Received: 2 December 2022 | Revised: 13 June 2023 | Accepted: 12 July 2023

DOI: 10.1002/wics.1631

OVERVIEW

WIREs COMPUTATIONAL STATISTICS WILEY

A Bayesian theory of mind approach to modeling cooperation and communication

Stephanie Stacy[1] | Siyi Gong[2] | Aishni Parab[1] | Minglu Zhao[1] | Kaiwen Jiang[1] | Tao Gao[1,2]

[1]Department of Statistics, UCLA, Los Angeles, California, USA

[2]Department of Communication, UCLA, Los Angeles, California, USA

Correspondence
Tao Gao, Department of Communication, UCLA, Los Angeles, CA, USA.
Email: taogao@ucla.edu

Funding information
ONR Vision-Language Integration Program

Edited by: David Scott, Review Editor and Commissioning Editor

Abstract

Language has been widely acknowledged as the benchmark of intelligence. However, evidence from cognitive science shows that intelligent behaviors in robust social interactions preexist the mastery of language. This review approaches human-unique intelligence, specifically cooperation and communication, from an agency-based theory of mind (ToM) account, emphasizing the ability to understand others' behaviors in terms of their underlying mental states. This review demonstrates this viewpoint by first reviewing a series of empirical works on the socio-cognitive development of young children and non-human primates in terms of their capacities in communication and cooperation, strongly suggesting that these capacities constitute the origin of human-unique intelligence. Following, it reviews how ToM can be formalized as a Bayesian inference of the mental states given observed actions. Then, it reviews how Bayesian ToM can be extended to model the interaction of minds in cooperation and communication. The advantage of this approach is that non-linguistic knowledge such as the visual environment can serve as the contextual constraint for multiple agents to coordinate with sparse and limited signals, thus demonstrating certain cognitive architectures underlying human communication.

This article is categorized under:

Applications of Computational Statistics > Psychometrics
Statistical Models > Bayesian Models
Statistical Models > Agent-Based Models

KEYWORDS

Bayesian inference, communication, cooperation, shared agency, theory of mind

1 | INTRODUCTION

Even at its onset, artificial intelligence (AI) has been concerned with uncovering the nature of the mind by understanding the psychological mechanisms underlying fundamental human capabilities such as reasoning and thinking (Newell et al., 1957; Simon & Newell, 1962). In fact, the initial goal of AI was to "find how to make machines use language, form

Stephanie Stacy and Siyi Gong contributed equally to this work.

https://doi.org/10.1002/wics.1631

abstractions and concepts, and solve problems now reserved for humans" (S. Russell, 2019, p. 15). While there has always been a deep connection between cognitive science and AI, this connection has been particularly fruitful in the last 20 years, where AI has focused increasingly on incorporating insights from developments in cognitive science (e.g., Lake et al., 2017; Zhu et al., 2020).

One major branch at the intersection of these fields builds AI that mimics the minds of young infants by formalizing insights from developmental psychology. Intelligence can be reverse-engineered by first understanding the remarkably rich physical and social knowledge exhibited by infants and young toddlers. Aligned with this perspective, we start by reviewing recent insights on the socio-cognitive development of young children and non-human primates in terms of their capacities for communication and cooperation. There is strong evidence showing that these abilities differentiate humans from the rest of the animal kingdom (Tomasello, 2009, 2010; see Tomasello (2019) for a comprehensive review). We find this particularly impressive in visually grounded scenarios, as robust cooperation emerges in humans even given only limited usage of language.

Following a review of empirical results, we describe an agency based modeling approach which formalizes how agents take actions in the world based on their beliefs and desires. Moreover, this approach relies on a utility calculus (Jara-Ettinger et al., 2015, 2016) defining the rewards and costs associated with behaviors, which can be derived from an intuitive understanding of affordances in the physical environment. The power of this formulation stems from its ability to connect mental states to the environment and thus captures aspects of social and physical commonsense. Recently, these models have extended beyond predicting others to capture more complex social interactions (see Ho et al. (2022) for review).

Here, we review two research directions that extend from this modeling approach: cooperation and communication. By reformulating models of agency—which traditionally take an individual perspective—to a shared perspective, we show how humans' cooperative behaviors are motivated and supported by a shared understanding of the self and others as a group. Moreover, this framework allows communicative signals to be viewed as a type of rational, cooperative action for the purpose of establishing common ground between agents, especially when collaboration becomes challenging. By reviewing a body of computational modeling and empirical work, we offer insight into how a cognitive approach to modeling can advance AI. These models can capture behaviors exhibited by already incredibly intelligent 2–3 years old toddlers who effectively cooperate and communicate based on intuitive visual, physical, and social information before they become masters of language. A visually grounded understanding of cooperation and communication has been proposed as the origin of human-unique intelligence (Tomasello, 2010), and models built on these principles can lay the foundation for formalizing more sophisticated social interactions.

2 | COGNITIVE UNDERPINNINGS OF COOPERATIVE COMMUNICATION

Historically, language has been used as a critical benchmark to evaluate both human and machine intelligence. Early psychological theories emphasize language as arguably the most decisive characteristic contributing to humans' unique cognition and behavior, an extreme being the language determinism view (e.g., De Villiers & De Villiers, 2000; Gordon, 2004; Levinson et al., 2002; see Pinker, 2003, Chapter 3 for review). For example, starting from the late 1800s, language has been argued as a human instinct that interacts and competes with others to enable flexible human intelligence (Pinker, 2003). Later, language was regarded as a means to shape reality through grammatical structure (e.g., Whorf, 2012) and as a biologically pre-programmed system laying the ground for various cognitive capacities (Chomsky, 1983). Meanwhile, early intelligent machines have also been primarily concerned with language, focusing on question answering: the well-known Turing test evaluates artificial intelligence by distinguishing natural language responses from a human versus a machine (Turing, 2009 [1950]). From both sides, language and communication have been treated as a cognitive divider to distinguish human-level intelligence from those yet to be. It is agreed upon that there is something intrinsically special about language that makes it a necessary step to approach a better understanding of human intelligence.

While this perspective is also supported by the recent trend of AI that uses large neural models (Devlin et al., 2018; Radford et al., 2018; Vaswani et al., 2017) trained on huge corpora of human language to generate realistic text production, this review takes a different approach to modeling human intelligence. We argue that language should not be the starting point; instead, formalizing intelligence should start with examining the remarkable human-unique social interactions that humans exhibit without the help of language.

Overall, the language-focused approach has also been challenged by studies that demonstrate animals' remarkably intelligent behaviors and cognitive capacities without mastery of linguistic skills. For example, corvids exhibit complex behaviors in a variety of domains (Emery & Clayton, 2004) such as using and even manufacturing tools (Hunt, 1996). Perhaps the most noteworthy results come from studies on chimpanzees, which have been demonstrated to exhibit sophisticated physical cognitive abilities (e.g., Barth & Call, 2006; Herrmann et al., 2007), but more importantly, complex social cognitive abilities, such as understanding others' goals (Warneken et al., 2007; Warneken & Tomasello, 2006; Yamamoto et al., 2012) and intentions (Buttelmann et al., 2007; Tomasello, Carpenter, Call, et al., 2005); mentally simulating and physically manipulating others' perception and knowledge (Hare et al., 2000, 2006; Melis et al., 2006); and exhibiting some understanding of false beliefs, the ability to interpret actions based on others' beliefs even when they contradict with reality (Buttelmann et al., 2017; Krupenye et al., 2016). These studies demonstrate how physical and social cognition is ubiquitous in the animal kingdom and offers an alternative starting point for modeling intelligent behavior and reasoning.

Despite the incredible achievements made by our closest evolutionary relatives, a gap between how apes and young human toddlers interact socially emerges around 2–3 years of development, during which human toddlers start to exhibit an increasing level of qualitatively more advanced socio-cognitive capacities (Wobber et al., 2014). Specifically, such divergence manifests primarily in two domains: cooperation and communication.

The key difference in how humans as opposed to chimpanzees cooperate is commitment (see Tomasello (2019) for review). Behavioral studies demonstrate that children maintain robust cooperation by monitoring and evaluating commitment to a shared goal: they act under a normative sense of rights and obligations demanded by acting cooperatively. Toddlers regulate (1) both others' commitment by attempting to re-engage their partners when a collaborative activity is interrupted (Warneken et al., 2006) and (2) their own commitment by continuing putting efforts into a joint activity after receiving their own share of rewards prematurely (Hamann et al., 2012). They also often share the spoils equitably with their partners even when given the opportunity to easily monopolize them (Hamann et al., 2011; Warneken et al., 2011). Moreover, children acknowledge when they break a commitment (Gräfenhain et al., 2009) and express guilt for doing so (Vaish et al., 2016). Apart from commitment, evidence also shows that toddlers engage in joint planning during collaboration, in which they represent the self and others' roles in a reversible, flexible manner (Carpenter et al., 2005; Fletcher et al., 2012), and plan toward a joint goal that extends beyond their own perspective to predict their partners' future actions (Warneken et al., 2014). Taken together, humans represent joint activity from a collective point of view early in development (Tomasello, 2019), viewing collaborators as equal partners with complementary, reversible roles and obligatory commitments to achieve the shared goal.

