


Review

Seeing social interactions

Emalie McMahon¹ and Leyla Isik ,^{1,2,*}

Seeing the interactions between other people is a critical part of our everyday visual experience, but recognizing the social interactions of others is often considered outside the scope of vision and grouped with higher-level social cognition like theory of mind. Recent work, however, has revealed that recognition of social interactions is efficient and automatic, is well modeled by bottom-up computational algorithms, and occurs in visually-selective regions of the brain. We review recent evidence from these three methodologies (behavioral, computational, and neural) that converge to suggest the core of social interaction perception is visual. We propose a computational framework for how this process is carried out in the brain and offer directions for future interdisciplinary investigations of social perception.

Recognizing core components of a social interaction

Sitting in a coffee shop, you see two people walk in. Immediately, you know that they are together and that they are arguing even though you cannot hear their voices and their faces are largely obscured. Incredibly complex computations are required to recognize and understand this **social interaction** (see [Glossary](#)), but growing evidence suggests that the most fundamental of these abilities rely primarily on visual computations.

Recognizing others' social interactions is a critical part of the human experience. This ability develops early in infancy [1] and guides how we act in the social world. In the first year of life, infants draw inferences from observed social interactions that inform their preferences [1] and potential social partners [2]. Non-human primates can also recognize social interactions even in minimal visual displays [3] and base important decisions about kinship and hierarchy on these observations [4].

We have known since the early work of Heider and Simmel [5] that this rich information about social interactions can be extracted from simple visual cues. However, these abilities have primarily been discussed in the context of recognizing each individual agent as being animate [6,7] or goal directed [8–10] rather than recognizing the interactions between agents. In fact, the ability to recognize social interactions is usually grouped with higher-level aspects of social cognition, like mentalization [11–14] and thought to require complex social inference. However, emerging evidence suggests that recognizing social interactions is fast, automatic, evolutionarily adaptive, and thus unlikely to rely solely on complex mental models. While there is clearly a mentalistic aspect to understanding others' social interactions, here we argue that recognizing **core components of social interactions** is visual in nature.

This core recognition includes detecting interactions (based on both physical and communicative contingencies) and extracting their valence and goal compatibility (e.g., cooperation vs. conflict or helping vs. hindering). These aspects of social interactions can be considered analogous to core object recognition [15] or scene gist [16]. Critically, they go beyond visual **social primitives** such as the distance between agents, the extent to which they are facing, and their contingent motion (Figure 1). Some evidence even suggests that higher-level information about the meaning

Highlights

Recognizing the social interactions of others is fundamental to everyday life. We argue that core components of a social interaction can be extracted by the human visual system.

The visual system represents visual precursors of social interactions, social interactions themselves, and even some higher-level features of interactions.

Computational neural network models of social interaction recognition match human judgements using only visual information and bottom-up architectures.

Social interactions activate visual regions of the brain that are functionally dissociated from other social visual stimuli and from social cognitive processes like theory of mind.

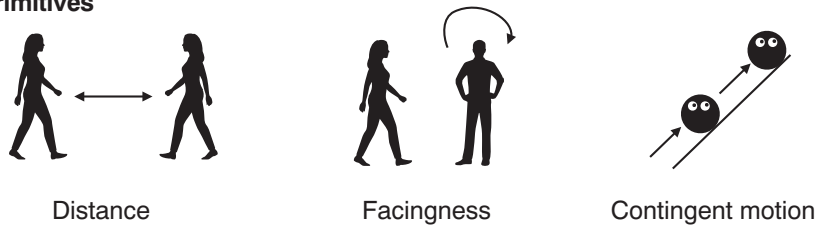
Applying a vision science approach to social interaction recognition can advance our computational and mechanistic understanding of this critical human ability.

¹Department of Cognitive Science, Johns Hopkins University, Baltimore, MD, USA

²Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD, USA

*Correspondence: lisik@jhu.edu (L. Isik).

Social primitives



Core components



Trends In Cognitive Sciences

Figure 1. Social interaction representations in the visual system. We argue that the visual system represents not only social primitive of social interactions (top row), including inter-agent distance, facingness, and contingent motion, but also higher-level ‘core components’ of a social interaction (bottom row), including detecting physical and communicative interactions and valence or goal compatibility. The information in each category is roughly ordered by increasing complexity from left to right.

of an interaction, including the type of interaction (e.g., talking vs. dancing) and social roles (e.g., agent vs. patient), may also be extracted visually. However, explicit representations of people’s mental states (e.g., wondering why the couple in the coffee shop is fighting) rely on theory of mind and thus fall outside of this definition [17].

Importantly, when we discuss visual information, we are not simply talking about low-level sensory features like contrast or motion. We also do not mean information that is extracted by the visual system and then processed cognitively. Instead, we are referring to an exciting middle ground that suggests our visual system contains rich, abstract representations of social interaction that go beyond low-level correlates of these representations [18]. We believe the visual system forms these representations based on hierarchical computations of visuospatial and motion cues; thus our arguments apply to social interactions that are based on relatively fast spatiotemporal contingencies, occurring on the order of less than one second, and happen in close physical space. Interactions like collaborating on a work project (long time-scale) or texting (spatially far) fall outside the scope of our arguments. Our arguments also focus specifically on social interactions, which differ from both object–object and person–object interactions in terms of low-level visual properties (e.g., social interactions are often highly dynamic and involve different motion patterns than physical interactions) and subsequent high-level processing.

We review interdisciplinary evidence in support of visual representations of social interactions across different levels of analysis [19]. First, a growing body of behavioral work suggests that recognizing social interactions is a computational goal of the visual system. Algorithmically, emerging evidence suggests that human social interaction perception is well modeled by **bottom-up**, discriminative models. Neurally, social interactions are processed in a manner that is hierarchical like other aspects of vision, overlapping with visual processing streams,

Glossary

Bottom-up: bottom-up processing is direct mapping between a stimulus and a particular feature of interest without generative processing or external top-down knowledge. This does not entail the absence of recurrent processing within a processing stage, distinguishing bottom-up from strictly feedforward processing.

Core components of social interactions: detecting the presence of a social interaction based on physical or communicative contingencies. This definition extends beyond visual social primitives and also includes extracting some features of the interaction such as valence or goal compatibility.

Facing dyads: minimal images of human bodies facing towards one another without contextual information.

Generative inverse planning: a computational framework that aims to recognize information about agents by inverting an internal model of the physical and social world.

Inductive bias: built-in assumptions or knowledge a learner uses for more effective generalization.

Second-person social interactions: a social interaction involving the participant.

Social chunking: grouping of two or more agents (depicted as bodies, faces, or point light figures) into a single unit during visual recognition.

Social interactions: actions between two or more people that are directed at and contingent upon each other. Unless specified, we focus here on third-person interactions (i.e., observations of others’ interactions, not involving the participant).

Social primitives: visual cues to the presence of a social interaction, which include inter-agent distance, whether agents are facing one another, and their motion congruency.

Sociality: the extent to which a person’s action is directed at another person.

and distinct from the theory of mind network. This converging evidence suggests that seemingly high-level aspects of social interactions are extracted by largely visual processes.

Visual signatures of social interaction recognition

The distinction between vision and cognition has been discussed extensively [7,20]. While visual percepts and higher-level judgements often covary, perception has many distinctive behavioral hallmarks, including rapid and automatic processing, attentional capture, efficient visual search, and recognition advantages. Using these approaches, researchers have found that many high-level attributes, including animacy [21,22] and causality [23], may in fact be extracted perceptually, and recent evidence has argued that such visual processing occurs not only for single entities but also their relations [24]. In the following sections, we outline the behavioral evidence that social relations in particular are processed visually.

Visual processing of social primitives

Recognizing social interactions requires first detecting two or more animate agents [25] and then next detecting visual cues indicative of social interactions such as whether agents are facing one another [26–28] and their motion congruency [21,29]. We have termed these social visual cues social primitives. While detecting at least one social primitive is necessary for recognizing a social interaction, social primitives alone do not constitute an interaction.

