2014], adolescents and adults with Asperger syndrome still have difficulties in mentalizing with others during naturalistic interactions [Ponnet et al. 2004].

2.1.3 Neural Correlates of Theory of Mind

Recent studies have revealed two main brain regions (see Figure 1) involved in ToM: the paracingulate cortex, involved in processing of own and others' mental states, and the temporoparietal junction, linked to identifying actions and intentions produced by biological agents. The wider network of brain areas involved in ToM tasks includes the dorsomedial prefrontal cortex, temporal-parietal junction, superior temporal sulcus, ventromedial prefrontal cortex, and posterior cingulate cortex [Amodio and Frith 2006, Frith and Frith 2006, Blakemore 2008, Van Overwalle and Baetens 2009]. These areas have been reported to be activated during various mentalizing tasks, such as making inferences about others' preferences [Mitchell et al. 2002, Jenkins et al. 2008], reading stories about others' mental states [Saxe and Kanwisher 2003], interactive games that require reasoning about intentions [Hampton et al. 2008, Chang et al. 2011, Sul et al. 2017], or watching movies that require inferring characters' mental states [Pantelis et al. 2015, Richardson et al. 2018].

2.2 Joint Attention

Cognitive science has also extensively investigated the phenomenon of JA, which is closely related to-, and a precursor of, ToM. JA takes place when two individuals coordinate their attentional processes to conjointly attend to the same object or situation in the environment. JA is fundamental for the acquisition of language, such that the caregiver says a word for a given object out loud and uses her/his gaze to guide the child's attention to the object in the environment, establishing an association between the spoken word and the object it represents [Baron-Cohen 1997]. JA is a pivotal precursor of joint action and mental state attribution as they are used to communicate a partner's focus of attention and allow inferences about his/her intentions (e.g., looking at a food item might mean that the gazer is hungry) and action goals (e.g., looking at a coffee cup might predict an upcoming grasping action).

2.2.1 Tasks for Measuring JA Capabilities

To examine attentional processes underlying JA, the gaze-cueing paradigm has been used extensively, where participants observe a centrally presented face on the screen that first looks at them and then changes its gaze direction to the left or right side of the screen to either validly or invalidly cue a subsequent target probe [Friesen and Kingstone 1998, Driver et al. 1999, Emery 2000, Frischen et al. 2007]. The standard observation is that targets appearing at the gazed-at (validly cued trials) location are processed faster and more accurately than targets appearing elsewhere (invalidly cued trials), which results in faster response times to targets located at the gazed-cued relative to other locations, where the difference in reaction times across these two conditions is termed the gaze-cueing effect. The gaze-cueing effect is explained as the consequence of enhanced attentional orienting in response to the change in gaze direction that functions as a spatial cue: when the gaze is directed to a location, the observer's attentional focus is shifted there, and this facilitates the sensory processing of stimuli that subsequently appear at the attended location. On the other hand, when a stimulus appears in an uncued location, the observer's attentional focus first needs to be shifted from the cued location to the target location; this additional time for shifting the attentional focus to the target is reflected in reaction times.

Although it has been argued that attention is reflexively oriented by social stimuli, such as changes in gaze direction (see Friesen and Kingstone [1998]), multiple studies (i.e., Wiese et al. [2014] and Kuhn et al. [2018]) have shown that attentional orienting to gaze cues can be top–down controlled when the context in which the interaction takes place is sufficiently social and provides information about the social relevance of the cue [Wiese et al. 2013]. For example, attentional orienting to gaze cues is enhanced when they are believed to be intentional as opposed to random or unpredictable [Teufel et al. 2010, Wiese et al. 2012, 2014, Perez-Osorio et al. 2015, Perez-Osorio et al. 2017; see Capozzi and Ristic 2020, for a review].

2.2.2 Neural Correlates of JA

Several specialized cerebral mechanisms have been postulated as the basis of socio-cognitive mechanisms related to gaze processing and gaze-induced JA. Neuroimaging studies in humans show that the superior temporal sulcus region is implicated in processing various face signals, such as changes in gaze direction or facial expression [Puce et al. 1998, Hooker et al. 2003, Pelphrey et al. 2003, see Allison et al. 2000, for a review]. The intraparietal sulcus (IPS), which generally is activated during covert shifts of attention [Nobre et al. 1997, Corbetta, 1998], is also involved in JA, specifically in shifting the observer's attention to the gazed-at location [Puce et al. 1998, Wicker et al. 1998, Hoffman and Haxby 2000, George et al. 2001, Hooker et al. 2003, Pelphrey et al. 2003]. In support of this notion, a functional magnetic resonance imaging study by Hoffmann and Haxby [2000] reported that passive viewing of faces that showed averted gaze elicited a significantly more robust response in bilateral IPS and left superior temporal sulcus than the passive viewing of faces with direct gaze, indicating that these brain areas are specialized in processing averted gaze. Further, Puce and colleagues [1998] showed that the inferior temporal sulcus is particularly sensitive to eye movements. Hence, inferior temporal sulcus and superior temporal sulcus seem to be specialized in processing gaze direction, while IPS may be preferentially involved in attentional orienting in response to gaze cues.

3. Theory of Mind in Artificial Agents

3.1 Evoking Human Social Cognition Mechanisms during HRI

3.1.1 Theory of Mind in HRI

Implementation of social cognition in SIA can be characterized by simulation of social interactions. This approach relies on scripted social behavior, usually created based on human-like behavior, to endow artificial agents with social signals, actions, and language. Through the combination of pre-scripted behavior, depending on the experimental conditions exhibited during social interactions or through experimenter-controlled protocols referred as Wizard-of-Oz studies [Dahlbäck et al. 1993], social behavior of artificial agents become dynamic and interactive within well-defined experimental scenarios. Multiple studies have shown that simulating social interactions with SRs can trigger ToM-related processes in HRI [Hegel et al. 2008, Byom and Mutlu 2013, De Graaf and Malle 2019]. Manipulating beliefs about an agent's capacity of having internal states or exploring the effects of communicative gestures in cooperative or competitive interactions, for instance, have been used to understand whether and under which conditions people mentalize with artificial agents. These studies help understand whether people would spontaneously adopt a mentalistic approach to artificial agents and what factors contribute to the likelihood of attributing mental states to SIA.

Many studies have examined to what extent humans interpret social signals displayed by SR in mentalistic terms and whether they are used to infer a robot's mental state. As an example, Mutlu and colleagues [2009] investigated whether non-verbal cues during an interactive game would elicit ToM inferences in participants, leading them to follow the cues of the SR. In this experiment, the participants' task was to guess which object the robot had chosen from objects depicted on top of a table. The results showed that participants were faster and more accurate when the robot used non-verbal social cues relative to no social cues. Interestingly, when asked afterwards, most of the participants did not explicitly pay attention to the cues or actively use them to complete the task. Other interactive protocols showed that humans tend to take into account the behavior and, arguably, internal states of SR when trying to coordinate or synchronize their actions with them during joint task execution [Xu et al. 2016, Ciardo et al. 2020].

Schreck et al. [2019], for instance, evaluated whether a SR's social behavior, the type of social signals it displayed, and the proxemics (how close would the SR get to people) affected the likelihood of ToM-related interpretations. They found that increased experience with a robot through continued interaction decreased the likelihood of mental state attributions, unless the robot showed more socially active behaviors (get close to people when interacting) as well as more human-like expressions, which triggered stable levels of mental state attributions across the experiment.

Instead of looking at the effect of social signals on mentalizing, more recent approaches ask a broader question, namely which physical and behavioral features SRs need to display in order to being perceived as an entity "with a mind," capable of displaying internal states [Epley et al. 2007, Gray et al. 2007, Złotowski et al. 2015]. When non-human agents are treated as agents with a mind, humans adopt the intentional stance [Dennett 1987] to them and interpret their behavior based on the assumption that it is motivated by internal states such as beliefs, desires, or intentions. Given the general capacity of SIAs to display physical and behavioral signs of human-likeness, and the human tendency to anthropomorphize non-human entities, it plausible to assume that humans would use mentalistic strategies to explain and predict the behaviors of artificial agents [Perez-Osorio and Wykowska 2019, for review]. While this assumption is theoretically plausible, empirical evidence is mixed. The seminal study by Gallagher and colleagues [2002] observed differential activation in the anterior cingulate cortex, previously linked with mentalizing tasks, when participants believed that they were playing rock-paper-scissors against a human compared to playing against a rule solving program or a random response generator. Importantly, all the conditions were controlled by the same algorithm and the only difference between them was the stance of the participants toward the opponent. However, Chaminade et al. [2012] used a similar paradigm and replaced the rule solving opponent with a humanoid robot and reported no differences in brain activation between humanoid robot and the random responses. Furthermore, Krach and colleagues [2008] who used the Prisoner's dilemma instead of rock-paper-scissors and found that the medial prefrontal and left temporo-parietal junction, associated with the attribution of intentional stance and part of the ToM network, were only activated in response to humans but not during the interaction with artificial agents (a humanoid robot, a functional robot without human-like appearance, and a computer). More recent evidence suggests that in addition to belief manipulations, other cues that manipulate social context information also have the potential to trigger participants to adopt the intentional stance toward SIA. For example, a recent study manipulating the intentional stance by presenting human and SR agents embedded in different scenes, and asking participants to score the plausibility of different explanations for the agents' behaviors, shows no difference in participants' tendencies to adopt mentalistic explanations toward human versus SRs [Thellman et al. 2017]. In a similar study, Marchesi et al. [2018] also showed that people spontaneously adopt the intentional stance toward humanoid robots under some contexts using a novel questionnaire, the Instance Questionnaire (ISQ), that was specifically developed to measure peoples' tendencies to adopt the intentional stance toward a robot [specifically, the humanoid robot iCub, Metta et al. 2010]. In the questionnaire, participants observe a series of pictures showing a sequence of events involving iCub and are then asked to judge whether its behavior was motivated by a mechanical (e.g., malfunction, calibration) or mentalistic reason (e.g., desire, curiosity), with the latter explanation indicating adoption of the intentional stance, see Figure 2. Results showed that although participants tended to give more often mechanistic explanations for iCub's behaviors, some behaviors evoked mentalistic interpretations. Interestingly, inter-individual might have also played a role in likelihood of adopting the intentional stance toward the robot.