This robust cooperative motive in humans is in sharp contrast with chimpanzees, whose collaboration is marked by "group behavior in I-mode" (Tuomela, 2007). Often, the dominant individual simply takes all rewards after a collaborative effort, demotivating the subordinate individual in future exchanges (Melis et al., 2011). When acting together in a collaborative task, chimpanzees do not regulate their own (Greenberg et al., 2010) or others' commitment (Warneken et al., 2006), nor do they represent a shared goal with complementary, reversible roles (Tomasello, Carpenter, & Hobson, 2005). It seems like chimpanzees use each other as social tools to achieve a greater individual reward (Tomasello, 2019), unlike humans who treat collaborating agents as equal partners with a shared goal in a joint activity.

It is argued in philosophy that robust, egalitarian cooperation has cognitive roots in a shared intentionality framework that is unique to humans (M. E. Bratman, 1992; Gilbert, 1992; Searle, 1990). Under shared intentionality, individuals view cooperators—including themselves—as a plural subject "We" with its own actions and mind. As a result, cooperation can be viewed as joint commitment to a shared goal while acting under that "We" mentality (Gilbert, 2013; Tomasello, 2010). This joint commitment, once established through a "readiness" from each participating agent, should be normatively upheld by each individual and cannot be rescinded unilaterally (Gilbert, 2013). Thus, a shared intention encompasses a collective representation of the group but is implemented at an individual level (Schweikard & Schmid, 2021).

Importantly, as an infrastructure for cooperation shared intentionality also provides a new perspective on the nature of human-unique communication. That is, communication serves as a mechanism for coordinating different agents' minds in a joint task so that each of their distinct versions of "We" can synchronize (e.g., Kleiman-Weiner et al., 2016; Tang et al., 2020; Wu et al., 2021). Communication can also be similarly understood through the lens of shared intentionality's account of cooperation. This logic is supported by studies showing that when cooperation is challenging and risky, human children—but not chimpanzees—use communication such as eye contact to offset challenges in coordination and achieve successful cooperation (Duguid et al., 2014; Siposova et al., 2018).

Communication enables increasingly complex, stable cooperation; at the same time, cooperation through shared intentionality offers shared knowledge in the context of a joint task which also imposes constraints on communication.

This makes it possible for humans to excel at efficiently disambiguating overloaded signals that carry multiple possible interpretations (Ferreira, 2008; Piantadosi et al., 2012; Winkler, 2015). This has been extensively studied in cognitive and developmental works on how humans produce and interpret pointing, gesturing, and pantomiming (Tomasello, 2010). For example, toddlers can adjust their communicative attempts according to listeners' levels of comprehension (Golinkoff, 1993), communicate about absent but mutually known entities by pointing (Liszkowski et al., 2009), and interpret ambiguous requests by referring to commonly known objects (Moll & Tomasello, 2006; Tomasello & Haberl, 2003). These tendencies, however, are not observed in chimpanzees. Chimpanzees effectively interpret a reaching pointing gesture when a task is framed as competitive (Hare & Tomasello, 2004) but surprisingly struggle to understand the same gesture when framed as cooperative because they fail to understand the helping intention behind pointing (Tomasello et al., 1997). Together, this suggests that overloaded communication, where a signal is consistent with multiple interpretations, is uniquely human as it is afforded by shared intentionality, which imposes strong assumptions on what is appropriate and relevant to communicate.

Shared intentionality is critical to modeling cooperation and communication, but in order to build models of shared agency, we first need to understand the mechanism of single agency by introducing the concept of theory of mind (ToM). In the following chapters, we first show how to model ToM through Bayesian inference (Chapter 2), and then we review ways to augment ToM to both cooperation (Chapter 2) and communication (Chapter 2) as well as how these augmented forms can be combined into a model of cooperative communication (Chapter 2).

3 | MODELING INDIVIDUAL INTENTIONALITY: BAYESIAN ToM

ToM refers to the ability to attribute an action to the underlying mental states that may have produced it, such as beliefs, desires, and intentions (Dennett, 1987; Gopnik & Meltzoff, 1997; Premack & Woodruff, 1978; Wellman, 1992). For example, if you see someone take a call in the library (action), this can be interpreted as (1) the person thinks it is appropriate to do so (belief), or (2) this call is extremely important and urgent to the person (desire). As a fundamental building block of social interaction, ToM plays a profound role in human society. For example, in our legal system, a guilty act (*actus reus*) alone is insufficient for conviction in many crimes: a guilty mind (*mens rea*) is also necessary (Duff, 1990). In addition, ToM has been extensively studied in developmental psychology (e.g., Gergely et al., 1995; Spelke & Kinzler, 2007) where various studies have supported the early onset of this social capacity in humans. For example, infants as young as 6 months can interpret desires and goals from others' reaching movement (Woodward, 1998), and 21-month-year-old toddlers can interpret an ambiguous request made by an adult by cooperatively inferring the adult's intention (Grosse et al., 2010).

Formally, ToM has been modeled using a Bayesian formulation (BToM; Baker et al., 2009). To provide an interpretation for an action, ToM first requires a model of how actions are generated. This can be captured by a forward planning process (Figure 1) unfolded through a generative model, which leverages the assumption that agents act rationally and consistently according to their mental states to maximize their expected utility in terms of the underlying mental states and constraints of the environment.

Crucially, Bayesian inference provides the formalism needed to capture the inverse planning process of ToM (Figure 1). Bayesian inference allows reasoning over potential generative models, allowing agents to infer the mind (m)—beliefs (b), desires (d), and intentions (i)—of others (Equation 1) that can best explain their observed actions (a) in the physical environment (w).

This formulation was first tested by modeling human inferences of an agent's goal in a 2-D grid world (Baker et al., 2009). By observing only a part of an agent's trajectory, the model successfully inferred the goal of the agent; moreover, the temporal dynamics of the model's inference matched those of humans. In the same study, this approach to goal inference was also extended to adapt to changing or complex goals made of multiple subgoals.

In follow-up studies (Baker et al., 2017; Baker & Tenenbaum, 2014), BToM has been extended to partially observable environments, showing that a rational interpretation of an agent's action requires reasoning over the agent's uncertainty and ignorance of the environment as well as the agent's goal preferences. In another study with a perceptual chasing paradigm (Gao et al., 2019), BToM was shown to capture human's effectiveness and efficiency in perceiving animacy by assuming that human ToM inference is rational but limited by attention and working memory. This capacity constraint of BToM model captures human's ability to detect predator–prey relations between pairs of targets in various conditions.

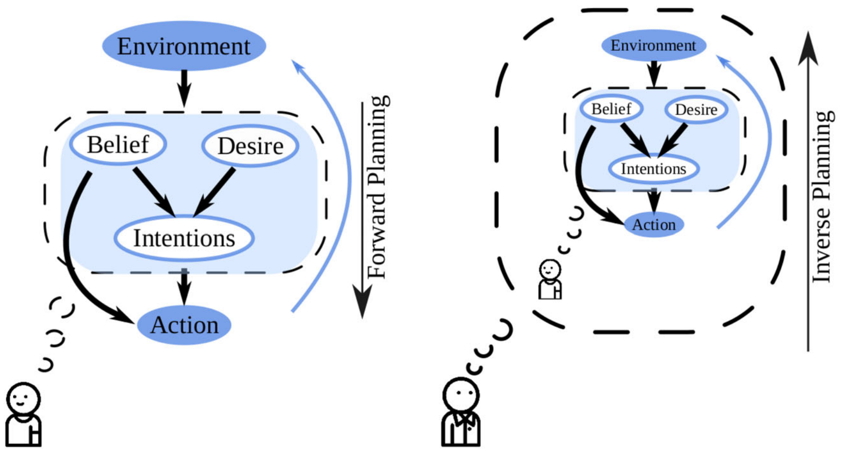

FIGURE 1 In this framework, action can be generated through a forward-planning process (left figure, Equation 2) by reasoning over the environment and mind to produce a rational decision. Actions are sampled from the soft-max of an agent's utility function, with $\beta \in [0, \infty)$ describing the agent's degree of rationality. When $\beta = 0$, the agent is modeled as acting randomly, and as $\beta \rightarrow \infty$, the agent deterministically chooses the action with the highest expected utility. Building on top of forward planning, ToM relies on inverse planning (right figure, Equation 3), in which an agent infers the mind given the action and the environment.

$$P(\mathrm{m}) = P(\mathrm{b}, \mathrm{d}, \mathrm{i}) = P(\mathrm{b})P(\mathrm{d})P(\mathrm{i} \mid \mathrm{b}, \mathrm{d})), \tag{1}$$

$$p(a \mid m, w) \propto e^{\beta E[U(a,m)|w]}, \tag{2}$$

$$P(\mathrm{m} \mid a, w) \propto P(a \mid \mathrm{m}, w)P(\mathrm{m}|w). \tag{3}$$

An inverse planning approach is also supported by developmental evidence showing that even at an early age children employ a naive utility calculus to infer agents' mental states from their actions by implicitly assuming that agents choose goals and actions following reward-maximization and cost-minimization (Jara-Ettinger et al., 2015, 2016). For example, infants as young as 10 months of age can infer the desirability of a goal for an agent from how much effort the agent would exert on achieving that goal (Liu et al., 2017).