A fast-growing body of work has shown **facing dyads** (bodies and faces) are subject to many effects that are hallmarks of perceptual processing. Facing bodies are recognized significantly better than bodies facing away from each other and are subject to an inversion effect where their recognition is significantly diminished when inverted [26] (Figure 2A). Facing dyads are also found more quickly in visual search tasks [27,30,31] and better remembered in subsequent memory tasks [30,32]. Together, these results suggest that facing dyads are processed preferentially as a visual unit (for more discussion and review see [33]).

Motion congruency between agents also indicates the presence of a social interaction, and when congruency is disrupted, the perception of the social interaction is as well. For instance, when congruency between agents' limb movements and the coordinated motion between agents is disrupted, descriptions shift from those of a social interaction (dancing) to an inanimate description (the 'dots were kind of paired up and they drifted to the left and right simultaneously') [29]. Further, the perception of chasing in displays of one arrow following another is highly sensitive to the degree of similarity of the agents' motion [21].

Social interactions are detected visually

Detecting social interactions also shares many behavioral signatures of visual processing. Social interactions have priority access in attention and working memory. Interacting dyads are the predominant percept in binocular rivalry tasks [34] (Figure 2B), viewed faster and longer in free viewing of natural images [35] (Figure 2C), and chunked as a single unit in attentional cueing and working memory tasks [36–39]. Further, **social chunking** does not require verbal labeling [38], happens automatically [39], and results in visual adaptation [40] (Figure 2D). Importantly, these effects are enhanced by the presence of meaningful social interactions (e.g., talking and fighting) in particular, not simply the agents being close or facing one another.

So far, we have reviewed evidence that the perception of social interactions has characteristically visual effects, including the influence of visual manipulations on social interaction perception. However, because visual perception is encapsulated from other aspects of cognition [20,41], the reverse argument can also be made: if a social interaction influences other visual percepts,

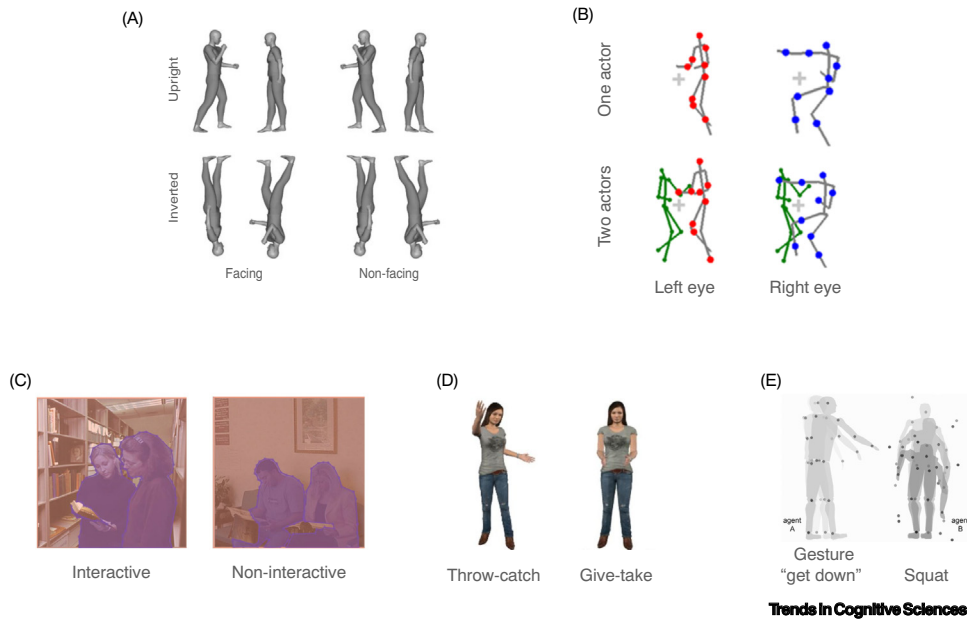


Figure 2. Visual effects for social interactions are observed across a range of stimuli and paradigms. (A) Facing dyads (top left) are recognized better than non-facing dyads (top right), but this advantage disappears when stimuli are inverted (bottom row) [26]. (B) Social interactions conveyed via physically contingent motion are the predominant stimulus perceived during a binocular rivalry task [42]. (C) In free viewing of natural images, people interacting (blue, left) are looked at faster and longer than objects (orange) or pairs of people not interacting (blue, right) [35]. (D) Adaptation effects are observed not only when someone views repeated presentations of a given action but also for the contingent social action of the adapted action, like throw and catch or give and take [40]. (E) Communicative interactions conveyed by point light dyads are recognized more accurately than individual actions [44].

it is evidence that social interactions are also visually perceived. Along these lines, the presence of a social interaction has been found to influence the visual discrimination of agents [42] as well as causality perception [43].

Most of the aforementioned studies investigate actions with physically synchronous motion (e.g., dancing or fighting, Figure 2C,D). One may argue that such motion-driven effects are not truly social. However, using point light walkers, Manera and colleagues [44] found that the presence of communicative actions (e.g., gestured ‘sit down’ or ‘help yourself’ and subsequent responses, Figure 2E), which lack physically synchronous motion, also enhance the visual discrimination of agents. These results suggest that the visual effects described hold for more abstract depictions of communicative interactions.

Automatic extraction of higher-level meaning of interactions

Emerging evidence suggests that the visual system not only detects social interactions but is also sensitive to features of interactions. In very brief, masked presentations, the class of an interaction (e.g., pushing) is automatically extracted even when it is task irrelevant [45]. Other studies suggest, however, that these visual category distinctions may happen only for privileged socially relevant categories (e.g., ‘chasing’ vs. other kinds of ‘stalking’ [21] or ‘giving’ vs. ‘taking’ [46]). Beyond categorization, the social role of the agents in the interaction (i.e., agent vs. patient or pusher vs. person who was pushed) is also perceived in brief, masked presentations even when attention is diverted to an orthogonal task [45,47]. Thus, there is evidence that at least some social interaction categories and roles of the agents in the interactions are automatically processed.

There is also preliminary evidence that valence or moral judgments of interactions can be extracted visually. Infants [1,48] and non-human primates [3] perceive the valence of an interaction ('help' vs. 'hinder') and use this information to form preferences and guide actions (Box 1). The extent to which this is visual or the result of a moral core [11,14] is debated, although recent computational work suggests these distinctions may be driven by visual statistical associations (computational algorithms and Box 2). Using a causality paradigm, researchers found evidence that whether an agent is blame-worthy is also visually determined [49]. However, in this study the blame-worthy 'agent' was a car crashing into a person, and the extent to which this is perceived as a social interaction is an open question.

Computational algorithms for recognizing social interactions

Having outlined the rich visual representations supporting social interaction recognition, an open question is how this information is extracted by the visual system. Computational models can

Box 1. Phylogenetic origins of social interaction recognition

While this article focuses on evidence that human adults recognize social interactions in a visual manner, here we briefly review evidence that both infants and non-human primates perceive and make inferences based on the interactions of others and offer preliminary evidence that this ability may be visual. In a seminal study, preverbal infants were found to understand the social interactions of other agents [1,48] (Figure 1). Later research has shown that like adults [36–39], infants perform social chunking to increase working memory capacity [109]. Further, preverbal infants understand the rationality and goals of social actions [110,111] and form expectations about how a stranger will interact with them based on observed third-party interactions with their caregiver [2]. There is also emerging evidence that infants are sensitive to social primitives, including proximity [112], facingness [113], and motion congruency [112], and assignment to social groups may be based on these visual cues [112]. More research is needed to elucidate the extent to which social interactions are processed visually in infancy, but an interesting open question is how an early developing perceptual ability may be bootstrapped for later emerging cognitive abilities like theory of mind.

Social interaction recognition is also shared with monkeys [114,115] and apes [3,116–118]. Apes in particular have been found to base future actions on their perceptions of interactions [3,116–118]. There is little research investigating the visual mechanisms underlying this ability as this literature largely focuses on social interaction recognition as a precursor to theory of mind (see [86,119] for examples). However, one study in primates does suggest that the underlying mechanism may be visual in nature. Atsumi and colleagues [115] showed stimuli similar to Gao and Scholl [21] depicting animated shapes engaged in chasing interactions to Japanese macaques (*Macaca fuscata*). The macaques were not only sensitive to the presence of social interactions, they were also sensitive to the motion congruency cues of the chaser and agent being chased. Further, macaque responses matched human responses across the different motion congruency conditions. This provides exciting, preliminary evidence that the visual basis of core social interaction perception may be shared broadly among primates, although with some possible differences (see [120] for counter examples).