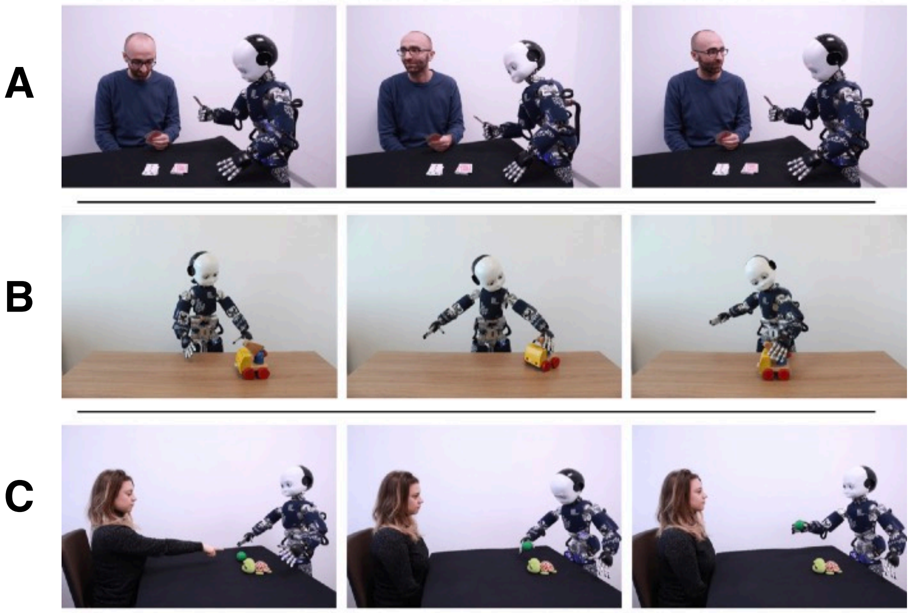


Figure 2. Examples of scenarios from the Instance Questionnaire. Under each scenario participants chose the explanation that would better describe the behavior of the robot, either a mentalistic or mechanistic statement. For example, in Panel A the options were "iCub was trying to cheat by looking at opponent's cards" for mentalistic description, and "iCub was unbalanced for a moment" for mechanistic description. Copyright © 2019 Marchesi, Ghiglino, Ciardo, Perez-Osorio, Baykara and Wykowska, Istituto Italiano di Tecnologia (CC BY 4.0).

Another crucial factor that has been hypothesized to influence adopting the intentional stance and that can be directly influenced via robotic design is whether the behavior of a robot seems human-like [Złotowski et al. 2015]. Wykowska et al. [2015], for instance, showed that variable temporal characteristics of gaze behavior lead participants to judge a robot's behavior as more human-like compared with less variable eye movements. Willemse et al. [2018] also reported that participants anthropomorphized and liked robots more that followed the participants' gaze during an interactive experiment that exhibited typical human-like reciprocity

Perez-Osorio et al. [2019] further showed that when participants had high expectations regarding the behavior of the robot, their scores in the ISQ increased after a brief observation of the robot; for participants with lower expectations ISQ scores decreased after the observation. Collectively, evidence suggests that people can (and do sometimes) attribute mental states to SRs and employ these attributions during social interactions. Human-like shape and behavior might facilitate the attribution of mental states to robots but might not be sufficient; rather, the type of interaction and the social signals exhibited play a crucial role on this process, as well as individual attitudes [De Graaf et al. 2016] and imageries of robots [De Graaf and Malle 2019].

3.1.2 Joint Attention in HRI

Several studies have examined JA in HAI, and have shown that people can identify and follow non-human gaze and discriminate whether the SR is looking at them or at a different location (see Admoni and Scasselatti [2017] for a review). Findings also suggest that robot gaze can be used to communicate information about relevant events and targets in the world—similar to human gaze. Although some studies have shown that the directional gaze of two different robots failed to elicit reflexive attentional orienting (i.e., Admoni et al. [2011] and Okumura et al. [2013]), several other studies have consistently shown engagement in JA with artificial agents. This occurs in screen-based studies [Wiese et al. 2012, 2019], but also when using embodied humanoid agents in interactive protocols [Wykowska et al. 2015, Kompatsiari et al. 2018, Chevalier et al. 2019 for review]. Various experimental conditions can modulate certain aspects of social attention, which, in consequence, results in variable empirical findings that can vary depending on which paradigm has been used and how it has been implemented. For instance, the robot's believed intentionality appear to modulate gaze effectiveness in orienting attention (i.e., the extent to which gaze orients attention) [Wiese et al. 2012, Wykowska et al. 2014]. Participants also respond more favorably to robots that display socially engaging gaze, for example, in the series of studies by Kompatsiari and colleagues [2017], gaze cueing effects were modulated by whether the robot engaged or not in mutual gaze with the participants prior to directing their attention to one of the locations where the target could appear (the gaze cueing procedure). Willemse et al. [2018] as well as Willemse and

Wykowska [2019] showed that degree of contingency of the robot's gaze on participants' gaze direction also influenced JA. Finally, Perez-Osorio et al. [2018] showed that action expectations also affect the magnitude of the gaze cueing effect. Huang and Mutlu [2012] also found that participants recalled the details of a story better when the SR used congruent speech and gaze cues than when the cues were spatially or temporally incongruent. Similarly, Mutlu et al. [2013] found that participants responded faster and understood instructions better when the SR used verbal and visual cues. These findings indicate that humans can engage in JA with artificial agents such as robots and that non-verbal social cues can be beneficial for human–robot interaction.

4. Modeling Social Cognition

4.1 Implementing Theory of Mind in SR

Considering the strong bias that humans have to interpret others' behaviors in anthropomorphic or mentalistic terms, it is natural to assume that in social interactions with artificial agents people would employ the same social-cognitive mechanisms as in interactions with humans. Roboticists might have followed a similar intuition when deciding how to best design SRs: they aim to create robots that can communicate using human-like social signals and to equip robots with social skills and cognitive capabilities comparable to humans in order to facilitate social interactions. For example, Scassellati [2002] suggested that endowing a robot with ToM would be very beneficial for social interactions as robots could use such a model not only to understand human behavior and communicate efficiently with humans, but also to learn from social interactions the same way that infants learn from their parents. Endowing a SR with ToM would not only allow robots to generate internal representations of humans' mental states, and to appropriately respond to these mental states, but it would also help robots to interact smoothly and fluently with humans. For that purpose, Scassellati extracted the most relevant aspects of traditional psychological models of ToM from developmental cognitive science [e.g., Baron-Cohen 1995, see Figure 3], and aimed to create analogous structures in artificial robot systems.

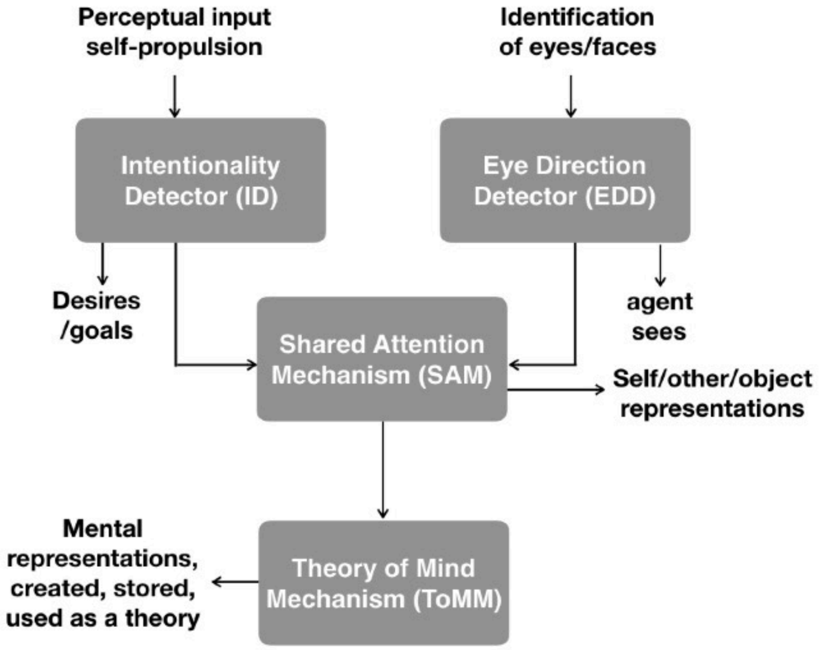

Perceptual input self-propulsion
Identification of eyes/faces
Intentionality Detector (ID)
Eye Direction Detector (EDD)
Desires /goals
agent sees
Shared Attention Mechanism (SAM)
Self/other/object representations
Mental representations, created, stored, used as a theory
Theory of Mind Mechanism (ToMM)

Figure 3. Theory of mind model based on Baron-Cohen [1995]. Detection of stimuli in the Intentional Detector (ID) module and the Eye Direction Detector (EDD) module constitute the basic input for the model. Representations created in this first layer feed the Shared Attention Mechanism (SAM) to build triadic representations. In the final layer, the Theory of Mind Mechanism encodes and stores representations to create a theory about others' mental states and beliefs. The levels increase in complexity and mature sequentially during development (based on Baron-Cohen [1995]).