Theoretically, ToM and planning are two distinct and inverse processes: planning is about synthesizing actions from the mind, while ToM is about inferring the mind from observed actions. This is supported empirically by the fact that many animals are able to plan for the future (e.g., Clayton et al., 2003), but it remains unclear how good their capacity for inferring others' minds is (see Penn & Povinelli, 2007 for a review of the debate). However, on an algorithmic level, there may be room for better integrating the inference and planning processes. In BToM approaches, the planning engine is treated as a black box with little concern about how planning is achieved. Recent research works show that the planning process itself can be modeled as a probabilistic inference process (e.g., Botvinick & Toussaint, 2012; Friston et al., 2021), suggesting that the Bayesian graphical model can act as a unified approach integrating the planning and BToM inference in the future.

4 | EXTENDING BToM FOR COOPERATION

ToM provides a useful scaffold for both inference and planning. At the same time, it has been shown to successfully model the dynamics between an actor in the environment and a disconnected observer trying to understand the mind of that actor in a variety of social situations. However, cooperation requires individuals to be both actors *and* observers at the same time. Thus, to capture the dynamic interaction inherent in cooperation, a mixture of inverse planning (for understanding partners) and forward planning (for generating one's own cooperative actions) is necessary. On top of that, while ToM models an agent's mind from an individual perspective, such that "my mind understands your mind," developmental studies show that children's cooperation is often marked by a joint perspective for a "meeting of minds" (Schelling, 1980). For this reason, one branch of modeling has extended the individual perspective of traditional ToM to incorporate shared agency, or shared intentionality, which is the focus of this section.

One challenge of modeling shared agency is its paradoxical nature. Philosophical discourse posits that there are two core principles of shared intentionality (Schweikard & Schmid, 2021). First, shared intentionality is a single concept that is inherently shared, making it qualitatively different than a summation or aggregation of individual minds. For example, "we walk together" cannot be reduced to "you and I walk on the same street." Second, because intentions are privately formed within one's mind, a shared intention can only be represented at the individual level. At the same time, sharing an intention is only possible when everyone voluntarily agrees and acknowledges its existence. These two claims give rise to tension—how can a collective intention that goes beyond any single individual simultaneously only makes sense at the individual level?

Philosophers reconcile this paradox from different angles. One approach focuses on joint commitment and proposes that a shared intention is only realized when two or more individuals are willing to be "jointly committed to espousing a goal as a body," or "plural subject" (Gilbert, 2013, p. 30, 37). Another approach, which highlights the coordination of plans produced by individual agents, proposes that a shared intention is an intricate mesh of individual intended plans and their interrelations, aligned with each other and commonly known to all (M. Bratman, 1987; M. E. Bratman, 1992, 2013).

Formulations of shared intentionality first arose in the 1990s and primarily used logical language. These modeling approaches heavily featured Bratman's joint planning account. One study (Grosz & Kraus, 1996) extended existing individual plans by formulating "sharedPlans" which contained both individual and mutual beliefs about how actions should be done. On the other hand, another model (Levesque et al., 1990) characterized the goal and intention of a group as "acting like a single agent" by defining a "joint persistent goal (JPG)" that replaced individual belief in a "personal goal (PGOAL)" with a mutual belief. However, early shared agency formulations remained largely rule-based and disconnected from more recent Bayesian approaches for modeling ToM.

More recently, a trend of probabilistic accounts of shared agency modeling based on BToM has emerged. Returning to the central challenge of shared agency, a shared agent is an irreducible single agent that does not exist physically. How is it possible to formalize a shared intention when it is not even real? The answer is in Bayesian inference's capacity to enable causal reasoning. Thus an agent can counterfactually reason: if the actions taken by myself and others had been rationally generated by a centralized controller, what were the most likely beliefs or desires that could explain our joint actions (Equation 4; Kleiman-Weiner et al., 2016)?

$$P(\mathrm{mind}_{\mathrm{We}}|\mathrm{action}_{\mathrm{We}}) \propto P(action_{We}|\mathrm{mind}_{\mathrm{We}})P(\mathrm{mind}_{\mathrm{We}}). \tag{4}$$

One approach to BToM modeling has treated shared intentionality as a "meshing of plans" (M. E. Bratman, 1992, 2013). One such study, developed in a grid-world coordination game, models how participants infer whether another agent is a cooperator or competitor based on whether past actions in their history of interactions can be explained by joint planning (Kleiman-Weiner et al., 2016). Specific to cooperation, the model augments each agent's individual plan to form a joint plan, taking the form of "I intend that we J" (where J is a joint action). Specifically, an agent aiming to achieve a group goal generates a joint policy that maximizes the joint utility of both agents, from which the agent derives their individual action by marginalizing out the partner's actions. This type of joint planning has also been proposed generically as a potential mechanism to make inferences about more complex team structures involving multiple levels of cooperative and competitive behavior (e.g., A is cooperating with B against C; Shum et al., 2019).

Another BToM approach uses a cooking domain to model shared agency coordination (Wu et al., 2021). Here, agents work together to quickly complete recipes in a 2-D environment in a game inspired by the video game *Overcooked*. In this game, each task is made of multiple subtasks with multiple possible orderings. To complete a recipe, agents work either in parallel on different subtasks or collectively coordinate on the same subtask. To decide one's action, an agent needs to first infer what all agents in the environment are likely working on by observing their actions. This is accomplished by a Bayesian Delegation (BD) model, in which each agent infers the most likely allocation of subtasks given a fictitious centralized planner, assuming a shared goal of collectively completing the recipe in the least amount of time. A low-level planner then helps the agent plan the next best action according to this belief distribution. BD agents are able to carry out their subtasks efficiently to collectively complete recipes across 2-D environments of ranging difficulties; they outperform alternative models during self-play and hybrid dyad collaboration and in predicting human inferences. Incorporating complex tasks and environment elements, this work captures the reactive dynamics of coordination under a shared task by using a joint form of ToM reasoning.

Another BToM work adopts the joint commitment account of shared intentionality established by philosophers (Gilbert, 2013) and is further supported by developmental work reviewed in Section 1. Named the Imagined We (IW), the model was first formulated in a study (Tang et al., 2020), where "We" is formalized as a single autonomous agent with its own beliefs, desires, and intentions. By observing the joint action of all agents in the shared environment, each agent infers its own version of “We” without explicit communication. Then, each agent forms a plan of how “We” would rationally pursue a goal by simulating (imagining) a centralized planner that outputs joint actions. In addition to taking an action, each agent also expects collaborators to take their respective actions as demanded by “We” and actively monitors these newly generated actions to update its inference of “We.” Each individual determines what “We” believes or wants by observing what “We” has done. Over time, different versions of “We” align across collaborators, allowing them to jointly commit to the shared goal. Importantly, this convergence occurs even when “We” lacks detailed information about each agent's action course, or when each agent actually sees a slightly different version of the world due to perceptual noise.

A recent follow-up study using a cooperative hunting setting along with psychophysics data from human adults (Tang et al., 2022) shows that the IW model effectively captures humans' robust teaming patterns in cooperation. When there are multiple equivalent targets, IW achieves greater accumulated rewards with higher quality hunts (more consecutive touches of a target) across games which vary in the number of available targets. The study also contrasts the IW model and humans with a baseline Reward-Sharing (RS) model implemented by the Multi-agent Reinforcement Learning (MARL) framework (Lowe et al., 2017) without any shared agency structure, showing that while the RS model performs well when presented with one target, its teaming falls apart in the face of multiple targets. Meanwhile, in a hybrid team simulation where a model hunter replaces the trajectory of a human hunter, the IW but not the RS model better mimics the intentions of the human hunters it replaces. Corroborating this finding, the necessity of shared agency is also demonstrated by another study (Zhao et al., 2021), in which without action cost, all RS agents are motivated to hunt and share the spoils. However, even with a minimal action cost, RS agents become reluctant to contribute and perform worse as a team, manifesting the free-rider problem (Olson, 1989). Together, these studies highlight how the capacity for shared agency in human cooperation can be realized using computational modeling and provide evidence that shared intentionality is an important cognitive structure that allows models to emulate aspects of intelligent human behavior not captured by purely reward-driven MARL approaches.

5 | EXTENDING BToM FOR COMMUNICATION

The second direction BToM has been extended is toward scenarios involving communication, another type of social interaction with a variety of situational uncertainties and dynamic exchange of actions and minds (Richardson et al., 2007; Shockley et al., 2009; Sperber & Wilson, 1986). Communication in real life is often sparse and ambiguous (Ferreira, 2008; Piantadosi et al., 2012). For example, a pointing gesture toward an empty glass means different things depending on context: a customer in a bar is likely asking for a refill by indicating the emptiness of the glass whereas a person washing dishes is likely asking someone to pass the glass by emphasizing its distant location. To approach the immense flexibility inherent in human communication, an agency-based framework is a promising approach as it imposes strong constraints on communication beyond signals. These extra constraints can facilitate both the production and interpretation of overloaded signals. The key insight to extending BToM to capture communication is to treat signaling as a type of rational action that can be planned and reasoned over (Austin, 1975; Cohen & Perrault, 1979; Franke, 2009; Parikh, 1991; Searle & Searle, 1969) instead of a code with a fixed mapping (MacKay, 2003). As a result, interpreting signals follows the same framework as generating actions from underlying mental states.