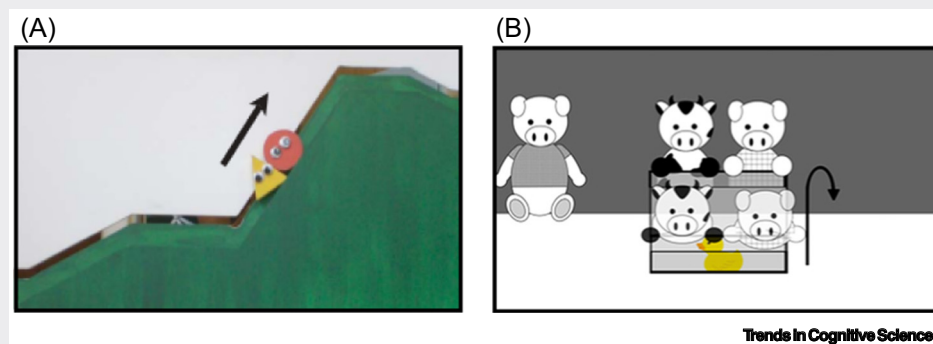


Figure 1. Social interaction stimuli used in studies of infants and apes. Infants [1,48] and bonobos (*Pan paniscus*) [3] can recognize social interactions and show subsequent preference for helping (infants) or hindering (bonobos) agent after viewing visual depictions of social interactions. These displays include (A) one agent helping (infants) or hindering (bonobos) agent up a hill (adapted from [1]), or (B) an agent unsuccessfully helping another open a box (adapted from [48]).

Box 2. Cue-based versus mentalistic accounts of social interaction recognition

In the past, perceptual or ‘cue-based’ accounts of how humans recognize social interactions (mostly focused on infants’ reaching preference for helping agents, see Box 1) have been largely dismissed based on two main arguments. First, infants can consistently recognize social relationships across a range of visual scenarios (see Figure I in Box 1). Second, when infants are presented with perceptually matched scenarios where one agent is replaced by a visually matched inanimate object (e.g., a red circle without eyes), helper preferences go away. Along similar lines, nearly identical visual events like giving and taking are differentially represented by infants based on their social nature (giving, in contrast to taking, necessitates a social interaction) [121]. This has led many to argue that instead infants rely on theory of mind (or an innate moral core [14]) to recognize social interactions. Such mental inference is often modeled with generative inverse planning models, which use explicit world knowledge of agents’ goals and the physical world to interpret social scenes [66] (Figure IA). In these models, judgements about a social interaction are made by inverting a generative model of agents’ interactions based on their goals and the physics of the world (i.e., comparing observed social scenes to different generated hypotheses from an internal world model). While there is now clear evidence that infants can use theory of mind to reason about social relationships when visual information is not diagnostic [14,122], we argue that prior cue-based accounts of interaction recognition have been overly simplified. Our visual system can extract information about agents, their intentions, and physical scenarios [7]. These higher-level visual features may serve as input to a system that then extracts social relationships in a bottom-up manner, even if the initial visual cues are not directly predictive of social interaction judgements (see Figure 3B in main text). Recent neural network models have been able to model many of these abilities using purely bottom-up computations without explicit simulation or representations of agents’ mental states and the physical world [67,70] (Figure IB, computational algorithms).

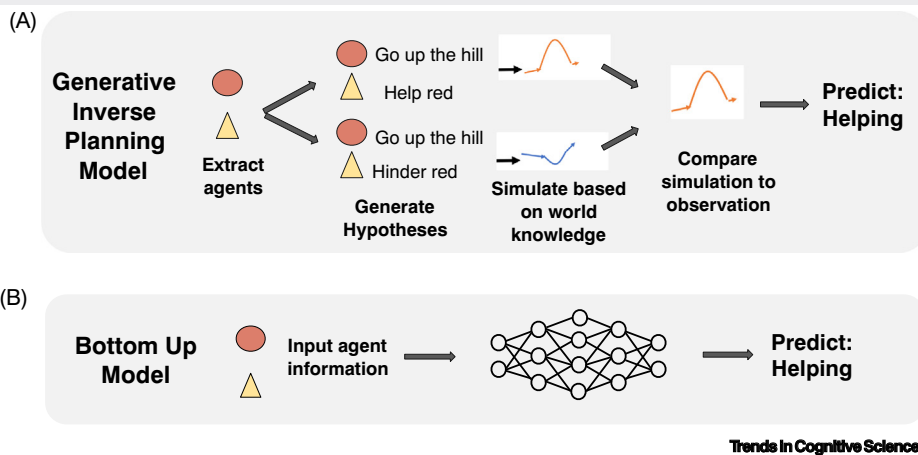


Figure I. Mentalistic versus cue-based accounts of social interaction recognition. (A) A common computational framework for modeling social interaction recognition with theory of mind, is generative inverse planning (figure adapted from [123]). Models implementing this framework aim to invert a generative model of agents’ goals and actions. They first extract visual information about agents (e.g., their position, size, velocity, animacy) and then generate hypotheses about their social relationship and goals using computational theory of mind [64–66]. They next perform simulations about how the social scene would play out under different hypotheses (requiring explicit world knowledge that is often implemented with a physics simulator) and make a prediction about the viewed scene based on matched to hypotheses. (B) In contrast, a purely bottom-up model would extract the same visual feature of agents and use that to directly predict the social relationship without any explicit hypothesis generation or knowledge of agents’ mental states. These are commonly implemented in neural network or connectionist models [67,71]. Prior cue-based accounts have ignored the possibility of additional visual processing and sought to predict interactions directly from visual information about the agent.

serve as an existence proof that a particular type of input information or algorithm is sufficient to solve a given task. In the following section, we review the types of algorithms that artificial intelligence (AI) and computer vision research have used to recognize social interactions. We present emerging evidence that: (i) visual social primitives are crucial precursors to social interaction recognition, and (ii) this task can be successfully modeled using visual algorithms without the need for higher-level cognitive models of mental state inference.

Bottom-up, discriminative models of high-level vision

To frame the following review and arguments, we first consider the question: what types of algorithms or computational models should be considered visual? Historically, visual processes are modeled with hierarchical, feedforward algorithms based on the architecture of visual cortex [50]. This is a known oversimplification [51] and more recent models of high-level vision include recurrence [52,53]. These algorithms, however, are still discriminative, meaning observations are directly mapped to a representation of a particular target in a bottom-up manner. In contrast, top-down, generative models aim to learn representations of the factors that generated the observed image. To recognize a particular target (like an object), generative models internally simulate possible observations (e.g., by rendering different 2D scenes from different possible 3D shapes and viewing conditions) and compare the true observation with their internal simulations to select the correct target. This process of ‘inverting’ a generative model is usually done using Bayesian inference [54,55], and these models of vision are often referred to as ‘inverse graphics’ or ‘analysis by synthesis’ approaches [56,57]. The extent to which such generative processes play a role in different aspects of high-level vision is an interesting open question [58]. In contrast to this approach, we argue that social interaction recognition can be solved specifically via discriminative or bottom-up (though not necessarily strictly feedforward) visual algorithms without the need for such generative processes.

A model does not have to be fully unstructured to be considered visual. While end-to-end learning systems with little structure (like deep neural networks) can solve most vision problems with enough training data, we know that humans use **inductive biases** to simplify learning and improve generalization [59,60] and that some inductive biases are present in the visual system. We consider models with inductive biases that can be extracted by the visual system (e.g., social primitives [33] and relations [24], both described later) to be visual. Models with added information extracted outside of the visual system (e.g., mental states) are considered extravisual.