The ToM formulated by Baron-Cohen [1995] proposes that humans develop a mindreading system consisting of four different modules: intentionality detector (ID), eye direction detector (EDD), shared attention mechanism (SAM), and theory of mind mechanism (ToMM). The ID recognizes entities in the environment that exhibit biological motion and is able to detect self-propelled motion and goal-oriented behaviors and thus can identify an organism with volition or agency [Premack 1990]. The EDD automatically detects the presence of eyes/face in the visual field and decides whether eye-like stimuli are "looking at me" or "looking at something else" and thus whether an agent shows mutual gaze (signaling readiness to engage) or averted gaze (trying to shift observer's attention to potential objects of interest). ID and EDD become functional earlier in development and precede the maturation of SAM and ToMM. SAM receives input from both ID and EDD to determine whether two biological interaction partners conjointly attend to the same event or object in the environment, thus creating a triadic representation (self, other, object) out of different dyadic representations (self/other, self/object, other/object). SAM usually develops at 9 to 14 months of age and allows JA behaviors, such as proto-declarative pointing and gaze monitoring. Importantly, SAM allows the agent to interpret the gaze change of others as intentional and, what follows, as intentional representations (i.e., "she wants to…"), which highlights the importance of gaze perception for the successful inference of others' mental states.

The most advanced module, the ToMM, enables representing and integrating the full set of mental state concepts into a "theory"; it creates representations of others' beliefs and desires but also allows for formulating knowledge states that are neither necessarily true nor match the knowledge of the agent (i.e., imagination and creativity). ToMM forms later in development (between 2 and 4 years of age) and allows pretend-play [Leslie 1987] as well as understanding false beliefs [Wimmer and Perner 1983] and the relationship between mental states [Wellman 1990]. Baron-Cohen's model has proven to be useful in interpreting typical and atypical development of social skills in humans, autism spectrum condition in particular. An important part of this model is that it is hierarchical with representations of different levels of complexity, starting with precursor functions like the detection of biological agents (ID) and gaze signals (EDD), continuing to the maturation of shared attention (SAM), and finally the successful representation and inference of others' mental states (ToMM).

In the process of implementing a ToM model in artificial agents, Scassellati first introduced the EDD and ID modules and over the years additional modules have been proposed: modules to distinguish animate from inanimate motion [Scassellati 2001], to share attention [Nagai et al. 2002, Scassellati 2002], to imitate actions as a method for learning motor skills and recognizing human actions [e.g., Schaal 1997, Demiris and Hayes 2002, Fod et al. 2002, Billard et al. 2004, Breazeal et al. 2005, Gray et al. 2005, Johnson and Demiris 2005], and to take others' perspective [Gray et al. 2005, Trafton et al. 2005, 2006]. The main challenge consists, however, in generating and integrating different motor and social skills into an articulated architecture able to cope with the changing environmental demands and able to adapt to multiple and variable social contexts.

Since this early work, the implementation of ToM models in socially interactive agents has progressed considerably. Most recent advancements (for a review, see Bianco and Ognibene [2019]) have resulted in the formulation of more complex cognitive architectures that aim at providing social skills to SIA. In recent years, computational models of social cognition have been used to understand cognitive mechanisms of ToM by simulating functioning ToM [Newell 1994, O'Reilly et al. 2012]. These models of ToM vary in their characteristics but typically highlight the importance of detecting social signals conveyed by others (with most of these signals being conveyed by eyes and faces), identification of goals and motivations (i.e., task-related goals), and creation of beliefs (based on state of the world and the estimation of others' knowledge). For example, Baker et al. [2009] propose a Bayesian theory of mind (BToM), a model that formalizes action understanding as a Bayesian inference problem. This approach models beliefs, goals, and desires as rational probabilistic planning in Markov decision problems (MDPs) and the goal inference is performed by the Bayesian inversion of this model of planning. The MDPs are a normative framework for modeling sequential decision-making processes under uncertainty, commonly used for human planning and reinforced learning [Dayan and Daw 2008]. MDPs allow creating representations of an agent's interaction within the environment and encode all relevant information about the configuration of the world and the agent as state variable, which allows capturing mental models of intentional agents' goal and environment-based planning. Further, MDPs represent actions permitted in the environment and determine a causal model of implications of these actions in the state of the world; they also represent subjective rewards or costs caused by the agent's actions in each state.

The model creates an agent's hypothetical representations of beliefs and desires that caused a behavior within that given environment using Bayesian inference; all hypotheses are associated with a particular goal. For each hypothesis, the model evaluates the likelihood of generating the observed behavior given the hypothesized belief or desire. Then, the model integrates this likelihood with the prior over mental states to infer the agent's joint belief and desire [Baker et al. 2009, 2017]. Although the model integrates beliefs, desires, and goals, it depends strongly on priors regarding the action goals (in contrast with other models of ToM). Importantly, and unlike other models of ToM, BToM performs Bayesian inferences over beliefs and desires simultaneously. The cognitive model incorporates the current perceptual states and beliefs' updates in order to modify the initial hypothesis and then generates new adjusted predictions in each iteration.

To evaluate the model, the authors tested whether the model is able to predict the mental states of an agent performing a decision-making task (i.e., choose a food truck) displayed in three action frames and then contrasted these predictions with both human and alternative models' performance on the same task. On each trial, an agent was looking for a food truck in three frames, starting point, transition, and goal, in different spatial configurations (layouts). After each trial, participants (and models) were asked to predict the agent's preferences regarding the food trucks and to rate how confident they were with their assessment. The authors contrasted the predictions of the BToM model with humans' performance and assessments and against two model alternatives—TrueBelief (a special case of BToM with a prior that assigns probability 1 to the true world state) and NoCost (another special case of BToM that tests the contribution of the principle of efficiency to ToM reasoning by assuming that the agent's cost of action is zero), as well as one cue-based alternative—Motion Heuristic (which tests whether social inferences are derived from processing of bottom-up perceptual features). They found that the proposed model successfully predicts the mental states of an agent and generates mental-state judgments similar to those of human participants in a wide variety of environment configurations. These findings obtained with Bayesian inversion of models of rational agents suggest that the brain might use similar principles to handle social information, infers others' mental states, and predicts their actions. Thus, Bayesian computational models offer a powerful tool to evaluate the inherent predictive functioning of the brain.

Cangelosi and colleagues have also been developing cognitive architectures incorporating ToM (e.g., Vinanzi et al. [2019]). The authors designed and implemented a biologically inspired artificial cognitive system that incorporates trust and ToM, which is supported by an episodic memory system and based on developmental robotics. The cognitive system integrates multimodal perception (visual and auditory stimuli) together with a motor module. The visual module detects and recognizes faces through machine learning algorithms Haar Cascade [Viola and Jones 2001] and Local Binary Pattern Histogram [Ojala et al. 2002]. This cognitive system also has a belief module based on Bayesian belief networks. Representations are stored in the episodic memory module to be retrieved and included in future interactions with new users. Interestingly, the architecture was tested using an experimental paradigm from developmental psychology. The paradigm has been developed to evaluate how much children trust an interaction partner [Vanderbilt et al. 2011]. Only children 5 years of age and older are typically able to pass this test, thanks to the emergence of the ToM. To pass the test, children need to differentiate people who give useful cues (helpers) from people who are lying (trickers). Interestingly, the proposed architecture satisfactorily identified helpers from trickers, thereby passing the test.

The creation of neurocognitive models of ToM mechanisms with computational simulations and architectures and the further observation of the effects of these models in interactive situations has two main advantages. As mentioned above, the implementation of ToM models in SIA would facilitate interaction with humans, as the models anchor the agents' behavior in predictive identification of action and proactive generation of responses. Furthermore, implementation of such models in SIA also provides a new tool for understanding the ToM mechanisms in humans during social interactions.

Finally, some cognitive architectures, like the work proposed by Rabinowitz et al. [2018], use ToM neural networks to infer mental states online based on a meta-learning approach. The application of strong prior results in inferences that require only a few observations, adapting quickly to different tasks and behaviors, which brings the architecture closer to human performance. It is important to mention that although this neural network receives only visual inputs, it can solve false belief tasks. Another notable example of biologically inspired architectures is the work of Kahl and Kopp [2018, 2019]. The authors propose a mentalizing system for attributing and inferring mental states together with a hierarchical predictive model for online action perception and production that represents the mirror system. While the mentalizing subsystem allows differentiating mental perspectives for "me," "you," and "us," the mirror subsystem adopts the Empirical Bayesian Belief Update model [Sadeghipour and Kopp 2011] for action observation and production. The architecture allows second-order ToM, that is, representations of beliefs about beliefs, in actual simulations of dynamic interaction.

4.2 Simulating Other Social Cognition Mechanisms that Support Theory of Mind in Artificial Agents

Simulation of higher-order socio-cognitive capabilities includes action recognition, imitation, memory, and learning. All these modules contribute to model a functional ToM. The model can create, store, retrieve, and track the counterpart's mental states and compare them with its own internal states online. This type of simulation allows making inferences about goals, predict actions, and also facilitates learning [Byom and Mutlu 2013]. Furthermore, these cognitive simulations allow (i) making inferences about the perspective of humans [Trafton et al. 2005] and (ii) distinguishing and storing particular sets of beliefs to help robots plan actions and to learn based on imitation [Breazeal et al. 2006] or activate motor-resonance mechanisms that facilitate generation of inferences about subsequent steps of actions sequences [Blakemore and Decety 2001]. Another example of cognitive simulation in the motor domain is a system developed by Gray et al. [2005] that monitors the human behavior in a collaborative task by simulating the observed behavior within the robot's own generative mechanisms. This enables the robot to perform task-level simulations, track the participant's progress, and anticipate the needs to accomplish the action goal.