This insight has been recently formalized in the rational speech act framework (RSA; Frank & Goodman, 2012; Goodman & Frank, 2016). According to an influential set of principles on linguistic pragmatics called the Gricean maxims, communication should be truthful, concise, relevant, and straightforward (Grice, 1975). The framework proposes that a signal is used to convey information about the states of the world in a maximally efficient way. In its initial formulation, RSA was used to solve referential signaling games (Lewis, 1969; Wittgenstein, 1953), where a speaker sends a potentially overloaded signal to help the listener identify an intended referent among a set of potential referents by reasoning about the linguistic context. For example, the cards in Figure 2 each share one of the two features with each other: the numbers (four and six) and the suits (clubs and spades). The question is, when a signal only conveys information about one feature which is shared by two cards, can a pragmatic listener identify the intended card better than a random guess? Adhering to the Gricean maxims, when a speaker says “clubs,” assuming the speaker is truthful, the options for the listener are limited to the four (left) or the six (middle). Assuming the speaker is concise and relevant, the listener realizes the signal itself is sufficient for him to make a choice. Given the two possible cards, the listener can then reason, “If the speaker refers to six, she could have said six, since it is the most straightforward, unique information among all cards; but

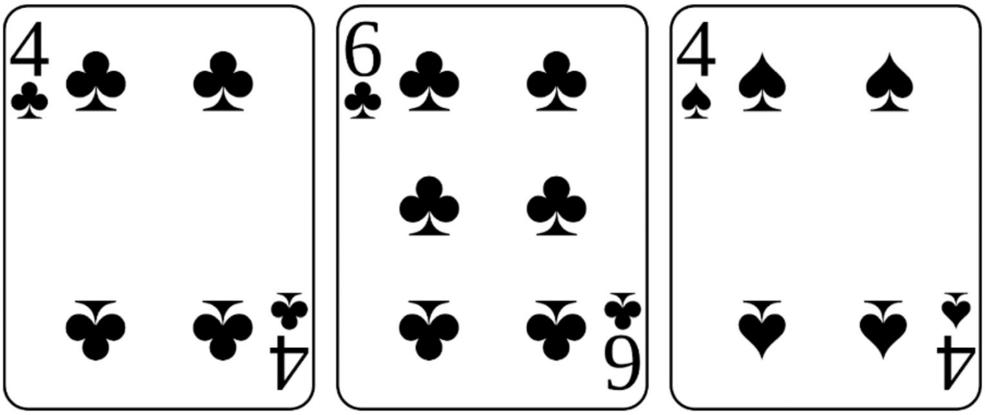

FIGURE 2 An example of the referential signaling games. Given the cards, the speaker says "clubs." While clubs can literally refer to two cards: the four and six, a pragmatic interpretation would choose the four of clubs by reasoning if the speaker had been referring to the six of clubs, "six" would have been a better signal to send.

she didn't, so it must be the four." As such, RSA models pragmatics as a Bayesian counterfactual reasoning process, not only focusing on the chosen signal but also on alternative options that are not chosen.

Mathematically, this reasoning can be captured by a utility function, integrating both the gain and cost of sending a signal. The gain of a signal is evaluated by the probability that the listener can correctly identify the intended card. Then, sending a signal can be understood as a decision-making problem, in which an agent chooses a signal to maximize its expected utility. Traditionally, this utility measures the listener's interpretation of the world state based on linguistic context. However, follow-up work has taken non-linguistic context into account during utility evaluation such as listener actions (e.g., Benz, 2004; Qing & Franke, 2015; Van Rooy, 2003) and visual context (e.g., Anderson & Dillon, 2019; R. D. Hawkins et al., 2021).

RSA is a recursive model: for the signaler to generate a signal, she models how the receiver will interpret it; for the receiver to interpret the signal, he models how the signaler would have generated it. Reasoning like this can develop iteratively with increasingly sophisticated signalers and receivers, each building on top of less sophisticated ones. Importantly, the entering point of this recursive process is a literal communicator, who only samples signals or interpretations based on whether they are true given the states of the world, thus requiring only a model of the world without any model of one's partner.

Formally, a literal speaker (p_{sl}) samples a *signal* that truthfully reflects the state of the world, *state*, according to a prior (which is often uniform). For example, if an object has two features, a literal speaker may randomly sample one of the features to send. From this entering point, more sophisticated communicators can be built on top.

Using the literal speaker, a listener model that uses Bayesian inference (p_l) can be built to infer the states that are consistent with the *signal* (Equation 5).

$$P_l(state|signal) \propto p_{sl}(signal|state)p(state). \tag{5}$$

A pragmatic RSA speaker (p_{sp}) uses this listener model to reason how signals will be interpreted in terms of the states it could refer to, selecting the signal that is expected to yield the maximal utility (Equation 6). Here, the degree of rationality in communication signal can be modeled by a noisy utility maximization (soft-max) by manipulating the $\beta \in [0,\infty)$ parameter, which determines the signal sent (Equation 7).

$$\mathbb{E}[U(signal,state)] = p_l(state|signal), \tag{6}$$

$$P_{sp}(signal|state) \propto e^{\beta \mathbb{E}[U(signal,state)]}. \tag{7}$$

Under this framework, models of more complex listeners can be built to support increasingly sophisticated speakers in a recursive process.

RSA was developed to model pragmatics and has been incredibly successful at capturing linguistic phenomena including metaphor (Kao et al., 2014), redundancy (Degen et al., 2020), and convention formation (R. X. D. Hawkins et al., 2017). On top of this, BToM expands context to include available *actions* in the environment as an additional constraint on what is rational to say, allowing agents to communicate with actions that have no predefined meaning.

For example, in many cases, we can gesture to demonstrate or teach something when the gesture itself does not have a single, set meaning. When I am walking with my young cousin on the beach and encounter a jellyfish, not only do I want to avoid stepping on it (instrumental action), but I also want to teach my young cousin that jellyfish are dangerous through my moves (communicative action). While stepping over the jellyfish is the most efficient action to avoid the jellyfish, exaggeratingly walking around it serves better to demonstrate my intention to my cousin. Acts such as these are communicative demonstrations, which are characterized by the distinction between *doing* something to achieve one's own goals and *showing* a task to someone else. In the case of showing, the utility of an action is evaluated by both its instrumental values as well as its effectiveness in communicating one's intention to others.

Communicative, or pedagogical, demonstrations through non-linguistic actions have been modeled in a grid-world task where the goal is to reach a certain location by walking over tiles colored based on their reward (Ho et al., 2016, 2021). Here the agent is an instrumental actor who plans a rational route to minimize costs. However, in some cases, the actor must also show the underlying reward structure of the world to an ignorant observer in which case she takes on the role of communicative demonstrator. Aligned with RSA, the process of generating communicative actions is cast as a decision-making problem. However, the key difference is that in a typical decision-making problem, a signaler only considers the effect of her actions in the physical states; but here, the model needs to consider the impact of a signaler's action on the receiver's mental states. As a result, the state space of decision making is augmented, now including both the fully observable environment as well as the receiver's beliefs. Since those beliefs are not directly observable to the signaler and must be inferred, the task becomes a partially observable MDP (POMDP) problem (Kaelbling et al., 1998). Thus, communicative actions can be solved by applying a traditional POMDP solver in a joint physical-belief space. The resulting agent of the pedagogical model manages to show its knowledge of the world to an observer while still achieving its own instrumental goals.

In this work, the integration of RSA and ToM is still limited, since the receiver only passively observes the signaler and does not have the capacity to act on his own beliefs. In cases where the receiver can act, the interaction between ToM and communication becomes more interesting as the potential actions the receiver can take now serves as additional context to constrain the interpretation of a signal. Here we review recent studies that have developed this integration.

The first work augments the classic RL multi-armed bandits paradigm, in which a receiver aims to maximize their rewards which are associated with features of the items they can choose, to include communication (Sumers et al., 2021). This extends classic referential signaling games (Lewis, 1969; Wittgenstein, 1953) to include both the receiver's beliefs and actions. To apply this game to the poker example, each feature of the card now has a different reward known privately by the signaler (see Figure 3). Importantly, one model proposed in this work solves it by defining the utility of the signal not based on truth value like RSA, but based on the beliefs of the receiver and the action he takes after receiving the signals. Thus, the utterance needs not to be absolutely truthful, but is encouraged to be useful to the receiver. In the poker case, the signaler wants the receiver to choose the four of spades which results in the highest combined feature score. By signaling "the spade feature is worth 3 points," the signaler offers false information about the world (spades are only worth 2 points) but leads the receiver to believe the four of spades will lead to the maximal reward making him more likely to select it. This strategy effectively allows the receiver to achieve better performance in the current context.

In parallel, another work (Jiang et al., 2021) has developed the integration of RSA and BToM in a way that departs from the referential signaling game. In cognitive science and cognitive linguistics, it has been long recognized that the key to human pragmatic reasoning is *relevance* (Sperber & Wilson, 1986). Humans excel at intelligently interpreting a signal based on what is relevant in the context, such as deciding whether pointing to an empty glass means passing it or filling it. However, relevancy has been a notoriously elusive concept to model. In this study, the relevance of a piece of information is evaluated by how much it can improve the receiver's well-being by introducing additional information into his belief states. This idea, which follows from Gricean maxims, implies that the rational interpretation of a signal is the one that induces the largest increase in the receiver's expected utility among all possible interpretations.