Visual social primitives help computer vision systems recognize social interactions

Unlike other aspects of visual recognition, social interactions (and social scene understanding more generally) have historically received relatively little attention from the AI community. Recently, however, social interactions have been identified as a critical area for future AI research [61] and AI and computer vision researchers have increasingly studied social interaction recognition. This work has revealed that several of the visual social primitive features identified by cognitive psychologists improve computer vision systems’ ability to recognize interactions. For example, adding gaze direction into motion-based computer vision models improves detection of social groups [62]. In addition, a model based on posture, inter-agent distance, and facingness can accurately predict human judgements of social groupings in static and dynamic scenes [63].

This work suggests that systems that explicitly represent social primitives in their input do better than unstructured end-to-end learning systems at detecting interactions, adding to the mounting behavioral evidence that social primitives are critical precursors to detecting interactions. This precursor strategy may be particularly important since the computer vision field does not currently have the type of large-scale datasets of social actions used in learning other visual tasks. While it is possible that with enough data end-to-end learning systems will achieve human-level performance without explicit representations of social primitives, the aforementioned studies point to the computational efficiency of using social primitives as precursors to interaction recognition.

Visual versus inverse planning models of social interaction recognition

Most successful computational models of human social interaction recognition have sought to model higher-level cognitive processes that go beyond vision. These **generative inverse**

planning models, while related to the generative models of vision described earlier, are often considered to be implementing a computational theory of mind [64] and thus extravisual. These models invert a generative model of social interactions based on explicit representations of agents' mental states and the physical world [65,66] (see Figure 1A in Box 2). Their success over visual, cue-based models (Box 2) in matching human behavior has been provided as evidence that humans use similar processes of mental inference to recognize social interactions.

More recent work, however, has shown that with proper inductive biases, bottom-up models can also explain human social interaction recognition. One study developed a neural network model to link input scenes of social interactions to output representations of agents using associative learning mechanisms [67]. Using minimal assumptions and supervision based on everyday interactions an infant is likely to encounter (e.g., agents acting in a 'concordant' or physically contingent manner are more likely to engage in subsequent interactions), this model can reproduce infant preferences for helping agents across a range of interaction scenarios. One outstanding question, however, is to what extent the agent and scene information input to this model could be extracted visually.

Indeed, standard visual deep neural network models that match many other aspects of human visual behavior do a poor job of recognizing social interactions in images or videos [68]. However, recent work incorporating relational inductive biases [69] into bottom-up neural network models found that these models can match human judgements of social interactions with only visual input and relatively little training data but require relational information to do so [70,71].

Developing image-computable models of human social interaction judgements hold great promise for more thoroughly characterizing the cognitive and neural computations underlying these abilities. In addition, they present new ways to test hypotheses about how social perceptual systems interact with cognitive systems for social scene understanding. For example, a bottom-up neural network could serve as input to a generative inverse planning system to model both interaction perception and theory of mind judgements in social scenes [59].

Neural basis of social interaction recognition

How might the aforementioned cognitive computations for recognizing social interactions be implemented in the brain? While the detailed neural implementation of these processes is still largely unknown, recent work has sought to understand the cortical organization of this information, which can provide evidence as to what functions are processed by separate versus overlapping neural systems [72]. In this section, we review evidence for the representation of social interactions in visual cortex, particularly the extrastriate body area (EBA), and in multimodal posterior superior temporal sulcus (pSTS). We also review evidence and discuss the extent to which these regions are distinct from those engaged in theory of mind processing (Figure 3A).

Social primitive representations in the brain

In addition to the behavioral work described earlier, growing neuroimaging research suggests that facing dyads are represented in visual cortex. In particular, EBA, a region in lateral occipitotemporal cortex (LOTc) that shows selective responses to bodies versus other categories of visual stimuli (including faces, objects, and scenes), responds more to dyads than individual bodies and more to facing than non-facing dyads [73]. In dynamic displays of point light figures, this preference can be seen in EBA and the pSTS [74]. EBA also shows other hallmarks of configural processing of dyads, including susceptibility to stimulus inversion and activity that is modulated based on behavioral measures of configural processing [75].

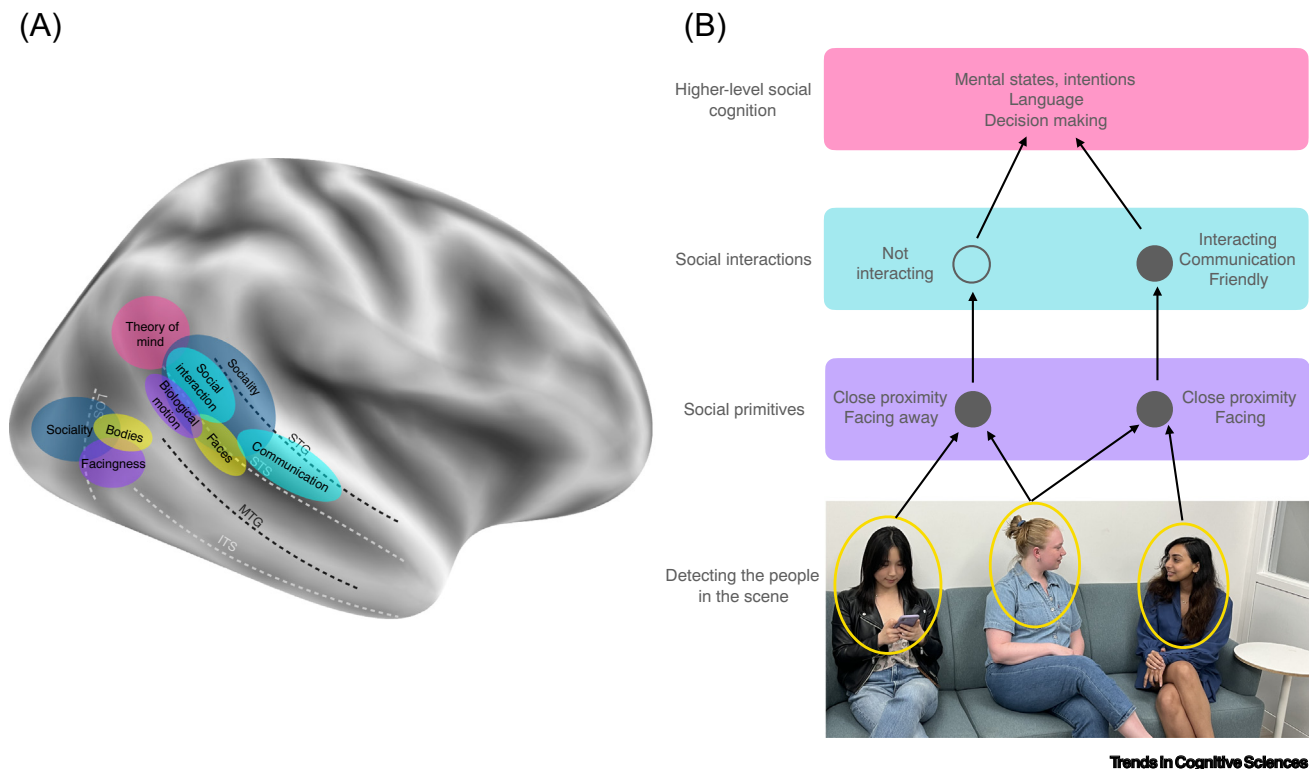


Figure 3. Neural computations of dyadic social interactions. (A) Approximate locations of regions in the brain representing different features of social interactions: configuration between bodies in lateral occipitotemporal cortex (LOTC) [73–75], sociality in LOTC [76,78], and posterior superior temporal sulcus (pSTS) [78], social interactions in pSTS [93,94], biological motion in pSTS [135], faces in pSTS [98], communication along the STS [105,133,134], and theory of mind in the temporoparietal junction (TPJ) [136]. Select anatomical landmarks (sulci: light gray, gyri: black) are annotated: lateral occipital sulcus (LOS), inferior temporal sulcus (ITS), middle temporal gyrus (MTG), superior temporal sulcus (STS), and superior temporal gyrus (STG). (B) A proposed computational framework for core social interaction recognition in the human brain. To recognize social interactions, people are first detected in body and face-selective areas [some of which are highlighted in yellow in (A)]. Next, social primitives (purple) are extracted based on the relative configuration of people in the scene. These computations occur in regions of LOTC and pSTS (A). While detecting at least one social primitive is necessary to recognize an interaction, this information is not sufficient (e.g., the nearby agents on the left are not interacting). This information is then transformed to recognize social interactions (teal) along the STS. We propose that core social interaction recognition is computed primarily based on bottom-up computations (i.e., information from one stage is directly fed to the next). This does not preclude recurrent computations that likely occur at each processing stage, particularly for recognizing interactions across time. The core components of an interaction are then fed to higher-level cognitive regions, including areas processing theory of mind and communicative language, at which top-down cognitive processes are engaged.