A considerable part of the initial work on ToM in robotics was focused on implementing the visual perspective-taking mechanism and the so-called "belief management system" in the SR in order to infer the humans' mental states (i.e., Scassellati [2002] and Berlin et al. [2006]). A prominent example of the integration of these modules was presented by Breazeal et al. [2009]. The authors aimed at incorporating mechanisms based on the simulation theory as a principle to support mind-reading skills and abilities. The proposed model is characterized by two modes of operation: one, which generates the mental states of the SR using the current state of the world, and a second one, which constructs and represents the mental states of the human counterpart. Importantly, both modes share the perception, belief, motor, intention, and body representation modules. The perception system can estimate what the other can see and transforms that information into the point of view of the robot. The motor system maps and represents the body positions of the human in terms of the SR joint space to perform action recognition. The belief system combined with perspective taking represents possible beliefs of the human. Finally, the intention system predicts the ideal action sequence to achieve a goal.

This information is combined with the perceptual cues and the current state of the environment to create inferences about mental states and to predict the actions of the human counterpart. According to the authors, the system to represent the human's mental states builds beliefs based on perceptual states using an embodied simulation together with higher-level knowledge about task-goals [Breazeal et al. 2009]. Their architecture was tested in a collaborative task and a learning-from-demonstration task. The SR was able to anticipate and generate inferences about human behavior and pointing at relevant objects during the collaborative task. It was also capable of recognizing rules demonstrated by the teacher in the learning task. The physical limitations of the SR platform (i.e., the social robot Leonardo) prevented physical interaction with the environment. For that reason, the capabilities of the architecture were tested also in a virtual reality environment showing successful results. The authors concluded that the system can infer and predict the beliefs of the interaction partner, although the range of these beliefs was limited compared to the capabilities of humans.

More recent cognitive architectures attempt to develop more flexible artificial ToM systems to enhance robots' capabilities to improve human–robot interactions, allowing SRs to take others' perspectives, generate predictive actions, support active perception, and reduce the dependence on external datasets to infer actions and mental states [Bianco and Ognibene 2019]. There are several types of cognitive architectures. For instance, multimodal architectures rely on collecting inputs from different modalities (i.e., visual, auditory, and proprioception) to predict and understand the behavior of the human and to reproduce movements. Inputs that include posture, location, facial expressions, visual perspective, and movements of the human can be used to determine whether the actions are intended or not. This inputs are also integrated with verbal commands and proprioceptive information to perform collaborative tasks like cleaning a table in the most efficient manner. The biomimetic architecture for situated social intelligence systems (BASSIS) proposed by Petit and colleagues [2013] provides a robot real-time adaptation during collaborative scenarios using multimodal inputs (visual, verbal, and proprioceptive) to infer the mental states of the human counterpart. This architecture is organized at three different levels of control: reactive, adaptive, and contextual, which are all based on the physical instantiation of the agent through its body. It is based on the Distributed Adaptive Control Architecture [Verschure et al. 2003] and was employed for multimodal learning with NAO and iCub platforms. It has shown great potential for collaborative environments. However, it is limited by the quality of teaching (as it does not tolerate errors from the tutor) and by limited long-term storing of the acquired learning [Verschure et al. 2003].

Another architecture that uses multimodal estimations was proposed by Görür et al. [2017]. In contrast with BASSIS, it integrates a ToM model into decision-making tasks. This architecture is composed of three modules: Sensing, Action State Estimation (ASE), and the Human–Robot Shared Planner (HRSP). The architecture receives primarily sensory data (visual and auditory/ verbal) and generates stochastic policies in the form of human–robot shared decisions from the robot's point of view. A combination of sensory data and generated policies produces an input that drives the remainder of the architecture. The ASE and the HRSP constitute the ToM part, but they rely closely on the Sensing module. The stochastic planner of the HRSP depends on partially observable Markov decision process (POMDP), which is a Bayesian ToM model inspired by Baker and Tenenbaum [2014]. Similarly, Devin and Alami [2016] designed a model that employs multiple inputs to estimate and maintain representations about the environment and the actions goals of the partner. It includes representations of the previous goal, plans, and actions, and holds them online to decrease unnecessary or redundant verbal communication with the human counterpart.

A variation of Devin and Alami's architecture [2016] has been proposed by Demiris and Khadhouri [2006], called HAMMER (Hierarchical Attentive Multiple Models for Execution and Recognition). This architecture is an example of multimodal estimation and hypothesis simulation. It has been designed to identify and execute goal-oriented actions and has an inverse model paired with a forward model. The inverse model processes the current state of the system and the target goal(s) and produces the control commands that are needed to achieve or maintain those goal(s); the forward model takes the current state of the system as input and a control command to be applied on it and outputs the predicted next state of the controlled system. This architecture is based on the visual perception of another agent's movements, which is controlled by top–down signals to orient the robot's attention toward information necessary to confirm its hypothesis concerning the demonstrator's action. This architecture was implemented and tested on a robot that conducted an action recognition task while observing a human demonstrator performing an object-oriented action. The robot successfully performed the task and the attentional mechanism acting over the inverse model was suggested to reduce robots' computational costs.

5. Comparison of IVAs and SRs

The physical embodiment of interaction partners is known to impact social interactions (see Li [2015], for review). However, there is no consensus regarding the question whether artificial agents' embodiment has an effect on ToM and JA specifically. Users typically find physically embodied SRs more engaging, enjoyable, informative, and credible [Kidd and Breazeal 2004] than virtual agents. Physically co-present embodied systems improve interactions over virtual systems [Bainbridge et al. 2011]. In general, a wide variety of studies in HRI support the idea that implementing human-like characteristics in SRs facilitates social interaction. However, there is sparse literature that supports a systematic comparison of the SRs and IVAs. Studies showed that SRs with human-like appearance and behavior are judged as more pleasant [Axelrod and Hone 2005], more usable [Riek et al. 2009], more accepted [Venkatesh and Davis 2000, Kiesler and Goetz 2002, Duffy 2003], easier to get acquainted with [Krach et al. 2008], and more engaging [Bartneck and Forlizzi 2004]. Further, SRs that communicate using social signals, such as facial expressions [Eyssel et al. 2010], other emotion displays, like ear or fin movements in human-like robots, [Gonsior et al. 2011], or turn-taking in conversations [Fussell et al. 2008] evoke stronger emotional responses and are preferred by users over SRs that do not show social signals. IVAs with these or similar characteristics would also be expected to facilitate the attribution of mental capacities to them, and studies indeed suggest that people can read intentions into IVAs' behaviors during social interactions. For instance, using the principles of animation, Takayama et al. [2011] designed the behavior of an IVA such that it either displayed intentions (i.e., hint with the gaze whether it is aiming at opening a door) or not, and was reactive (or not) to the events in the environment during action execution. They found that when an IVA showed forethought (time to "think") before executing an action, the outcome was judged as more competent and intelligent, and the agent was perceived as more appealing. This suggests that people are sensitive to the intentional hints from IVAs and can interpret them accordingly.

Several studies have focused on whether virtual agents can communicate using the gaze and engage participants during interaction. For example, Andrist and colleagues [2012] used a model to control the gaze shifts of a virtual character; two main conditions, mutual and referential gaze, were developed. Predominant mutual gaze elicited subjective positive feelings of connection, and referential gaze improved participants' recall of information in the environment. This suggests that similar to findings with SRs, participants seem to follow the gaze of IVAs and engage in mutual eye gaze with IVAs [Andrist et al. 2012]. In a similar study, Wilms et al. [2010] showed that when IVAs' gaze was contingent with participants' gaze, JA evoked higher activity in the medial prefrontal cortex and posterior cingulate cortex relative to disjoint conditions. More recently, Willemse and colleagues [2018], using a gaze leading task on screen (iCub followed or not participants' gaze), found that participants preferred the robot that exhibited JA behavior relative to the robot with a disjoint attention behavior. The robot with JA behavior was also rated as more human-like and as more likeable. These results showed a similar pattern to findings obtained with a physically embodied robot [Willemse and Wykowska 2019].

Development of cognitive architectures makes it now possible to implement ToM in IVAs. For instance, Buchsbaumm et al. [2005] proposed a framework inspired by simulation theory and hierarchical action structures to help IVAs understand human actions and emotions. The framework includes a motivational system in which certain actions (get/search/find) are associated with certain drives (feeding, self-defense, or socializing), and increasingly specific, sequentially organized actions can be defined for satisfying associated drives (e.g., feeding —% eat/search/get food —% jump/reach item). The system is designed to learn by imitation, that is, associate an observed action with a particular goal, and is able to identify goals, learn, and predict actions. Other approaches have implemented BToM to predict the behavior of the users, aiming at facilitating the navigation and increasing the users' satisfaction within an immersive virtual environment with multiple agents, rather than a one-to-one interaction in a social context. The algorithm proposed by Narang et al. [2019] uses a probabilistic model that integrates observed social cues and actions with statistical priors regarding the user's mental states. The model, used in a real-time algorithm, endows multiple virtual agents with a ToM model. The algorithm perceives the proxemics and the gaze-based social cues from the users to reliably infer their underlying implicit intentions. For instance, the algorithm can differentiate between a user who is passing and a user who is aiming to talk/interact with a virtual agent. Altogether, this suggests that both IVAs and SRs can communicate using social signals and language. Further research employing the methods of cognitive neuroscience should be carried out to evaluate the added value of physical embodiment in SRs as compared to IVAs.