Imagine a scenario where a young and an experienced hunter both see the same broken branch while hunting. The younger hunter passes by the branch, thinking it has been broken by a strong wind. However, the experienced hunter ostensively points out the broken branch. Now, the young hunter's interpretation of the broken branch changes

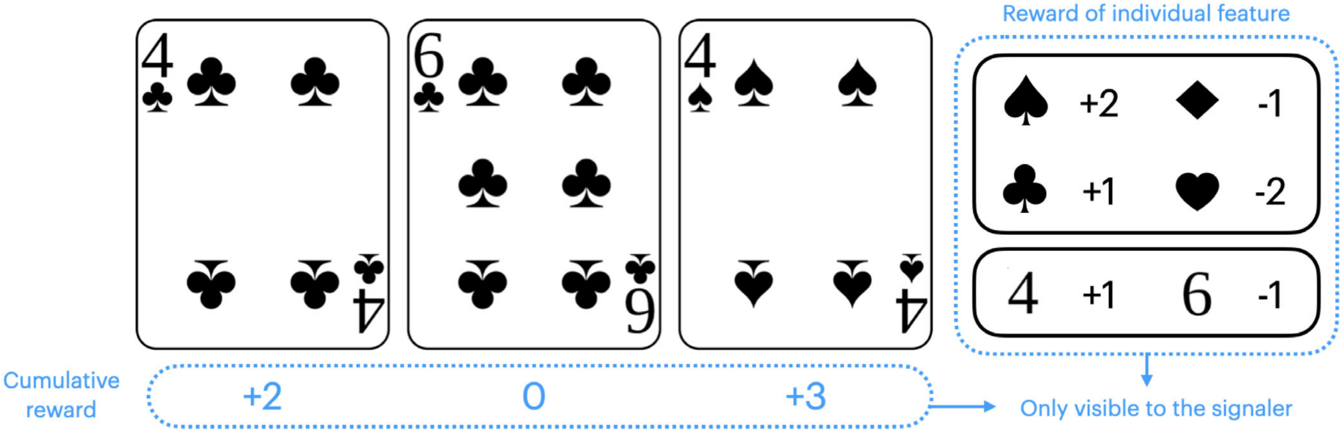

FIGURE 3 A poker example of the extended referential signaling game. The reward of each object at the bottom is determined by the cumulative reward of its individual features (suit and number) on the right, which is only known to the signaler but not the receiver. The goal of the signaler is to produce an utterance to communicate the reward of a feature to maximize the receiver's overall reward. Based on the observed utterance, the receiver selects one out of three possible cards that always shares one feature with the other.

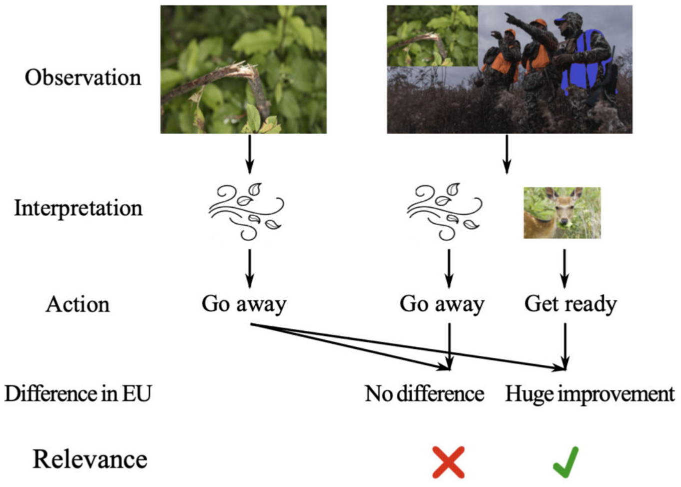

FIGURE 4 Different interpretations of a signal may lead to different action consequences. The most relevant interpretation should be the one that elicits the most improvement on one's expected utility.

dramatically—the branch indicates possible nearby prey. This scenario can be intuitively understood based on the definition of relevance. Suppose that there are two possible interpretations of the pointing gesture: the broken branch is caused by (1) the prey or (2) the wind. Learning that the branch was broken by the wind would not affect the hunter's expected utility as he would have passed by regardless. However, learning that the branch was broken by prey would lead to dramatically different expected utilities stemming from either a successful or failed hunt. Therefore, the more relevant interpretation of the pointing gesture in this scenario should be the cause of prey (Figure 4). This intuition can be tested in the Wumpus world, a classic AI paradigm modeling decision making in a partially observable world (S. J. Russell et al., 2010). In this task, a hunter aiming to hunt the invisible monster, Wumpus, can only infer its location through the stench it emits to its surroundings. The study augments the Wumpus world by adding a guide with knowledge of the location of the Wumpus. This guide can help the receiver, but only with extremely limited and overloaded communication—pointing to what the hunter already observed or not. The pointing gesture cannot provide any additional observation to the hunter other than emphasizing what the hunter already knows about the world. Still, this highly overloaded information can greatly improve the hunter's estimation of the world and his overall performance, especially when his perception of the world is noisy.

These works share a common theme: communication is treated as a social tool to coordinate minds and actions under asymmetric information. From a theoretical perspective, this can be connected to the idea of paternalistic helping, a concept originating from developmental psychology that deals with overriding others' desires or actions because "I know better" or "it's good for you," often occurring in a caregiver-child context (Jacobsson et al., 2007; Sibicky et al., 1995). For example, a mother might say no to her young child who thinks it is fun to run around with a rake, because the parent thinks it is dangerous. In this case, the signaler (the mother) knows more information about the world than the receiver (the child); she predicts the future action of the child (running around with a rake) given the child's desire (to have fun), and evaluates the child's action (it is dangerous as the child could easily get hurt). Then, the parent produces a signal ("no") that is the most relevant in the scenario (prevent the child from getting hurt).

The challenge, however, lies in how to evaluate the receiver's actions. This is because the utility function may not be straightforward, making it hard to decide what is "good" for the child. While one approach is to evaluate actions based on what the receiver wants and knows, this creates a dilemma. When the signaler's utility evaluation is based on the receiver's knowledge, the signaler will not deliver bad news because it does not improve the utility of the receiver, leaving the receiver "happily ignorant." Similarly, the child might get upset if his mom prevents him from playing with the rake, but that is not necessarily bad because he does not know how dangerous the rake can be. To approach this problem, a solution for the signaler in this case is to *predict* actions using the receiver's beliefs and desires, but then *evaluate* those actions using her private knowledge due to her more comprehensive knowledge about the world. This action evaluation of a crossing-over of the signaler's and the receiver's minds is treated as paternalistic evaluation.

One work focusing on this idea integrates paternalistic evaluation and RSA into a model of Paternalistic Communication (PaCo; Stacy et al., 2022). While RSA provides a framework for rational communication, paternalistic evaluation provides a stronger mechanism to disambiguate overloading in signals by comparing at the change in the receiver's actions before and after communication. PaCo was tested in a cooperative communication task where the goal was to collect rewards and avoid punishments stored in boxes (Misyak et al., 2016). The task included asymmetric abilities (only the receiver could open boxes), private information (only the signaler knew which boxes contained rewards), and shared information (e.g., both agents sometimes knew the total number of boxes containing rewards). Moreover, the signaler could mark a limited number of boxes with tokens to help the receiver. However, the meaning of a token was completely overloaded: it could mean "open this reward" or "avoid this punishment" depending on the scenario. By integrating paternalistic action evaluation with signaling, PaCo was able to reason flexibly to perform well at the task and successfully capture human-like signaling patterns (see Figure 5 for example).

In summary, RSA and ToM form two essential building blocks that allow promising formulations of overloaded communication. This relies on two points: (1) communication should be treated as an action with a utility just like other instrumental actions; and (2) understanding the meaning of a signal can be formulated as a ToM problem, where one infers the latent mind of an agent to help interpret the signal. Moreover, it is noteworthy that communication refers to more than beliefs—agents are also capable of acting upon the world based on the outcomes of communication. The actions agents can take and the consequences of those actions can then be used as constraints to facilitate the interpretation of ambiguous signals. This makes human communication not only overloaded but also indirect, introducing a gap between what is said (e.g., pointing to a broken branch) and what is meant (e.g., getting ready to hunt).