Sociality representations in the brain

A large body of fMRI work has investigated how actions generally are represented in the human brain. Several studies have now shown that **sociality** is an important organizing dimension of action representations behaviorally and in the brain, particularly in lateral occipital visual regions. These effects have been observed across a range of controlled fMRI studies (e.g., [76], for a recent review, see [77]). Sociality was also identified as a key organizing dimension in data-driven investigations of fMRI responses to naturalistic videos [78]. Sociality also explains significant variance in unguided, intuitive behavior similarity judgments of natural videos [79]. However, the role of sociality in explaining neural responses in other naturalistic studies has been somewhat mixed since it is difficult to dissociate sociality from strongly covarying signals such as the number of agents in a video [79,80]. While prior studies controlled for the number of agents [76], it remains an open question to what extent sociality, like facingness, is a representation of social interaction or closely correlated visual cues.

Social interactions in the STS

Extensive neuroscience research has now revealed brain regions that respond more to people acting together versus alone across a range of stimuli, including images [81], video clips [82], and point light figures [83–85]. Responses are found most consistently in the pSTS. Some studies also identified social interaction responses in the theory of mind network, including medial prefrontal cortex (mPFC) and precuneus [13,82–84]. Similar brain networks have also been found in non-human primates [86]. In general, interacting and non-interacting stimuli in these studies both include two people and interactivity is conveyed through movements and configurations beyond simple visual cues such as facingness (see Box 3 for discussion of participant-directed social interactions depicted with single actors).

Several prior studies have also used Heider and Simmel [5] style animated shaped videos to contrast social motion with physical interactions or random motion. These studies often find both the STS and the theory of mind network, including mPFC, precuneus, and the temporoparietal junction (TPJ), are activated when viewing social interactions [87–92]. However, these results are often interpreted in terms of sensitivity to animacy or goal-directed behavior and, indeed, the contrasted shape conditions vary across these factors as well as the presence of social interactions.

More recent studies have identified a region in the pSTS that responds significantly more to interacting point light dyads than two agents acting independently but does not show preferential responses to other social stimuli (like faces or false belief tasks). Similarly, nearby regions representing dynamic faces or theory of mind do not show a preference for interacting versus non-interacting dyads [93–95]. This work suggests that when controlling for social content and task demands, there is a clear dissociation between recognizing social interactions and theory of mind. It is important to note that theory of mind regions in these experiments were identified using a false belief task and that broader definitions of theory of mind include not only information about others' beliefs but also their goals and emotions [96,97]. While representations of goals and emotions have also been found in the STS [98], this information appears to be in distinct regions from those showing social interaction selectivity [93]. The functional separation between social interactions and theory of mind can even be seen during natural movie viewing when the two features are strongly correlated [99].

Box 3. Second-person social interaction perception

This review focuses on the perception of third-person social interactions, interactions between others that are only observed by the participant. However, a separate body of work has investigated perception of **second-person social interactions**, real or simulated interactions between another person and the participant [124]. There is evidence that understanding the actions of a partner in an interaction is also based on visual cues. People rely on body motion to predict the target of an interactive opponent's reach [125,126]. Further, in an interactive context, the facing direction of other agents towards a participant affects performance even for agents the participant is not directly interacting with and who are therefore task irrelevant [22]. Finally, gaze direction, which has been a large topic of research for decades, is a social visual cue. While infants are sensitive to direct gaze at birth [127], gaze following requires visual experience [128] and is diminished when gaze is not a developmentally informative social cue such as infants born to blind parents [129]. In adults, gaze cueing is also diminished when it can be 'explained away' by other social factors like gaze deflection [130], suggesting the visual system is sensitive to intentions behind gaze changes.

Whether or not the extent to which perception of second- and third-person interactions have the same underlying mechanism is largely an open question. Some of the same brain regions representing third-person interactions have also been implicated in representing second-person social interaction. For example, third-party social interactions [93,94,99], mutual gaze [131,132], and communicativeness of a gesture [133] or face [134] activate regions of the STS (see Figure 3A in main text) though these have never been compared in individual subjects. Future research should investigate whether the STS has shared representations for second- and third-person interactions.

While the STS, particularly posterior portions, is a key region involved in dynamic visual processing [100], it also process multimodal information and is, thus, not purely visual. For example, some preliminary evidence suggests that the pSTS may also respond to auditory stimuli depicting social interactions [101,102] (see [Outstanding questions](#)). On the other hand, some work has suggested that even lower-level visual regions, particularly EBA, may also be involved in processing social interactions in point light dyads [95,103] and animated shape videos [91,104]. The latter study even found that subjects were more likely to describe ambiguous mechanical stimuli as social than non-social and that responses in visual cortex predicted these behavioral judgements in individual subjects. Together, this evidence points to a clear involvement of visual cortex in detecting others' social interactions and recent proposals have suggested that these posterior lateral visual regions serve as input to the STS in a 'third' visual pathway (in addition to the classic ventral and dorsal pathways) dedicated to social perception [77,100,105].

Higher-level feature representations in visual brain regions

A few studies suggest that EBA and pSTS represent not only the presence of a social interaction but also higher-level judgements of that interaction. For example, neural patterns in the EBA and pSTS can distinguish between different types of dyadic interactions [95]. In addition, activity in pSTS can predict goal compatibility of an interaction (e.g., helping vs. hindering or cooperation vs. competition) [93,94]. It is important to note though that this information about goal compatibility is also decodable from TPJ in the theory of mind network. Given the low temporal resolution of fMRI, it is difficult to tell from these studies alone if these help versus hinder representations are visually driven or based on top-down signals from the theory of mind network.

Social interaction representations in the theory of mind network

An alternative view argues that social interaction representations in the brain can be attributed to mentalization and activity in the theory of mind network. Computationally, this is supported by the generative inverse planning models described earlier. Neurally, one recent study made this argument based on meta-analysis of several studies investigating social interaction, action recognition, and theory of mind tasks [17]. Critically, while the pSTS and TPJ are separable in individual subjects [93], they are extremely close spatially and thus likely to be blurred together in a group or meta-analysis. In addition, several tasks or stimuli can engage both perceptual and mentalization systems (e.g., judging helping vs. hindering scenarios [93,94] as described earlier or watching a natural movie [13,99]), so it is important to separate them using appropriate tasks or analyses. Finally, task demands that correlate with stimulus properties in an fMRI experiment (e.g., pressing a button in response to interacting stimuli [84,85]) may lead to prefrontal brain activity due to attention or preparatory motor responses. When these considerations are taken into account (using a within-subjects design, without task confounds, and stimuli or analysis methods that separate social interaction recognition from theory of mind), social interactions and mentalization appear to be processed in distinct neural networks [93,99].

Some of our prior work has suggested that social interactions are represented on a relatively slow timescale by the brain [68,79], a potential argument against visual processing. For example, one study used tightly controlled pairs of natural images either with versus without a social interaction and found that the presence of a social interaction could not be read out from magnetoencephalography signals until at least 300 ms after image onset, a time period generally considered to be outside the range of feedforward visual processing [68]. It is possible that the challenging nature of these matched natural image pairs required recurrent visual processing, similar to those engaged in challenging object recognition [53,106,107], or that the distinction of subtle gaze shifts fall outside of core interaction recognition. Importantly, in these studies, social interaction

information was extracted spontaneously by the brain, even in the absence of an explicit task, suggesting some degree of automatic visual processing.