6. Current Challenges

One of the central questions in HRI is: What are the necessary conditions for a robot to evoke similar social-cognitive mechanisms as in human–human interaction? Which robot features make us socially attune or synchronize with it, and represent its actions and interpret them in mentalistic terms? Which features are more impactful, physical or behavioral, and would their impact be different in the short-term versus long-term interactions? What role do preexisting expectations, stereotypes, and individual differences play when interpreting and reacting to observed robot behaviors (see Marchesi et al. [2019]; Spatola and Wykowska, [2021], and Bossi et al. [2020])? Answering these questions by no means suggests that it is desirable under all circumstances to design robots that evoke socio-cognitive mechanisms to the same extent as other humans. Whether it is ethically and morally acceptable, and in which application contexts, is open to ethical debate. On the one hand, robots that evoke similar social schemes as human interaction partners may have positive effects as SRs are perceived as more friendly, which might lead to higher acceptance, for instance, in elderly care. In fact, it might be that a senior person is more likely to, for example, follow medical recommendations (e.g., taking pills at a prescribed time of the day) when a robot elicits social attunement, as compared to when the robot is perceived as a machine. Whether this is indeed the case remains to be tested in future research. However, there might be application scenarios where social attunement is not desirable. For example, when a person is working side by side with a robot in a factory or wants to use the robot for a specific service. Studies that have examined the fit between a robot and the task it is supposed to execute suggest that anthropomorphic design features might help for tasks that require “core” human skills, like reading emotions, but might be disadvantageous for tasks that require the robot to execute actions that a human would not want to execute (e.g., Smith et al. [2016] and Hertz and Wiese [2018]). Finally, it is also important to evaluate the ethical implications of creating the "illusion" that a robot is a social being "with a mind," similar or equivalent to another human. Humans should always remain aware of the difference between a robot, which is just machine, and another human. The challenge is to make sure that ethical debate goes hand in hand with technical and scientific development and research.

References

H. Admoni and B. Scassellati. 2017. Social eye gaze in human–robot interaction: A review. J. Hum. Robot Interact. 6, 1, 25–63. DOI: https://doi.org/10.5898/JHRI.6.1.Admoni.

H. Admoni, C. Bank, J. Tan, M. Toneva, and B. Scassellati. 2011. Robot gaze does not reflexively cue human attention. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 33, No. 33).

T. Allison, A. Puce, and G. McCarthy. 2000. Social perception from visual cues: Role of the STS region. Trends Cogn. Sci. 4, 267–278. DOI: https://doi.org/10.1016/S1364-6613(00)01501-1.

D. M. Amodio and C. D. Frith. 2006. Meeting of minds: The medial frontal cortex and social cognition. In Discovering the Social Mind. Psychology Press, 183–207.

S. Andrist, T. Pejsa, B. Mutlu, and M. Gleicher. 2012. Designing effective gaze mechanisms for virtual agents. In Proceedings of the ACM Annual Conference on Human Factors in Computing Systems (CHI). ACM Press, Austin, TX, 705–714.

L. Axelrod and K. Hone. 2005. Uncharted passions: User displays of positive affect with an adaptive affective system. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). DOI: https://doi.org/ 10.1007/11573548_ 114.

W. A. Bainbridge, J. W. Hart, E. S. Kim, and B. Scassellati. 2011. The benefits of interactions with physically present robots over video-displayed agents. Int. J. Soc. Robotics 3, 1, 41–52. DOI: https: //doi.org/10.1007/s12369-010-0082-7.

C. L. Baker, R. Saxe, and J. B. Tenenbaum. 2009. Action understanding as inverse planning. Cognition 113, 3, 329–349. DOI: https://doi.org/10.1016/j.cognition.2009.07.005.

C. L. Baker, J. Jara-Ettinger, R. Saxe, and J. B. Tenenbaum. 2017. Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. Nat. Hum. Behav. 1. DOI: https:// doi.org/10. 1038/s41562-017-0064.

S. Baron-Cohen. 1997. Mindblindness: An Essay on Autism and Theory of Mind. MIT Press, Cambridge, MA. DOI: https://doi.org/10.7551/mitpress/4635.001.0001.

S. Baron-Cohen, R. Campbell, A. Karmiloff-Smith, J. Grant, and J. Walker. 1995. Are children with autism blind to the mentalistic significance of the eyes? Br. J. Dev. Psychol. 13, 379–398. DOI: http s://doi.org/10.1111/j.2044-835x.1995.tb00687.x.

C. Bartneck and J. Forlizzi. 2004. A design-centred framework for social human–robot interaction. In Proceedings—IEEE International Workshop on Robot and Human Interactive Communication. DOI: https://doi.org/10.1109/roman.2004.1374827.

A. P. Bayliss and S. P. Tipper. 2006. Gaze cues evoke both spatial and object-centered shifts of attention. Percept. Psychophys. 68, 2, 310–318. DOI: https://doi.org/10.3758/BF03193678.

A. P. Bayliss, A. Frischen, M. J. Fenske, and S. P. Tipper. 2007. Affective evaluations of objects are influenced by observed gaze direction and emotional expression. Cognition 104, 3, 644–653. DOI: https://doi.org/10.1016/j.cognition.2006.07.012.

G. Beattie and A. W. Ellis. 2017. The Psychology of Language and Communication. Taylor & Francis. DOI: https://doi.org/10.4324/9781315187198.

M. Berlin, J. Gray, A. L. Thomaz, and C. Breazeal. 2006. Perspective taking: An organizing principle for learning in human–robot interaction. In Proceedings of the National Conference on Artificial Intelligence.

F. Bianco and D. Ognibene. 2019. Functional advantages of an adaptive theory of mind for robotics: A review of current architectures. 2019 11th Computer Science and Electronic Engineering Conference, CEEC 2019—Proceedings. 139–143. DOI: https://doi.org/10.1109/CEEC47804.2019. 8974334.

A. Billard, Y. Epars, S. Calinon, S. Schaal, and G. Cheng. 2004. Discovering optimal imitation strategies. Rob. Auton. Syst. 47, 2-3, 69–77. DOI: https://doi.org/10.1016/j.robot.2004.03.002.

P. Billeke and F. Aboitiz. 2013. Social cognition in schizophrenia: From social stimuli processing to social engagement. Front. Psychiatry 4, 4. DOI: https://doi.org/10.3389/fpsyt.2013.00004.

S. J. Blakemore. 2008. The social brain in adolescence. Nat. Rev. Neurosci. 9, 4, 267–277. DOI: https: //doi.org/10.1038/nrn2353.

S. J. Blakemore and J. Decety. 2001. From the perception of action to the understanding of intention. Nat. Rev. Neurosci. 2, 8, 561–567. DOI: https://doi.org/10.1038/35086023.

F. Bossi, C. Willemse, J. Cavazza, S. Marchesi, V. Murino, & A. Wykowska. 2020. The human brain reveals resting state activity patterns that are predictive of biases in attitudes toward robots. Sci Robs, 5(46).

C. Breazeal, D. Buchsbaum, J. Gray, D.Gatenby & B. Blumberg, 2005. Learning from and about others: Towards using imitation to bootstrap the social understanding of others by robots. Art life, 11(1-2), 31-62.

C. Breazeal, M. Berlin, A. Brooks, J. Gray, and A. L. Thomaz. 2006. Using perspective taking to learn from ambiguous demonstrations. Rob. Auton. Syst. 54, 5, 385–393. DOI: https://doi.org/10.1016/j.ro bot.2006.02.004.

C. Breazeal, J. Gray, and M. Berlin. 2009. An embodied cognition approach to mindreading skills for socially intelligent robots. Int. J. Rob. Res. 28, 5, 656–680. DOI: https://doi.org/10.1177/0278364909102796.

D. Buchsbaumm, B. Blumberg, C. Breazeal, and A. N. Meltzoff. 2005. A simulation-theory inspired social learning system for interactive characters. IEEE International Workshop on Robot and Human Interactive Communication, 2005. Nashville, TN, 2005, 85–90. DOI: https://doi.org/10.1109/RO MAN.2005.1513761.

L. J. Byom and B. Mutlu. 2013. Theory of mind: Mechanisms, methods, and new directions. Front. Hum. Neurosci. 7, 413. DOI: https://doi.org/10.3389/fnhum.2013.00413.

F. Capozzi and J. Ristic. 2020. Attention AND mentalizing? Reframing a debate on social orienting of attention. Visual Cognit. 28, 97–105. DOI: https://doi.org/10.1080/13506285.2020.1725206.

T. Chaminade, D. Rosset, D. Da Fonseca, B. Nazarian, E. Lutcher, G. Cheng, and C. Deruelle. 2012. How do we think machines think? An fMRI study of alleged competition with an artificial intelligence. Front. Hum. Neurosci. 6, 103. DOI: https://doi.org/10.3389/fnhum.2012.00103.

L. J. Chang, A. Smith, M. Dufwenberg, and A. G. Sanfey. 2011. Triangulating the neural, psychological, and economic bases of guilt aversion. Neuron 70, 3, 560–572. DOI: https://doi.org/10.1016/j.ne uron.2011.02.056.

P. Chevalier, K. Kompatsiari, F. Ciardo, and A. Wykowska. 2019. Examining joint attention with the use of humanoid robots—A new approach to study fundamental mechanisms of social cognition. Psychon. Bull. Rev. 27, 2, 217–236. DOI: https://doi.org/10.3758/s13423-019-01689-4.

F. Ciardo, F. Beyer, D. De Tommaso, and A. Wykowska. 2020. Attribution of intentional agency towards robots reduces one's own sense of agency. Cognition 194, 104109. DOI: https://doi.org/10. 1016/j.cognition.2019.104109.

M. Corbetta. 1998. Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural systems? Proc Nat Ac Sci, 95 (3) 831-838; DOI: 10.1073/pnas.95.3.831

N. Dahlbäck, A. Jönsson, and L. Ahrenberg. 1993. Wizard of Oz studies: Why and how. In Proceedings of the 1st International Conference on Intelligent user Interfaces. (1993 February) 193–200.