6 | A SHARED AGENCY MODEL OF COOPERATIVE COMMUNICATION

Thus far, we have discussed two extensions of ToM that can be used to model social interactions. One direction is to transform individual ToM into a shared intentionality framework to model cooperation, and the other is to integrate ToM with RSA to model communication. So far, these two aspects of ToM have been largely treated as independent, mutually exclusive problems. In this section, we explore how these two lines of research have begun to converge in a shared agency model capable of handling overloaded signaling. This integration stems from deep theoretical roots: communication is intrinsically cooperative, and arguably humans evolved the capacity of communication to coordinate increasingly sophisticated cooperation (Tomasello, 2010). As a result, communication can be viewed as another type of social tool allowing people to get things done together (Bruner, 1985; Vygotsky & Cole, 1978). Thus, models of communication can be built from models of shared agency which allow agents to reason about joint goals, joint attention, and common ground. The cooperative foundation of communication has been limited in current works due to asymmetric capacities between receivers and signalers: either there are no actions (classical RSA) or only receivers can act (more recent works). However, a truly cooperative task should embody an action-oriented partnership where all communicators can also participate in the joint task, giving rise to a communication model based on

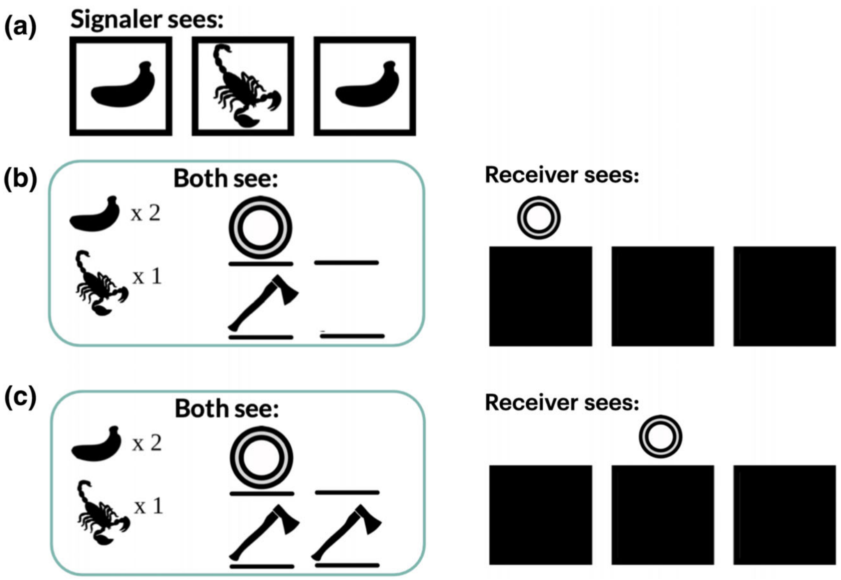

FIGURE 5 Signaling task examples: using the logic of paternalistic helping, the signaler predicts which boxes the receiver will open using common information including how many total boxes contain rewards/punishments (bananas/scorpions), how many boxes the receiver can open (axes), and how many signals are available (tokens; two example setups shown in b and c). The signaler then evaluates how good that action is with private knowledge (shown in a). (a) Signaler's private knowledge of where the rewards and punishment are located. (b) One ax setup common knowledge: both partners know two boxes contain rewards, the signaler can mark one box, and the receiver can open two boxes. Humans and both model signalers are most likely to place a token on a reward, for example on the leftmost box (shown), for the receiver to see. (c) Inversion setup common knowledge: both partners know two boxes contain rewards, the signaler can mark one box, but now the receiver can open two boxes. Humans and PaCo, but not RSA, tend to flip the meaning of their signal to indicate punishment, placing a token on the middle box for the receiver to see. This signal is expected to maximize the expected action utility of the receiver.

shared agency. This reflects the essential idea of this section: that communication is for and by cooperation. Below we review how to study communication under a shared agency approach.

Cooperation is often spontaneous—agents may decide the interests of the group on the spot, based on all other agents' observed actions so far. Thus, cooperation becomes challenging when agents have asymmetric knowledge and perspectives, as well as when they face an overwhelming amount of information from the environment. In these cases, communication is especially important as it serves to help agents align their individual versions of the group and converge on a shared goal. At the same time, communication is also afforded by cooperation, as the assumption of cooperation constrains how signals should be generated and interpreted. That is, signals should be (1) maximally informative (in disambiguating overloaded information), and, equally importantly, (2) jointly efficient for all agents. This means the relevance of a signal's interpretation is no longer calculated by the improvement of utility from any single individual's perspective but from our shared perspective. Thus, on top of Gricean maxims, one must also treat one's partner respectively and approach problems from a joint perspective. In other words, it is unreasonable for you to ask me to do things that are, from a "We" perspective, my responsibility (i.e., more efficient for me). From this, modeling cooperative communication under a shared agency framework follows naturally.

This reasoning is demonstrated in a two-agent interactive signaling task where agents execute joint plans in a grid-world environment (Stacy et al., 2021). Similar to the existing referential signaling games introduced in Section 5 (Lewis, 1969; Wittgenstein, 1953), the task involves a target referent only known to the signaler, and the signaling content is limited to only one feature. Here, the shared goal is not just for the receiver to infer the correct reference, but for either agent to reach the target item. Thus, joint planning which helps decide who should walk to the target adds extra constraints to the features of the items. In certain environments, it is maximally efficient for the signaler herself to walk to the target; in other cases, it is more efficient for the receiver to walk, in which case the signaler may send a signal. With a single feature alone, the signal can be too ambiguous for identification even using RSA pragmatic reasoning. For example, in Figure 6, suppose the green square is the target, it is impossible for a receiver to simply use pragmatics to identify the target as the number of shared features all equal to two—no feature is more informative than the other.

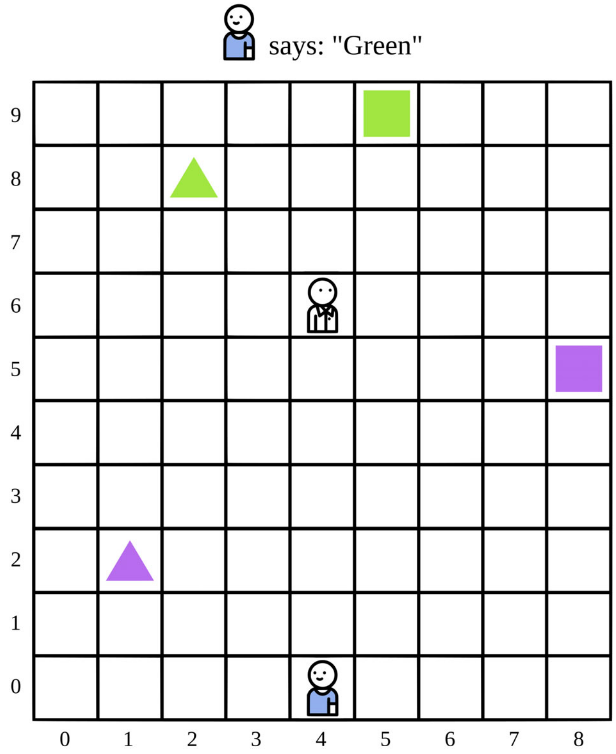

FIGURE 6 An example of a scenario where it is impossible to disambiguate between the green triangle and the green square solely using pragmatics. The receiver has to use the joint utility as knowledge to re-reason over the possible options.

Instead, using the logic of joint planning, the fact that there is a signal sent regardless of its content eliminates the purple triangle, because a respectful, cooperative partner should not ask the receiver to walk somewhere that is closer to herself. With the three items left (green triangle, green square, purple square), the receiver can now easily rely on the pragmatics of the signal ("green") to determine that the target is the green square (otherwise it would have been more straightforward for the signaler to say "triangle" for the green triangle). This is a scenario in which a combination of RSA and joint utility is required for successful, cooperative communication.

This combination of RSA and joint planning has been recently formalized in the Imagined We for Communication (IWc) model (Stacy et al., 2021). This model is essentially an extension of paternalistic evaluation with a crossing of minds between action predictions and action evaluations, only that the mind for action prediction is expanded from an individual mind to a joint "We" mind. Specifically, for each signal, the IWc signaler predicts a distribution of joint actions for both agents by sampling a joint mind; she then evaluates the actions using her private knowledge and selects the signal that maximizes the joint outcome (Figure 7).

Compared to the RSA model, IWc exhibits several advantages in the results. First, as the number of items in the environment increases with the signal content limiting to one feature only, IWc is robust to the escalated amount of overloading and difficulty of reasoning, sending, and interpreting ambiguous signals with much higher success. This result shows that joint planning provides powerful constraints on producing and interpreting the relevance of signals from a "We" perspective, which is absent in the individual reasoning in RSA. Second, when agents can signal all the features in one utterance, a mild cost added to the act of communication is enough for IWc to keep the signaling short (using fewer bits). This result is not trivial, as in the fields of machine communication and information theory, much effort has been made to minimize the information needed to transmit a message by optimizing the coding schemes (MacKay, 2003). IWc achieves a similar goal from a different approach, through constraining the interpretations of signals, not by making a better predefined coding scheme, but by using cooperation as a context for joint planning on the fly. In addition, a second experiment tests IWc's performance with increased recursive reasoning in communication compared to RSA. Despite its effectiveness, RSA relies on recursive partner reasoning to make increasingly sophisticated inferences about one's partner, which can become computationally expensive with deeper recursions. While IWc also includes reasoning of "We," the "We" agent in this case serves as an imagined objective truth supposedly shared by all and thus does not reason any real agents' minds in return. This setup effectively avoids the trap of endless recursive