Concluding remarks

Looking around, we cannot help but make complex inferences about the structure of the social world. Here we present converging evidence from behavioral, computational, and neural studies that rich, abstract information about the social interactions of others is computed by the visual system. Synthesizing across different levels of analysis, we put forth a proposal that social interactions are computed in a bottom-up manner across different regions in the lateral visual pathway [100] (Figure 3B). Many open questions and challenges remain (see Outstanding questions). Critically, although we liken core social interaction recognition to core object recognition, a fundamental challenge remains in characterizing the computational goal of social perception. Indeed, prior work suggests that social interactions cannot be broken down into the same taxonomy of categories as objects [108], suggesting social interactions require more flexible representations that might be better understood in terms of more abstract concepts. Further, additional work is needed linking computational models that recapitulate human behavior with visuo-social brain responses and improving our mechanistic understanding of these neural computations. Finally, we must understand how these visual social representations are integrated with information from other perceptual modalities and higher-level cognitive systems. Interdisciplinary studies of social interaction recognition can help bridge the gap between high-level vision and social cognition and will be critical to understanding humans' rich social visual abilities.

Acknowledgments

This work was funded in part by NSF GRFP DGE-1746891 awarded to E.M. and NIMH R01MH132826 awarded to L.I. We would like to thank the following people for useful discussion and feedback on earlier drafts of this paper: Barbara Landau, Chaz Firestone, Ian Phillips, Mick Bonner, Chris Krupenye, Lindsey Powell, Manasi Malik, Sholei Croom, Zekun Sun, Rui Zhe Goh, Lana Milman, and Luz Carvajal Villalobos.

Declaration of interests

No interests are declared.

References

1. Hamlin, J.K. *et al.* (2007) Social evaluation by preverbal infants. *Nature* 450, 557–559
2. Thomas, A.J. *et al.* (2022) Infants infer potential social partners by observing the interactions of their parent with unknown others. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2121390119
3. Krupenye, C. and Hare, B. (2018) Bonobos prefer individuals that hinder others over those that help. *Curr. Biol.* 28, 280–286
4. Bergman, T.J. *et al.* (2003) Hierarchical classification by rank and kinship in baboons. *Science* 302, 1234–1236
5. Heider, F. and Simmel, M. (1944) An experimental study of apparent behavior. *Am. J. Psychol.* 57, 243–259
6. Scholl, B.J. and Gao, T. (2013) Perceiving animacy and intentionality: visual processing or higher-level judgment? In *Social Perception: Detection and Interpretation of Animacy, Agency, and Intention* (Rutherford, M.D. and Kuhlmeier, V.A., eds), The MIT Press
7. Scholl, B.J. and Tremoulet, P.D. (2000) Perceptual causality and animacy. *Trends Cogn. Sci.* 4, 299–309
8. Castelli, F. *et al.* (2002) Autism, Asperger syndrome and brain mechanisms for the attribution of mental states to animated shapes. *Brain* 125, 1839–1849
9. Csibra, G. *et al.* (1999) Goal attribution without agency cues: the perception of 'pure reason' in infancy. *Cognition* 72, 237–267
10. Gergely, G. *et al.* (1995) Taking the intentional stance at 12 months of age. *Cognition* 56, 165–193
11. Hamlin, J.K. (2013) Moral judgment and action in preverbal infants and toddlers: evidence for an innate moral core. *Curr. Dir. Psychol. Sci.* 22, 186–193
12. Ullman, S. *et al.* (2012) From simple innate biases to complex visual concepts. *Proc. Natl. Acad. Sci. U. S. A.* 109, 18215–18220
13. Wagner, D.D. *et al.* (2016) The dorsal medial prefrontal cortex responds preferentially to social interactions during natural viewing. *J. Neurosci.* 36, 6917–6925
14. Woo, B.M. *et al.* (2022) Human morality is based on an early-emerging moral core. *Annu. Rev. Dev. Psychol.* 4, 41–61
15. DiCarlo, J.J. *et al.* (2012) How does the brain solve visual object recognition? *Neuron* 73, 415–434
16. Greene, M.R. and Oliva, A. (2009) The briefest of glances: the time course of natural scene understanding. *Psychol. Sci.* 20, 464–472
17. Arioli, M. and Canessa, N. (2019) Neural processing of social interaction: coordinate-based meta-analytic evidence from human neuroimaging studies. *Hum. Brain Mapp.* 40, 3712–3737
18. Baker, B. *et al.* (2022) Three aspects of representation in neuroscience. *Trends Cogn. Sci.* 26, 942–958
19. Marr, D. (1982) *Vision*, W.H. Freeman
20. Firestone, C. and Scholl, B.J. (2016) Cognition does not affect perception: evaluating the evidence for "top-down" effects. *Behav. Brain Sci.* e229, 1–77
21. Gao, T. *et al.* (2009) The psychophysics of chasing: a case study in the perception of animacy. *Cognit. Psychol.* 59, 154–179
22. Gao, T. *et al.* (2010) The wolfpack effect: perception of animacy irresistibly influences interactive behavior. *Psychol. Sci.* 21, 1845–1853
23. Rolfs, M. *et al.* (2013) Visual adaptation of the perception of causality. *Curr. Biol.* 23, 250–254

Outstanding questions

Beyond detection and goal compatibility, what other information about social interactions is extracted by the visual system? The answer to this question will help characterize the computational-level goals of social perception. One promising direction may be identifying divisions within the representational space of social interactions in the brain and behavior.

How do the visual system and higher-level social cognitive systems, including the theory of mind network, interact to form more complex social evaluations? Comparing computational models of these systems with behavioral and neural responses can help test theories of how these systems interact.

How are social interactions represented in other perceptual modalities? Are neural representations for social interactions multimodal? Preliminary research has shown that auditory social interactions activate similar regions as visual presentations of social interactions. However, more work is needed to understand the extent to which these are cross-modal representations.

Are the same neural systems and computations used to process social interactions depicted in simplified stimuli, such as static images or animated geometric shapes, the same as those used for processing naturalistic social interactions?

How do we detect social interactions among crowds of people? Are there differences in social interaction perception for dyads compared with larger interacting groups?

How are social primitives and social interactions represented at the level of neural circuits? Collecting spatiotemporally-resolved neural data, such as electrophysiology in humans and non-human primates, will be an important next step in understanding the neural basis of social interaction perception.