K. Dautenhahn, C. L. Nehaniv. 2002. Imitation as a Dual-Route Process Featuring Predictive and Learning Components: A Biologically Plausible Computational Model. In Imitation in Animals and Artifacts , MIT Press, pp.327-361.

P. Dayan and N. D. Daw. 2008. Decision theory, reinforcement learning, and the brain. Cogn. Affect. Behav. Neurosci. 8, 429–453. DOI: https://doi.org/10.3758/CABN.8.4.429.

M. M. A. De Graaf and B. F. Malle. 2019. People's explanations of robot behavior subtly reveal mental state inferences, 2019 14th ACM/IEEE International Conference on Human–Robot

Interaction (HRI), Daegu, Korea (South). 239–248. DOI: https://doi.org/10.1109/HRI.2019.8673308.

M. M. A. De Graaf, S. B. Allouch, and S. Lutfi. 2016. What are people's associations of domestic robots? Comparing implicit and explicit measures. In 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN). IEEE, 1077–1083.

L. M. J. De Sonneville, C. A. Verschoor, C. Njiokiktjien, V. Op het Veld, N. Toorenaar, and M. Vranken. 2002. Facial identity and facial emotions: Speed, accuracy, and processing strategies in children and adults. J. Clin. Exp. Neuropsychol. 24, 2, 200–213. DOI: https://doi.org/10.1076/jcen .24.2.200.989.

Y. Demiris and B. Khadhouri. 2006. Hierarchical attentive multiple models for execution and recognition of actions. Rob. Auton. Syst. 54, 361–369. DOI: https://doi.org/10.1016/j.robot.2006.02.003.

D. C. Dennett. 1971. Intentional systems. J. Philos. 68, 87–106. DOI: https://doi.org/10.2307/2025382.

D. C. Dennett. 1987. The Intentional Stance. MIT Press.

S. Devin and R. Alami. 2016. An implemented theory of mind to improve human–robot shared plans execution. In ACM/IEEE International Conference on Human–Robot Interaction. DOI: https://doi.org/10.1109/HRI.2016.7451768.

J. Driver, G. Davis, P. Ricciardelli, P. Kidd, E. Maxwell, and S. Baron-Cohen. 1999. Gaze perception triggers reflexive visuospatial orienting. Vis. Cogn. 6, 5, 509–540. DOI: https://doi.org/10.1080/ 135062899394920.

B. R. Duffy. 2003. Anthropomorphism and the social robot. In Robotics and Autonomous Systems. DOI: https://doi.org/10.1016/S0921-8890(02)00374-3.

N. J. Emery. 2000. The eyes have it: The neuroethology, function and evolution of social gaze. Neurosci. Biobehav. Rev. 24, 581–604. DOI: https://doi.org/10.1016/S0149-7634(00)00025-7.

N. Epley, A. Waytz, and J. T. Cacioppo. 2007. On seeing human: A three-factor theory of anthropomorphism. Psychol. Rev. 114, 864–886. DOI: https://doi.org/10.1037/0033-295X.114.4.864.

F. Eyssel, F. Hegel, G. Horstmann, and C. Wagner. 2010. Anthropomorphic inferences from emotional nonverbal cues: A case study. In Proceedings—IEEE International Workshop on Robot and Human Interactive Communication. DOI: https://doi.org/10.1109/ROMAN.2010.5598687.

A. Fod, M. J. Mataric´, and O. C. Jenkins. 2002. Automated derivation of primitives for movement classification. Auton. Rob. 12, 39–54. DOI: https://doi.org/10.1023/A:1013254724861.

C. K. Friesen and A. Kingstone. 1998. The eyes have it! Reflexive orienting is triggered by nonpredictive gaze. Psychon. Bull. Rev. 5, 3, 490–495. DOI: https://doi.org/10.3758/BF03208827.

A. Frischen, A. P. Bayliss, and S. P. Tipper. 2007. Gaze cueing of attention: Visual attention, social cognition, and individual differences. Psychol. Bull. 133, 4, 694–724. DOI: https://doi.org/ 10.1037/ 0033-2909.133.4.694.

K. Friston. 2005. A theory of cortical responses. Philos. Trans. R. Soc. Lond. B Biol. Sci. 360, 1456, 815–836. DOI: https://doi.org/10.1098/rstb.2005.1622.

C. D. Frith and U. Frith. 2006. The neural basis of mentalizing. Neuron 50, 4, 531–534. DOI: https:// doi.org/10.1016/j.neuron.2006.05.001.

U. Frith, J. Morton, and A. M. Leslie. 1991. The cognitive basis of a biological disorder: Autism. Trends Neurosci. 14, 10, 433–438. DOI: https://doi.org/10.1016/0166-2236(91)90041-r.

S. R. Fussell, S. Kiesler, L. D. Setlock, and V. Yew. 2008. How people anthropomorphize robots. In

HRI 2008—Proceedings of the 3rd ACM/IEEE International Conference on Human–Robot Interaction: Living with Robots. DOI: https://doi.org/10.1145/1349822.1349842.

H. L. Gallagher, A. I. Jack, A. Roepstorff, and C. D. Frith. 2002. Imaging the intentional stance in a competitive game. NeuroImage 16, 814–821. DOI: https://doi.org/10.1006/nimg.2017.

N. George, J. Driver, and R. J. Dolan. 2001. Seen gaze-direction modulates fusiform activity and its coupling with other brain areas during face processing. NeuroImage 13, 1102–1112. DOI: https://do i.org/10.1006/nimg.2001.0769.

D. Ghiglino, C. Willemse, D. De Tommaso, F. Bossi, A. Wykowska. 2020. At first sight: robots' subtle eye movement parameters affect human attentional engagement, spontaneous attunement and perceived human-likeness. Paladyn. Journal of Behavioral Robotics, 11:31-39

O. Golan, S. Baron-Cohen, J. J. Hill, and Y. Golan. 2006. The "reading the mind in films" task: Complex emotion recognition in adults with and without autism spectrum conditions. Soc. Neurosci. 1, 2, 111–123. DOI: https://doi.org/10.1080/17470910600980986.

B. Gonsior, S. Sosnowski, C. Mayer, J. Blume, B. Radig, D. Wollherr, and K. Kuhnlenz. 2011. Improving aspects of empathy and subjective performance for HRI through mirroring facial expressions. In Proceedings—IEEE International Workshop on Robot and Human Interactive Communication. DOI: https://doi.org/10.1109/ROMAN.2011.6005294.

O. C. Görür, B. S. Rosman, G. Hoffman, and S. Albayrak. 2017. Toward integrating theory of mind into adaptive decision-making of social robots to understand human intention. 12th ACM/ IEEE International Conference on Human–Robot Interaction (HRI).

J. Gray, C. Breazeal, M. Berlin, A. Brooks, and J. Lieberman. 2005. Action parsing and goal inference using self as simulator. In ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, (August 2005). IEEE, 202–209.

H. M. Gray, K. Gray, D. M. Wegner. 2007. Dimensions of mind perception. Science 315, 5812, 619. DOI: https://doi.org/10.1126/science.1134475.

A. N. Hampton, P. Bossaerts, and J. P. O'Doherty. 2008. Neural correlates of mentalizing-related computations during strategic interactions in humans. Proc. Nat. Acad. Sci. 105, 18, 6741–6746. DOI: https://doi.org/10.1073/pnas.0711099105.

F. G. Happé. 1994. An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. J. Autism Dev. Disord. 24, 2, 129–154. DOI: https://doi.org/10.1007/BF02172093.

F. Happé, and U. Frith. 2014. Annual research review: Towards a developmental neuroscience of atypical social cognition. J. Child Psychol. Psychiatry 55, 553–577. DOI: https://doi.org/ 10.1111/jcpp.12162.

F. Hegel, S. Krach, T. Kircher, B. Wrede and G. Sagerer. 2008. Theory of mind (ToM) on robots: A functional neuroimaging study, 2008 3rd ACM/IEEE International Conference on Human–Robot Interaction (HRI), Amsterdam, 335–342. DOI: https://doi.org/ 110.1145/1349822.1349866.

F. Heider. 1958. The Psychology of Interpersonal Relations. Psychology Press. DOI: https:// doi.org/10. 4324/9780203781159.

N. Hertz and E. Wiese. 2018. Under pressure: Examining social conformity with computer and robot groups. Hum. Factors 60, 8, 1207–1218. DOI: https://doi.org/ 10.1177/0018720818788473.

E. A. Hoffman and J. V. Haxby. 2000. Distinct representations of eye gaze and identity in the distributed human neural system for face perception. Nat. Neurosci. 3, 80–84. DOI: https:// doi.org/10. 1038/71152.

C. I. Hooker, K. A. Paller, D. R. Gitelman, T. B. Parrish, M. M. Mesulam, and P. J. Reber. 2003. Brain networks for analyzing eye gaze. Cogn. Brain Res. 17, 2, 406–418. DOI: https:// doi.org/10.1016/ S0926-6410(03)00143-5.

C. M. Huang and B. Mutlu. 2012. Robot behavior toolkit: Generating effective social behaviors for robots. In HRI'12—Proceedings of the 7th Annual ACM/IEEE International Conference on Human– Robot Interaction. DOI: https://doi.org/10.1145/2157689.2157694.

B. Huebner. 2010. Commonsense concepts of phenomenal consciousness: Does anyone care about functional zombies? Phenomenol. Cogn. Sci. 9, 133–155. DOI: https://doi.org/10.1007/ s11097-00991266.

A. C. Jenkins, C. N. Macrae, and J. P. Mitchell. 2008. Repetition suppression of ventromedial prefrontal activity during judgments of self and others. Proc. Natl. Acad. Sci. U. S. A. 105, 11, 4507– 4512. DOI: https://doi.org/10.1073/pnas.0708785105.