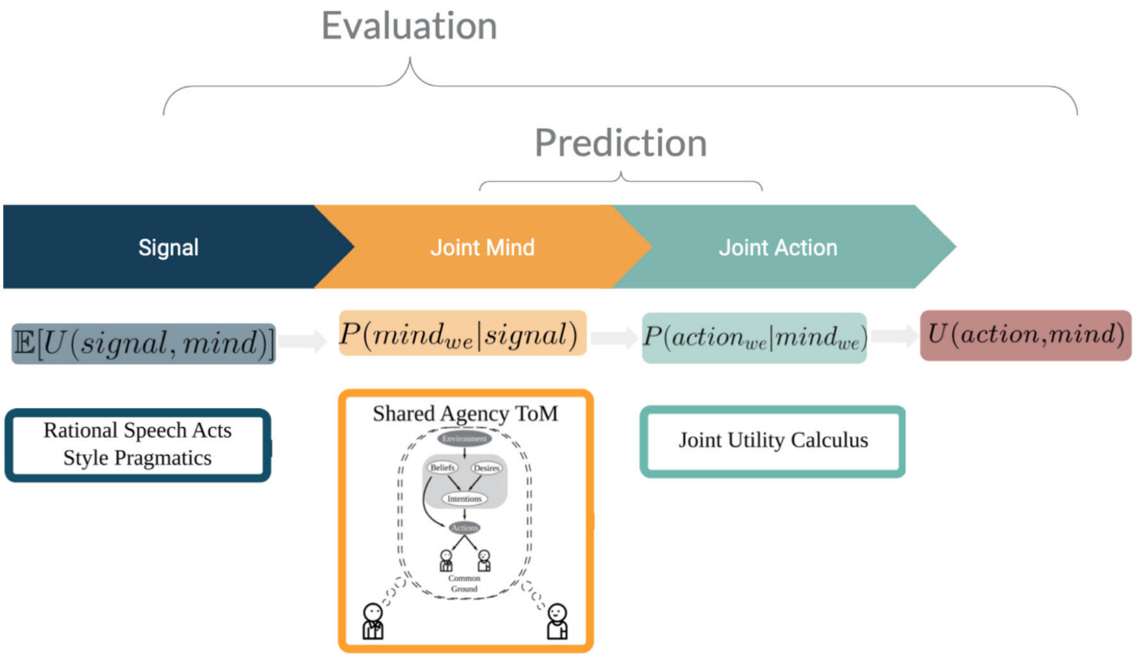

FIGURE 7 IW integrates RSA, ToM, and shared agency to connect signals to joint actions under a cooperative framework.

reasoning. In fact, the results show that when signalers and receivers of different reasoning levels play each other, IWc outperforms RSA at all levels of equivalent recursion. Moreover, even the simplest IWc model outperforms the most complex version of RSA tested, with an additional two layers of recursion. These findings indicate that integrating reasoning over multiple types of context from different sources can dramatically improve overloaded communication. In this case, joint planning over potential goals significantly reduces the need for deep recursion in communication. This aligns with the idea that communication, which occurs at a rate that is likely too fast (Levinson & Torreira, 2015) for deep recursion in RSA to keep up with, should be intuitive and computationally affordable.

7 | CONCLUSION

The goal of this review is to demonstrate the potential of shared agency models for cooperation and communication. We first reviewed that cooperation is achieved by extending individual ToM to a shared mind that can be imagined and inferred by individual agents. Then, we reviewed how ToM can also be modified and applied to communication where the production and interpretation of signals depended on an agent's reasoning of others' beliefs and evaluation of their action consequences. Lastly, we reviewed how these two lines of work can be combined to model cooperative communication, where shared agency serves as a natural constraint that helps overcome ambiguity in overloading signals. These works show how intelligent cooperation and communication can be modeled in tasks purely based on non-linguistic knowledge in a visual environment. This knowledge requires non-trivial cognitive capacities such as ToM and joint utility reasoning that are yet to be fully captured in more sophisticated communication involving language (Collins et al., 2022; Trott et al., 2022). Such evidence supports the developmental insights that human-unique cooperation and communication originate from vision and preexist before any mastery of language (Tomasello, 2010), and this visually grounded cognitive architecture can then be later augmented to the linguistic domain. Thus, the cognitive models reviewed here can potentially serve as the foundation for learning and using language.

This review focuses on the use of non-linguistic context as a key component of cooperation and communication. This lies in contrast with the growing trend of vision-language integration in artificial intelligence which often relies on a huge amount of linguistic input. Large Language Models (LLMs), which leverage an attention-based transformer network architecture, have been hugely successful in purely linguistic contexts (Brown et al., 2020; Devlin et al., 2018; Radford et al., 2018; Vaswani et al., 2017), are also able to provide a unified framework for multi-modal data and have been adapted for tasks involving reasoning over both text and images such as text-to-image generation (Ramesh et al., 2021; Rombach et al., 2022; Saharia et al., 2022) and Visual Question Answering (VQA; Chen et al., 2020; Tan & Bansal, 2019). While impressive in their ability to solve traditional computer vision tasks such as detection and recognition, these models still exhibit limitations toward reasoning and inference over how people think and talk about the world. For example, VQA models are biased by how questions are asked (Sharma & Jalal, 2021) and the reasoning behind their output is often opaque (Khan et al., 2022). Thus, it is difficult to interpret the errors they make or whether their reasoning incorporates any structural elements of either individual or shared agency. In contrast, a ToM-based approach explicitly represents the causal reasoning of agents based on beliefs, desires, and intentions and how they plan rationally. These two approaches, however, should be viewed as complementary: LLMs offer powerful linguistic knowledge to initialize the content of beliefs and desires that are essential to the ToM models in real-world situations; then, a shared agency model can organize that linguistic knowledge into a structured causal model that supports transparent, interpretable, and consistent reasoning, planning, and inference. Together, they can scale up the case studies reviewed above and potentially process real-world scenarios in which machines can cooperate and communicate in a way that humans find intuitive and trustworthy.

AUTHOR CONTRIBUTIONS

Stephanie Stacy: Conceptualization (lead); investigation (lead); visualization (lead); writing – original draft (lead). **Siyi Gong:** Conceptualization (lead); visualization (lead); writing – original draft (lead). **Aishni Parab:** Writing – original draft (supporting). **Minglu Zhao:** Writing – original draft (supporting). **Kaiwen Jiang:** Writing – original draft (supporting). **Tao Gao:** Conceptualization (lead); funding acquisition (lead); investigation (lead); methodology (lead); project administration (lead); supervision (lead); writing – original draft (supporting).

CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

Tao Gao https://orcid.org/0000-0003-2523-7867

REFERENCES

Anderson, C. J., & Dillon, B. W. (2019). Guess who's coming (and who's going): Bringing perspective to the rational speech acts framework. *Proceedings of the Society for Computation in Linguistics*, *2*(1), 185–194.

Austin, J. L. (1975). *How to do things with words*. Oxford University Press.

Baker, C. L., Jara-Ettinger, J., Saxe, R., & Tenenbaum, J. B. (2017). Rational quantitative attribution of beliefs, desires and percepts in human mentalizing. *Nature Human Behaviour*, *1*, 1–10.

Baker, C. L., Saxe, R., & Tenenbaum, J. B. (2009). Action understanding as inverse planning. *Cognition*, *113*, 329–349.

Baker, C. L., & Tenenbaum, J. B. (2014). Modeling human plan recognition using Bayesian theory of mind. In G. Sukthankar, C. Geib, H. Bui, D. Pynadath, & R. P. Goldman (Eds.), *Plan, activity, and intent recognition: Theory and practice (Chapter 7)* (pp. 177–204). Elsevier.

Barth, J., & Call, J. (2006). Tracking the displacement of objects: A series of tasks with great apes (pan troglodytes, pan paniscus, gorilla gorilla, and pongo pygmaeus) and young children (homo sapiens). *Journal of Experimental Psychology: Animal Behavior Processes*, *32*(3), 239–252.

Benz, A. (2004). Questions, plans, and the utility of answers. *Proceedings of Sinn und Bedeutung*, *8*, 51–66.

Botvinick, M., & Toussaint, M. (2012). Planning as inference. *Trends in Cognitive Sciences*, *16*(10), 485–488.

Bratman, M. (1987). *Intention, plans, and practical reason* (Vol. 10). Harvard University Press.

Bratman, M. E. (1992). Shared cooperative activity. *The Philosophical Review*, *101*(2), 327–341.

Bratman, M. E. (2013). *Shared agency: A planning theory of acting together*. Oxford University Press.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., … Amodei D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Bruner, J. (1985). Child's talk: Learning to use language. *Child Language Teaching and Therapy*, *1*(1), 111–114.

Buttelmann, D., Buttelmann, F., Carpenter, M., Call, J., & Tomasello, M. (2017). Great apes distinguish true from false beliefs in an interactive helping task. *PLoS One*, *12*(4), e0173793.

Buttelmann, D., Carpenter, M., Call, J., & Tomasello, M. (2007). Enculturated chimpanzees imitate rationally. *Developmental Science*, *10*(4), F31–F38.

Carpenter, M., Tomasello, M., & Striano, T. (2005). Role reversal imitation and language in typically developing infants and children with autism. *Infancy*, *8*(3), 253–278.

Chen, Y.-C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Yu, C., & Liu, J. (2020). Uniter: Universal image-text representation learning. In *Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Part XXX* (pp. 104–120). Springer.

Chomsky, N. (1983). Mental representations. *Syracuse Scholar (1979–1991)*, *4*(2), 2.

Clayton, N. S., Bussey, T. J., & Dickinson, A. (2003). Can animals recall the past and plan for the future? *Nature Reviews Neuroscience*, *4*(8), 685–691.

Cohen, P. R., & Perrault, C. R. (1979). Elements of a plan-based theory of speech acts. *Cognitive Science*, *3*(3), 177–212.