24. Hafri, A. and Firestone, C. (2021) The perception of relations. *Trends Cogn. Sci.* 25, 475–492
25. van Buren, B. *et al.* (2017) What are the underlying units of perceived animacy? Chasing detection is intrinsically object-based. *Psychon. Bull. Rev.* 24, 1604–1610
26. Papeo, L. *et al.* (2017) The two-body inversion effect. *Psychol. Sci.* 28, 369–379
27. Papeo, L. *et al.* (2019) Visual search for people among people. *Psychol. Sci.* 30, 1483–1496
28. Hochmann, J.-R. and Papeo, L. (2021) How can it be both abstract and perceptual? Comment on Hafri, A., & Firestone, C. (2021), The perception of relations, Trends in Cognitive Sciences. *PsyArXiv* Published online July 13, 2021. <https://doi.org/10.31234/osf.io/hm49p>
29. Thurman, S.M. and Lu, H. (2014) Perception of social interactions for spatially scrambled biological motion. *PLoS One* 9, e112539
30. Vestner, T. *et al.* (2019) Bound together: social binding leads to faster processing, spatial distortion, and enhanced memory of interacting partners. *J. Exp. Psychol. Gen.* 148, 1251–1268
31. Vestner, T. *et al.* (2021) Visual search for facing and non-facing people: the effect of actor inversion. *Cognition* 208, 104550
32. Vestner, T. *et al.* (2022) Remembered together: social interaction facilitates retrieval while reducing individuation of features within bound representations. *Q. J. Exp. Psychol.* 75, 1593–1602
33. Papeo, L. (2020) Twos in human visual perception. *Cortex* 132, 473–478
34. Su, J. *et al.* (2016) Social interactions receive priority to conscious perception. *PLoS One* 11, e0160468
35. Skripkauskaitė, S. *et al.* (2022) Attentional bias towards social interactions during viewing of naturalistic scenes. *Q. J. Exp. Psychol. (Hove)* 76, 2197–2430
36. Ji, H. *et al.* (2020) Selective attention operates on the group level for interactive biological motion. *J. Exp. Psychol. Hum. Percept. Perform.* 46, 1434–1442
37. Ding, X. *et al.* (2017) Two equals one: two human actions during social interaction are grouped as one unit in working memory. *Psychol. Sci.* 28, 1311–1320
38. Paparella, I. and Papeo, L. (2022) Chunking by social relationship in working memory. *Vis. Cogn.* 30, 354–370
39. Lu, X. *et al.* (2022) Is the social chunking of agent actions in working memory resource-demanding? *Cognition* 229, 105249
40. Fedorov, L.A. *et al.* (2018) Adaptation aftereffects reveal representations for encoding of contingent social actions. *Proc. Natl. Acad. Sci. U. S. A.* 115, 7515–7520
41. Fodor, J.A. (1983) *The Modularity of Mind: An Essay on Faculty Psychology*, MIT Press
42. Neri, P. *et al.* (2006) Meaningful interactions can enhance visual discrimination of human agents. *Nat. Neurosci.* 9, 1186–1192
43. Zhou, J. *et al.* (2012) Perceived causalities of physical events are influenced by social cues. *J. Exp. Psychol. Hum. Percept. Perform.* 38, 1465–1475
44. Manera, V. *et al.* (2011) The second-agent effect: communicative gestures increase the likelihood of perceiving a second agent. *PLoS One* 6, 1–7
45. Hafri, A. *et al.* (2013) Getting the gist of events: recognition of two-participant actions from brief displays. *J. Exp. Psychol. Gen.* 142, 880–905
46. Yin, J. *et al.* (2022) Structural asymmetries in the representation of giving and taking events. *Cognition* 229, 105248
47. Hafri, A. *et al.* (2018) Encoding of event roles from visual scenes is rapid, spontaneous, and interacts with higher-level visual processing. *Cognition* 175, 36–52
48. Hamlin, J.K. (2013) Failed attempts to help and harm: intention versus outcome in preverbal infants' social evaluations. *Cognition* 128, 451–474
49. De Freitas, J. and Alvarez, G.A. (2018) Your visual system provides all the information you need to make moral judgments about generic visual events. *Cognition* 178, 133–146
50. Riesenhuber, M. and Poggio, T. (1999) Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025
51. Felleman, D.J. and Van Essen, D.C. (1991) Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1, 1–47
52. Spoerer, C.J. *et al.* (2020) Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS Comput. Biol.* 16, e1008215
53. Kar, K. *et al.* (2019) Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* 22, 974–983
54. Lake, B.M. *et al.* (2015) Human-level concept learning through probabilistic program induction. *Science* 350, 1332–1338
55. Feldman, J. and Singh, M. (2006) Bayesian estimation of the shape skeleton. *Proc. Natl. Acad. Sci. U. S. A.* 103, 18014–18019
56. Yildirim, I. *et al.* (2020) Efficient inverse graphics in biological face processing. *Sci. Adv.* 6, eaax5979
57. Yuille, A. and Kersten, D. (2006) Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* 10, 301–308
58. DiCarlo, J.J. *et al.* (2021) How does the brain combine generative models and direct discriminative computations in high-level vision? In *Cognitive Computational Neuroscience Generative Adversarial Collaborations 2021*, CCN
59. Lake, B.M. *et al.* (2017) Building machines that learn and think like people. *Behav. Brain Sci.* 40, E253
60. Spelke, E.S. and Kinzler, K.D. (2007) Core knowledge. *Dev. Sci.* 10, 89–96
61. Bolotta, S. and Dumas, G. (2022) Social neuro AI: social interaction as the "dark matter" of AI. *Front. Comput. Sci.* 4, 846440
62. Leach, M. *et al.* (2014) Detecting social groups in crowded surveillance videos using visual attention. In *2014 IEEE Conference: CVPR Workshop on Computational Models for Social Interactions and Behavior*, pp. 467–473, IEEE
63. Zhou, C. *et al.* (2019) A social interaction field model accurately identifies static and dynamic social groupings. *Nat. Hum. Behav.* 3, 847–855
64. Baker, C.L. *et al.* (2017) Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nat. Hum. Behav.* 1, 1–10
65. Ullman, T. *et al.* (2009) Help or hinder: Bayesian models of social goal inference. *Adv. Neural Inf. Process. Syst.* 22, 1874–1882
66. Netanyahu, A. *et al.* (2021) PHASE: PHysically-grounded Abstract Social Events for machine social perception. *arXiv* Published online March 19, 2021. <https://doi.org/10.48550/arXiv.2103.01933>
67. Benton, D.T. and Lapan, C. (2022) Moral masters or moral apprentices? A connectionist account of sociomoral evaluation in preverbal infants. *Cogn. Dev.* 62, 101164
68. Isik, L. *et al.* (2020) The speed of human social interaction perception. *NeuroImage* 215, 116844
69. Battaglia, P.W. *et al.* (2018) Relational inductive biases, deep learning, and graph networks. *arXiv* Published online October 17, 2018. <https://doi.org/10.48550/arxiv.1806.01261>
70. Malik, M. and Isik, L. Social inference from relational visual information: an investigation with graph neural network models. In *2022 Conference on Cognitive Computational Neuroscience*, CCN
71. Malik, M. and Isik, L. (2022) Relational visual information explains human social inference: a graph neural network model for social interaction recognition. *PsyArXiv* Published online November 3, 2022. <https://doi.org/10.31234/osf.io/5cuyr>
72. Kanwisher, N. (2010) Functional specificity in the human brain: a window into the functional architecture of the mind. *Proc. Natl. Acad. Sci. U. S. A.* 107, 11163–11170
73. Abassi, E. and Papeo, L. (2020) The representation of two-body shapes in the human visual cortex. *J. Neurosci.* 40, 852–863
74. Bellot, E. *et al.* (2021) Moving toward versus away from another: how body motion direction changes the representation of bodies and actions in the visual cortex. *Cereb. Cortex* 31, 2670–2685
75. Abassi, E. and Papeo, L. (2022) Behavioral and neural markers of visual configural processing in social scene perception. *NeuroImage* 260, 119506
76. Wurm, M.F. *et al.* (2017) Action categories in lateral occipitotemporal cortex are organized along sociality and transitivity. *J. Neurosci.* 37, 562–575
77. Wurm, M.F. and Caramazza, A. (2022) Two 'what' pathways for action and object recognition. *Trends Cogn. Sci.* 26, 103–116
78. Tarhan, L. and Konkle, T. (2020) Sociality and interaction envelope organize visual action representations. *Nat. Commun.* 11, 3002