M. Johnson and Y. Demiris. 2005. Perceptual perspective taking and action recognition. Int. J. Adv. Rob. Syst. 2, 4, 32. DOI: https://doi.org/https://doi.org/10.5772/5775.

S. Kahl, & S. Kopp. 2018. A predictive processing model of perception and action for self-other distinction. Front Psych, 9, 2421.

C. D. Kidd and C. Breazeal. 2004. Effect of a robot on user perceptions. In 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). DOI: https://doi.org/ 10.1109/iros .2004.1389967.

S. Kiesler and J. Goetz. 2002. Mental models and cooperation with robotic assistants. CHI'02 Extended Abstracts on Human Factors in Computing Systems. DOI: https://doi.org/10.1145/506443.506491.

K. Kompatsiari, V. Tikhanoff, F. Ciardo, G. Metta, and A. Wykowska. 2017. The importance of mutual gaze in human–robot interaction. In A. Kheddar, E. Yoshida, S. S. Ge, K. Suzuki, J.-J. Cabibihan, F. Eyssel, and H. He (Eds.), Social Robotics. Springer International Publishing, Cham, 443–452.

K. Kompatsiari, J. Perez-Osorio, D. De Tommaso, G. Metta, and A. Wykowska. 2018. Neuroscientificallygrounded research for improved human–robot interaction. In IEEE International Conference on Intelligent Robots and Systems. DOI: https://doi.org/10.1109/IROS.2018.8594441.

K. Kompatsiari, F. Bossi & A. Wykowska. 2021. Eye contact during joint attention with a humanoid robot modulates oscillatory brain activity. Soc Cog Aff Neu (SCAN), 16, 383-392.

S. Krach, F. Hegel, B. Wrede, G. Sagerer, F. Binkofski, & T. Kircher. 2008. Can machines think? Interaction and perspective taking with robots investigated via fMRI. PloS one, 3(7), e2597.

G. Kuhn, I. Vacaityte, A. D. C. D'Souza, A. C. Millett, and G. G. Cole. 2018. Mental states modulate gaze following, but not automatically. Cognition 80, 1–9. DOI: https://doi.org/10.1016/j.cognition. 2018.05.020.

A. M. Leslie. 1987. Pretense and representation: The origins of "theory of mind." Psychol. Rev. 94, 4, 412-426. DOI: https://doi.org/10.1037/0033-295X.94.4.412.

J. Li. 2015. The benefit of being physically present: A survey of experimental works comparing copresent robots, telepresent robots and virtual agents. Int. J. Hum. Comput. Stud. 77, 23-37. DOI: https: //doi.org/10.1016/j.ijhcs.2015.01.001.

S. Marchesi, D. Ghiglino, F. Ciardo, J. Perez-Osorio, E. Baykara, and A. Wykowska. 2019. Do we adopt the intentional stance toward humanoid robots? Front. Psychol. 10. DOI: https://doi.org/10. 3389/fpsyg.2019.00450.

G. Metta, L. Natale, F. Nori, G. Sandini, D. Vernon, L. Fadiga, ... L. Montesano. 2010. The iCub humanoid robot: An open-systems platform for research in cognitive development. Neural Netw. 23, Q14 1125–1134. DOI: https://doi.org/10.1016/j.neunet.2010.08.010.

J. P. Mitchell, T. F. Heatherton, and C. N. Macrae. 2002. Distinct neural systems subserve person and object knowledge. Proc. Natl. Acad. Sci. U. S. A. 99, 23, 15238–15243. DOI: https://doi.org/10. 1073/pnas.232395699.

B. Mutlu, F. Yamaoka, T. Kanda, H. Ishiguro, and N. Hagita. 2009. Nonverbal leakage in robots: Communication of intentions through seemingly unintentional behavior. In Proceedings of the 4th ACM/IEEE International Conference on Human Robot Interaction. (March 2009). 69–76.

B. Mutlu, A. Terrell, and C. Huang. 2013. Coordination mechanisms in human–robot collaboration. In Proceedings of the HRI 2013 Workshop on Collaborative Manipulation.

Y. Nagai, M. Asada, and K. Hosoda. 2002. Developmental learning model for joint attention. In: IEEE International Conference on Intelligent Robots and Systems.

S. Narang, A. Best, and D. Manocha. 2019. Inferring user intent using Bayesian theory of mind in shared avatar–agent virtual environments. IEEE Trans. Vis. Comput. Graph. 25, 5, (May 2019). 2113–2122. DOI: https://doi.org/10.1109/TVCG.2019.2898800.

A. Newell. 1994. Unified Theories of Cognition. Harvard University Press.

A. C. Nobre, G. N. Sebestyen, D. R. Gitelman, M. M. Mesulam, R. S. Frackowiak, and C. D. Frith. 1997. Functional localization of the system for visuospatial attention using positron emission tomography. Brain 120, 3, 515–533. DOI: https://doi.org/10.1093/brain/120.3.515.

S. Nowicki Jr, and J. Carton. 1993. The measurement of emotional intensity from facial expressions. J. Soc. Psychol. 133, 5, 749–750. DOI: https://doi.org/10.1080/00224545.1993.9713934.

R. O'Reilly, T. Hazy, and S. Herd. 2012. The Leabra Cognitive Architecture: How to Play 20 Principles with Nature and Win! The Oxford Handbook of Cognitive Science. DOI: https://doi.org/10. 1093/oxfordhb/9780199842193.013.8.

T. Ojala, M. Pietikainen, and T. Maenpaa. 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Trans. Pattern Anal. Mach. Intell. 24, 971–987. DOI: https://doi.org/10.1109/TPAMI.2002.1017623.

Y. Okumura, Y. Kanakogi, T. Kanda, H. Ishiguro, and S. Itakura. 2013. Infants understand the referential nature of human gaze but not robot gaze. J. Exp. Child Psychol. 116, 1, 86–95. DOI: https://doi. org/10.1016/j.jecp.2013.02.007.

P. C. Pantelis, L. Byrge, J. M. Tyszka, R. Adolphs, and D. P. Kennedy. 2015. A specific hypoactivation of right temporo-parietal junction/posterior superior temporal sulcus in response to socially awkward situations in autism. Soc. Cogn. Affect. Neurosci. 10, 10, 1348–1356. DOI: https://doi.org/10. 1093/scan/nsv021.

K. Pelphrey, J. Singerman, T. Allison, and G. McCarthy. 2003. Brain activation evoked by perception of gaze shifts: The influence of context. Neuropsychologia 41, 156–170. DOI: https://doi.org/10. 1016/s0028-3932(02)00146-x.

J. Perez-Osorio and A. Wykowska. 2019. Adopting the intentional stance toward natural and artificial agents. Philos. Psychol. 33, 1-27. DOI: https://doi.org/10.1080/09515089.2019.1688778.

J. Perez-Osorio, H. J. Müller, E. Wiese, and A. Wykowska. 2015. Gaze following is modulated by expectations regarding others' action goals. PLoS One 10, e0143614. DOI: https://doi.org/10.1371/ journal.pone.0143614.

J. Perez-Osorio, H. J. Möller, and A. Wykowska. 2017. Expectations regarding action sequences modulate electrophysiological correlates of the gaze-cueing effect. Psychophysiology 54, 942–954. DOI: https://doi.org/10.1111/psyp.12854 .

J. Perez-Osorio, D. De Tommaso, E. Baykara, and A. Wykowska. 2018. Joint action with iCub: A successful adaptation of a paradigm of cognitive neuroscience to HRI. In 27th IEEE

International Symposium on Robot and Human Interactive Communication (RO-MAN), Nanjing – Tai'an, 6 Pages. DOI: https://doi.org/10.1109/ROMAN.2018.8525536.

J. Perez-Osorio, S. Marchesi, D. Ghiglino, M. Ince, and A. Wykowska. 2019. More than you expect: Priors influence on the adoption of intentional stance toward humanoid robots. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). DOI: https://doi.org/10.1007/978-3-030-35888-4_12.

M. Petit, S. Lallée, J. D. Boucher, G. Pointeau, P. Cheminade, D. Ognibene, E. Chinellato, U. Pattacini, I. Gori, U. Martinez-Hernandez, H. Barron-Gonzalez, M. Inderbitzin, A. Luvizotto, V. Vouloutsi, Y. Demiris, G. Metta, and P. F. Dominey. 2013. The coordinating role of language in realtime multimodal learning of cooperative tasks. IEEE Trans. Auton. Ment. Dev. 5, 3–17. DOI: https: //doi.org/10.1109/TAMD.2012.2209880.

K. S. Ponnet, H. Roeyers, A. Buysse, A. de Clercq, and E. van der Heyden. 2004. Advanced mindreading in adults with Asperger syndrome. Autism 8, 249–266. DOI: https://doi.org/ 10.1177/ 1362361304045214.

D. Premack and G. Woodruff. 1978. Premack and Woodruff: Chimpanzee theory of mind. Behav. Brain Sci. 4, 1978, 515–526. DOI: http://dx.doi.org/10.1017/S0140525X00076512.

D. Premack. 1990. The infant's theory of self-propelled objects. Cognition, 36(1), 1-16.

A. Puce, T. Allison, S. Bentin, J. C. Gore, and G. McCarthy. 1998. Temporal cortex activation in humans viewing eye and mouth movements. J. Neurosci. 18, 6, 2188–2199. DOI: https:// doi.org/10. 1523/JNEUROSCI.18-06-02188.1998.

N. C. Rabinowitz, F. Perbet, H. F. Song, C. Zhang, and M. Botvinick. 2018. Machine theory of mind. In 35th International Conference on Machine Learning, ICML 2018.

H. Richardson, G. Lisandrelli, A. Riobueno-Naylor, and R. Saxe. 2018. Development of the social brain from age three to twelve years. Nat. Commun. 9, 1, 1–12. DOI: https://doi.org/10.1038/ s41467-018-03399-2.

L. D. Riek, T. C. Rabinowitch, B. Chakrabartiz, and P. Robinson. 2009. Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. In 2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops, ACII 2009. DOI: https:// doi.org/10.1109/AC II.2009.5349423.

A. Sadeghipour and S. Kopp. 2011. Embodied gesture processing: Motor-based integration of perception and action in social artificial agents. Cognit. Comput. 3, 3, 419–435. DOI: https:// doi.org/10. 1007/s12559-010-9082-z.

R. Saxe and N. Kanwisher. 2003. People thinking about thinking people: The role of the temporoparietal junction in "theory of mind." NeuroImage 19, 4, 1835–1842. DOI: https:// doi.org/10.1016/ s1053-8119(03)00230-1.

B. M. Scassellati. 2001. Foundations for a Theory of Mind for a Humanoid Robot. MIT Press. DOI: https://doi.org/10.1037/e446982006-001.

B. Scassellati. 2002. Theory of mind for a humanoid robot. Auton. Robots 12, 1, 13–24. DOI: https:// doi.org/10.1023/A:1013298507114.

S. Schaal. 1997. Learning from demonstration. In Advances in Neural Information Processing Systems. DOI: https://doi.org/10.1007/978-1-4419-1428-6_4646.

K. R. Scherer and U. Scherer. 2011. Assessing the ability to recognize facial and vocal expressions of emotion: Construction and validation of the Emotion Recognition Index. J. Nonverbal Behav. 35, 4, 305. DOI: https://doi.org/10.1007/s10919-011-0115-4.

L. Schilbach, B. Timmermans, V. Reddy, A. Costall, G. Bente, T. Schlicht, and K. Vogeley. 2013. Toward a second-person neuroscience. Behav. Brain Sci. 36, 4, 393–414. DOI: https:// doi.org/10. 1017/S0140525X12000660.

J. L. Schreck, O. B. Newton, J. Song, and S. M. Fiore. 2019. Reading the mind in robots: How theory of mind ability alters mental state attributions during human–robot interactions. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 63, 1, 1550–1554. DOI: https://doi.org/ 10.1177/1071181319631414.

T. Singer. 2012. The past, present and future of social neuroscience: A European perspective. NeuroImage 61, 2, 437–449. DOI: https://doi.org/10.1016/j.neuroimage.2012.01.109.

M. A. Smith, M. M. Allaham, and E. Wiese. 2016. Trust in automated agents is modulated by the combined influence of agent and task type. Proc. Hum. Factors Ergon. Soc. Annu. Meet. 60, 1, 206–210. DOI: https://doi.org/10.1177/1541931213601046.

N. Spatola, and A. Wykowska. 2021. The personality of anthropomorphism: How the need for cognition and the need for closure define attitudes and anthropomorphic attributions toward robots. Comp Hum Beh, 122. https://doi.org/10.1016/j.chb.2021.106841.

S. Sul, B. Güroğ̆lu, E. A. Crone, and L. J. Chang. 2017. Medial prefrontal cortical thinning mediates shifts in other-regarding preferences during adolescence. Sci. Rep. 7, 1, 1–10. DOI: https:// doi.org/ 10.1038/s41598-017-08692-6.

L. Takayama, D. Dooley, and W. Ju. 2011. Expressing thought: Improving robot readability with animation principles. In Proceedings of the ACM/IEEE International Conference on Human–Robot Interaction (HRI). (March 2011). ACM Press. Lausanne, Switzerland, 69–76.

C. Teufel, P. C. Fletcher, and G. Davis. 2010. Seeing other minds: Attributed mental states influence perception. Trends Cognit. Sci. 14, 8, 376–382. DOI: https://doi.org/10.1016/ j.tics.2010.05.005.

S. Thellman, A. Silvervarg, and T. Ziemke. 2017. Folk-psychological interpretation of human vs. humanoid robot behavior: Exploring the intentional stance toward robots. Front. Psychol. 8, 1962. DOI: https://doi.org/10.3389/fpsyg.2017.01962.

J. G. Trafton, A. C. Schultz, N. L. Cassimatis, L. M. Hiatt, D. Perzanowski, D. P. Brock, M. D. Bugajska, F. E. Mintz, and W. Adams. 2006. Communicating and collaborating with robotic agents. Cognition and multi-agent interaction: From cognitive modeling to social simulation. Sun, R. (Ed.). Cambridge University Press, 252-278.

F. Van Overwalle and K. Baetens. 2009. Understanding others' actions and goals by mirror and mentalizing systems: A meta-analysis. NeuroImage 48, 3, 564–584. DOI: https://doi.org/ 10.1016/j.neuroi mage.2009.06.009.

K. E. Vanderbilt, D. Liu, and G. D. Heyman. 2011. The development of distrust. Child Dev. 82, 5, 1372–1380. DOI: https://doi.org/10.1111/j.1467-8624.2011.01629.x.

V. Venkatesh and F. D. Davis. 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. Manage. Sci. 46, 2, 186–204. DOI: https://doi.org/10.1287/mnsc.46.2. 186.11926.

P. F. Verschure, T. Voegtlin, and R. J. Douglas. 2003. Environmentally mediated synergy between perception and behaviour in mobile robots. Nature 425, 620–624. DOI: https://doi.org/10.1038/nature 02024.

S. Vinanzi, M. Patacchiola, A. Chella, and A. Cangelosi. 2019. Would a robot trust you? Developmental robotics model of trust and theory of mind. Philos. Trans. R. Soc. B. 374, 1771, 20180032. DOI: https://doi.org/10.1098/rstb.2018.0032.

P. Viola and M. Jones. 2001. Rapid object detection using a boosted cascade of simple features. In Computer Vision and Pattern Recognition, 2001. Proceedings of the 2001 IEEE Computer Society Conference on CVPR 2001, Vol. 1. IEEE, New York, NY, I–511.

H. Wellman. 1990. Children's Theories of Mind. MIT Press, Cambridge, MA.
B. Wicker, F. Michel, M. A. Henaff, and J. Decety. 1998. Brain regions involved in the perception of gaze: A PET study. NeuroImage 8, 2, 221–227. DOI: https://doi.org/10.1006/nimg.1998.0357.

E. Wiese, A. Wykowska, J. Zwickel, and H. J. Möller. 2012. I see what you mean: How attentional selection is shaped by ascribing intentions to others. PLoS One 7, 9, e45391. DOI: https://doi.org/ 10.1371/journal.pone.0045391.

E. Wiese, J. Zwickel, and H. J. Müller. 2013. The importance of context information for the spatial specificity of gaze cueing. Atten. Percept. Psychophys. 75, 967–982. DOI: https://doi.org/10.3758/ s13414-013-0444-y.

E. Wiese, A. Wykowska, and H. J. Müller. 2014. What we observe is biased by what other people tell us: Beliefs about the reliability of gaze behavior modulate attentional orienting to gaze cues. PLoS One 9, 4. DOI: https://doi.org/10.1371/journal.pone.0094529.

E. Wiese, A. Abubshait, B. Azarian, and E. J. Blumberg. 2019. Brain stimulation to left prefrontal cortex modulates attentional orienting to gaze cues. Philos. Trans. R. Soc. B 374, 1771, 20180430. DOI: https://doi.org/10.1098/rstb.2018.0430.

C. Willemse and A. Wykowska. 2019. In natural interaction with embodied robots, we prefer it when they follow our gaze: A gaze-contingent mobile eyetracking study. Philos. Trans. R. Soc. B 374, 1771, 20180036. DOI: https://doi.org/10.1098/rstb.2018.0036.

C. Willemse, S. Marchesi, and A. Wykowska. 2018. Robot faces that follow gaze facilitate attentional engagement and increase their likeability. Front. Psychol. 9, 70. DOI: https://doi.org/10.3389/fpsyg. 2018.00070.

M. Wilms, L. Schilbach, U. Pfeiffer, G. Bente, G. R. Fink, and K. Vogeley. 2010. It's in your eyes—Using gaze-contingent stimuli to create truly interactive paradigms for social cognitive and affective neuroscience. Soc. Cogn. Affect. Neurosci. 5, 98–107. DOI: https://doi.org/10.1093/scan/nsq024.

H. Wimmer and J. Perner. 1983. Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. Cognition 13, 1, 103–128. DOI: http s://doi.org/10.1016/0010-0277(83)90004-5.

A. Wykowska, E. Wiese, A. Prosser, and H. J. Müller. 2014. Beliefs about the minds of others influence how we process sensory information. PLoS One 9, 4, e94339.

A. Wykowska, J. Kajopoulos, M. Obando-Leitón, S. S. Chauhan, J. J. Cabibihan, and G. Cheng. 2015. Humans are well tuned to detecting agents among non-agents: Examining the sensitivity of human perception to behavioral characteristics of intentional systems. Int. J. Soc. Rob. 7. DOI: https://doi.or g/10.1007/s12369-015-0299-6.

A. Wykowska, 2021. Robots as mirrors of the human mind. Current Directions in Psychological Science, 30 (1), 34-40.

A. Wykowska, 2020. Social robots to test flexibility of human social cognition. International Journal of Social Robotics, 12, 1203–1211

T. Xu, H. Zhang, and C. Yu. 2016. See you see me: The role of eye contact in multimodal human–robot interaction. ACM Trans. Interact. Intell. Syst. (TiiS) 6, 1, 1–22. DOI: https://doi.org/ 10.1145/ 2882970.

J. Złotowski, D. Proudfoot, K. Yogeeswaran, and C. Bartneck. 2015. Anthropomorphism: Opportunities and challenges in human–robot interaction. Int. J. Soc. Rob. 7, 347–360. DOI: https://doi.org/ 10.1007/s12369-014-0267-6.