79. Dima, D.C. *et al.* (2022) Social-affective features drive human representations of observed actions. *eLife* 11, e75027
80. Tucciarelli, R. *et al.* (2019) The representational space of observed actions. *eLife* 8, e47686
81. Quadflieg, S. *et al.* (2015) The neural basis of perceiving person interactions. *Cortex* 70, 5–20
82. Iacoboni, M. *et al.* (2004) Watching social interactions produces dorsomedial prefrontal and medial parietal BOLD fMRI signal increases compared to a resting baseline. *NeuroImage* 21, 1167–1173
83. Petrini, K. *et al.* (2014) Look at those two! The precuneus role in unattended third-person perspective of social interactions. *Hum. Brain Mapp.* 35, 5190–5203
84. Centelles, L. *et al.* (2011) Recruitment of both the mirror and the mentalizing networks when observing social interactions depicted by point-lights: a neuroimaging study. *PLoS One* 6, e15749
85. Sapey-Triomphe, L.-A. *et al.* (2016) Deciphering human motion to discriminate social interactions: a developmental neuroimaging study. *Soc. Cogn. Affect. Neurosci.* 30, nsw117
86. Sliwa, J. and Freiwald, W.A. (2017) A dedicated network for social interaction processing in the primate brain. *Science* 356, 745–749
87. Castelli, F. *et al.* (2000) Movement and mind: a functional imaging study of perception and interpretation of complex intentional movement patterns. *NeuroImage* 12, 314–325
88. Schultz, R.T. *et al.* (2003) The role of the fusiform face area in social cognition: implications for the pathobiology of autism. *Philos. Trans. R. Soc. B Biol. Sci.* 358, 415–427
89. Gobbin, M.I. *et al.* (2007) Two takes on the social brain: a comparison of theory of mind tasks. *J. Cogn. Neurosci.* 19, 1803–1814
90. Santos, N.S. *et al.* (2010) Animated brain: a functional neuroimaging study on animacy experience. *NeuroImage* 53, 291–302
91. Gao, T. *et al.* (2012) Dissociating the detection of intentionality from animacy in the right posterior superior temporal sulcus. *J. Neurosci.* 32, 14276–14280
92. Lee, S.M. *et al.* (2014) Attributing intentions to random motion engages the posterior superior temporal sulcus. *Soc. Cogn. Affect. Neurosci.* 9, 81–87
93. Isik, L. *et al.* (2017) Perceiving social interactions in the posterior superior temporal sulcus. *Proc. Natl. Acad. Sci. U. S. A.* 114, E9145–E9152
94. Walbrin, J. *et al.* (2018) Neural responses to visually observed social interactions. *Neuropsychologia* 112, 31–39
95. Walbrin, J. and Koldewyn, K. (2019) Dyadic interaction processing in the posterior temporal cortex. *NeuroImage* 198, 296–302
96. Bloom, P. and German, T.P. (2000) Two reasons to abandon the false belief task as a test of theory of mind. *Cognition* 77, B25–B31
97. Tomasello, M. (2018) How children come to understand false beliefs: a shared intentionality account. *Proc. Natl. Acad. Sci. U. S. A.* 115, 8491–8498
98. Allison, T. *et al.* (2000) Social perception from visual cues: role of the STS region. *Trends Cogn. Sci.* 4, 267–278
99. Masson, H.L. and Isik, L. (2021) Functional selectivity for social interaction perception in the human superior temporal sulcus during natural viewing. *NeuroImage* 245, 118741
100. Pitcher, D. and Ungerleider, L.G. (2021) Evidence for a third visual pathway specialized for social perception. *Trends Cogn. Sci.* 25, 100–110
101. Landsiedel, J. and Koldewyn, K. (2023) Auditory dyadic interactions through the ‘eye’ of the social brain: how visual is the posterior STS interaction region? *Imaging Neurosci.* 1, 1–20
102. Olson, H. *et al.* (2023) Left-hemisphere cortical language regions respond equally to dialogue and monologue. *bioRxiv* Published online August 16, 2023. <https://doi.org/10.1101/2023.01.30.526344>
103. Landsiedel, J. *et al.* (2022) The role of motion in the neural representation of social interactions in the posterior temporal cortex. *NeuroImage* 262, 119533
104. Varrier, R.S. and Finn, E.S. (2022) Seeing social: a neural signature for conscious perception of social interactions. *J. Neurosci.* 42, 9211–9226
105. McMahon, E. *et al.* (2023) Hierarchical organization of social action features along the lateral visual pathway. *PsyArXiv* Published online March 23, 2023. <https://doi.org/10.31234/osf.io/x3avb>
106. Rajaei, K. *et al.* (2019) Beyond core object recognition: recurrent processes account for object recognition under occlusion. *PLoS Comput. Biol.* 15, e1007001
107. Tang, H. *et al.* (2014) Spatiotemporal dynamics underlying object completion in human ventral visual cortex. *Neuron* 83, 736–748
108. de la Rosa, S. *et al.* (2014) Visual categorization of social interactions. *Vis. Cogn.* 22, 1233–1271
109. Stahl, A.E. and Feigenson, L. (2014) Social knowledge facilitates chunking in infancy. *Child Dev.* 85, 1477–1490
110. Gredebäck, G. and Melinder, A. (2010) Infants’ understanding of everyday social interactions: a dual process account. *Cognition* 114, 197–206
111. Fawcett, C. and Gredebäck, G. (2013) Infants use social context to bind actions into a collaborative sequence. *Dev. Sci.* 16, 841–849
112. Powell, L.J. and Spelke, E.S. (2013) Preverbal infants expect members of social groups to act alike. *Proc. Natl. Acad. Sci. U. S. A.* 110, E3965–E3972
113. Goupil, N. *et al.* (2022) Visual perception grounding of social cognition in preverbal infants. *Infancy* 27, 210–231
114. Atsumi, T. and Nagasaka, Y. (2015) Perception of chasing in squirrel monkeys (*Saimiri sciureus*). *Anim. Cogn.* 18, 1243–1253
115. Atsumi, T. *et al.* (2017) Goal attribution to inanimate moving objects by Japanese macaques (*Macaca fuscata*). *Sci. Rep.* 7, 40033
116. Herrmann, E. *et al.* (2013) Direct and indirect reputation formation in nonhuman great apes (*Pan paniscus*, *Pan troglodytes*, *Gorilla gorilla*, *Pongo pygmaeus*) and human children (*Homo sapiens*). *J. Comp. Psychol.* 127, 63–75
117. Russell, Y.I. *et al.* (2008) Image scoring in great apes. *Behav. Process.* 78, 108–111
118. Subiaul, F. *et al.* (2008) Do chimpanzees learn reputation by observation? Evidence from direct and indirect experience with generous and selfish strangers. *Anim. Cogn.* 11, 611–623
119. Lewis, L.S. and Krupenye, C. (2022) Theory of mind in nonhuman primates. In *Primate Cognitive Studies* (Beran, M.J. and Schwartz, B.L., eds), pp. 439–482, Cambridge University Press
120. Schafroth, J.L. *et al.* (2021) No evidence that monkeys attribute mental states to animated shapes in the Heider–Simmel videos. *Sci. Rep.* 11, 3050
121. Tatone, D. *et al.* (2015) Giving and taking: representational building blocks of active resource-transfer events in human infants. *Cognition* 137, 47–62
122. Hamlin, J.K. *et al.* (2013) The mentalistic basis of core social cognition: experiments in preverbal infants and a computational model. *Dev. Sci.* 16, 209–226
123. Shu, T. *et al.* (2021) AGENT: a benchmark for core psychological reasoning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 9614–9625, PMLR
124. Redcay, E. and Schilbach, L. (2019) Using second-person neuroscience to elucidate the mechanisms of social interaction. *Nat. Rev. Neurosci.* 20, 495–505
125. Vaziri-Pashkam, M. *et al.* (2017) Predicting actions from subtle preparatory movements. *Cognition* 168, 65–75
126. McMahon, E. *et al.* (2019) Subtle predictive movements reveal actions regardless of social context. *J. Vis.* 19, 16
127. Farroni, T. *et al.* (2002) Eye contact detection in humans from birth. *Proc. Natl. Acad. Sci. U. S. A.* 99, 9602–9605
128. Zohary, E. *et al.* (2022) Gaze following requires early visual experience. *Proc. Natl. Acad. Sci. U. S. A.* 119, e2117184119
129. Senju, A. *et al.* (2015) Early social experience affects the development of eye gaze processing. *Curr. Biol.* 25, 3086–3091
130. Colombatto, C. *et al.* (2020) Gaze deflection reveals how gaze cueing is tuned to extract the mind behind the eyes. *Proc. Natl. Acad. Sci. U. S. A.* 117, 19825–19829
131. Redcay, E. *et al.* (2012) Look at this: the neural correlates of initiating and responding to bids for joint attention. *Front. Hum. Neurosci.* 6, 169
132. Pelphrey, K.A. *et al.* (2004) When strangers pass: processing of mutual and averted social gaze in the superior temporal sulcus. *Psychol. Sci.* 15, 598–603

133. Redcay, E. *et al.* (2016) Perceived communicative intent in gesture and language modulates the superior temporal sulcus. *Hum. Brain Mapp.* 37, 3444–3461
134. Deen, B. *et al.* (2020) Processing communicative facial and vocal cues in the superior temporal sulcus. *NeuroImage* 221, 117191
135. Grossman, E. *et al.* (2000) Brain areas involved in perception of biological motion. *J. Cogn. Neurosci.* 12, 711–720
136. Saxe, R. and Kanwisher, N. (2003) People thinking about thinking people: the role of the temporo-parietal junction in “theory of mind”. *NeuroImage* 19, 1835–1842