Collins, K. M., Wong, C., Feng, J., Wei, M., & Tenenbaum, J. B. (2022). *Structured, flexible, and robust: Benchmarking and improving large language models towards more human-like behavior in out-of-distribution reasoning tasks*. arXiv preprint arXiv:2205.05718.

De Villiers, J. G., & De Villiers, P. A. (2000). Linguistic determinism and the understanding of false. In P. Mitchell & K. Riggs (Eds.), *Children's reasoning and the mind* (191–228). Psychology Press.

Degen, J., Hawkins, R. D., Graf, C., Kreiss, E., & Goodman, N. D. (2020). When redundancy is useful: A Bayesian approach to “overinformative” referring expressions. *Psychological Review*, *127*(4), 591.

Dennett, D. C. (1987). *The intentional stance*. MIT Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *Bert: Pre-training of deep bidirectional transformers for language understanding*. arXiv preprint arXiv:1810.04805.

Duff, R. A. (1990). *Intention, agency and criminal liability: Philosophy of action and the criminal law*. Blackwell.

Duguid, S., Wyman, E., Bullinger, A. F., Herfurth-Majstorovic, K., & Tomasello, M. (2014). Coordination strategies of chimpanzees and human children in a stag hunt game. *Proceedings of the Royal Society B: Biological Sciences*, *281*(1796), 20141973.

Emery, N. J., & Clayton, N. S. (2004). The mentality of crows: Convergent evolution of intelligence in corvids and apes. *Science*, *306*(5703), 1903–1907.

Ferreira, V. S. (2008). Ambiguity, accessibility, and a division of labor for communicative success. *Psychology of Learning and Motivation*, *49*, 209–246.

Fletcher, G. E., Warneken, F., & Tomasello, M. (2012). Differences in cognitive processes underlying the collaborative activities of children and chimpanzees. *Cognitive Development*, *27*(2), 136–153.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998.

Franke, M. (2009). *Signal to act: Game theory in pragmatics*. University of Amsterdam.

Friston, K., Da Costa, L., Hafner, D., Hesp, C., & Parr, T. (2021). Sophisticated inference. *Neural Computation*, *33*(3), 713–763.

Gao, T., Baker, C. L., Tang, N., Haokui, X., & Tenenbaum, J. B. (2019). The cognitive architecture of perceived animacy: Intention, attention, and memory. *Cognitive Science*, *43*(8), e12775.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*(2), 165–193.

Gilbert, M. (1992). *On social facts*. Princeton University Press.

Gilbert, M. (2013). *Joint commitment: How we make the social world*. Oxford University Press.

Golinkoff, R. M. (1993). When is communication a ‘meeting of minds’? *Journal of Child Language*, *20*(1), 199–207.

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, *20*, 818–829.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories*. Mit Press.

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science*, *306*(5695), 496–499.

Gräfenhain, M., Behne, T., Carpenter, M., & Tomasello, M. (2009). Young children's understanding of joint commitments. *Developmental Psychology*, *45*(5), 1430–1443.

Greenberg, J. R., Hamann, K., Warneken, F., & Tomasello, M. (2010). Chimpanzee helping in collaborative and noncollaborative contexts. *Animal Behaviour*, *80*(5), 873–880.

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. Morgan (Eds.), *Syntax and semantics, volume 3 of Speech acts* (pp. 41–58). Academic Press.

Grosse, G., Moll, H., & Tomasello, M. (2010). 21-Month-olds understand the cooperative logic of requests. *Journal of Pragmatics*, *42*(12), 3377–3383.

Grosz, B. J., & Kraus, S. (1996). Collaborative plans for complex group action. *Artificial Intelligence*, *86*(2), 269–357.

Hamann, K., Warneken, F., Greenberg, J. R., & Tomasello, M. (2011). Collaboration encourages equal sharing in children but not in chimpanzees. *Nature*, *476*(7360), 328–331.

Hamann, K., Warneken, F., & Tomasello, M. (2012). Children's developing commitments to joint goals. *Child Development*, *83*(1), 137–145.

Hare, B., Call, J., Agnetta, B., & Tomasello, M. (2000). Chimpanzees know what conspecifics do and do not see. *Animal Behaviour*, *59*(4), 771–785.

Hare, B., Call, J., & Tomasello, M. (2006). Chimpanzees deceive a human competitor by hiding. *Cognition*, *101*(3), 495–514.

Hare, B., & Tomasello, M. (2004). Chimpanzees are more skilful in competitive than in cooperative cognitive tasks. *Animal Behaviour*, *68*(3), 571–581.

Hawkins, R. D., Gweon, H., & Goodman, N. D. (2021). The division of labor in communication: Speakers help listeners account for asymmetries in visual perspective. *Cognitive Science*, *45*(3), e12926.

Hawkins, R. X. D., Frank, M., & Goodman, N. D. (2017). Convention-formation in iterated reference games. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Herrmann, E., Call, J., Hernández-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans have evolved specialized skills of social cognition: The cultural intelligence hypothesis. *Science*, *317*(5843), 1360–1366.

Ho, M. K., Cushman, F., Littman, M. L., & Austerweil, J. L. (2021). Communication in action: Planning and interpreting communicative demonstrations. *Journal of Experimental Psychology: General*, *150*, 2246–2272.

Ho, M. K., Littman, M., MacGlashan, J., Cushman, F., & Austerweil, J. L. (2016). Showing versus doing: Teaching by demonstration. *Advances in Neural Information Processing Systems*, *29*, 3027–3035.

Ho, M. K., Saxe, R., & Cushman, F. (2022). Planning with theory of mind. *Trends in Cognitive Sciences*, *26*, 959–971.

Hunt, G. R. (1996). Manufacture and use of hook-tools by new caledonian crows. *Nature*, *379*(6562), 249–251.

Jacobsson, F., Johannesson, M., & Borgquist, L. (2007). Is altruism paternalistic? *The Economic Journal*, *117*(520), 761–781.

Jara-Ettinger, J., Gweon, H., Schulz, L. E., & Tenenbaum, J. B. (2016). The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in Cognitive Sciences*, *20*, 589–604.

Jara-Ettinger, J., Gweon, H., Tenenbaum, J. B., & Schulz, L. E. (2015). Children's understanding of the costs and rewards underlying rational action. *Cognition*, *140*, 14–23.

Jiang, K., Stacy, S., Wei, C., Chan, A., Rossano, F., Zhu, Y., & Gao, T. (2021). Individual vs. joint perception: A pragmatic model of pointing as communicative Smithian helping. In *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Kaelbling, L. P., Littman, M. L., & Cassandra, A. R. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, *101*(1–2), 99–134.

Kao, J., Bergen, L., & Goodman, N. (2014). Formalizing the pragmatics of metaphor understanding. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Khan, S., Naseer, M., Hayat, M., Zamir, S. W., Khan, F. S., & Shah, M. (2022). Transformers in vision: A survey. *ACM Computing Surveys (CSUR)*, *54*(10s), 1–41.

Kleiman-Weiner, M., Ho, M. K., Austerweil, J. L., Littman, M. L., & Tenenbaum, J. B. (2016). Coordinate to cooperate or compete: Abstract goals and joint intentions in social interaction. In D. Mirman, A. Papafragou, D. Grodner, & J. C. Trueswell (Eds.), *Proceedings of the 38th Annual Conference of the Cognitive Science Society* (pp. 1679–1684). Cognitive Science Society.

Krupenye, C., Kano, F., Hirata, S., Call, J., & Tomasello, M. (2016). Great apes anticipate that other individuals will act according to false beliefs. *Science*, *354*(6308), 110–114.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253.

Levesque, H. J., Cohen, P. R., & Nunes, J. H. T. (1990). *On acting together*. SRI International.

Levinson, S. C., Kita, S., Haun, D. B. M., & Rasch, B. H. (2002). Returning the tables: Language affects spatial reasoning. *Cognition*, *84*(2), 155–188.

Levinson, S. C., & Torreira, F. (2015). Timing in turn-taking and its implications for processing models of language. *Frontiers in Psychology*, *6*, 731.

Lewis, D. (1969). *Convention: A philosophical study*. Harvard University Press.

Liszkowski, U., Schäfer, M., Carpenter, M., & Tomasello, M. (2009). Prelinguistic infants, but not chimpanzees, communicate about absent entities. *Psychological Science*, *20*(5), 654–660.

Liu, S., Ullman, T. D., Tenenbaum, J. B., & Spelke, E. S. (2017). Ten-month-old infants infer the value of goals from the costs of actions. *Science*, *358*(6366), 1038–1041.

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, *30*, 6379–6390.

MacKay, D. J. C. (2003). *Information theory, inference and learning algorithms*. Cambridge University Press.

Melis, A. P., Call, J., & Tomasello, M. (2006). Chimpanzees (pan troglodytes) conceal visual and auditory information from others. *Journal of Comparative Psychology*, *120*(2), 154–162.

Melis, A. P., Schneider, A.-C., & Tomasello, M. (2011). Chimpanzees, pan troglodytes, share food in the same way after collaborative and individual food acquisition. *Animal Behaviour*, *82*(3), 485–493.

Misyak, J., Noguchi, T., & Chater, N. (2016). Instantaneous conventions: The emergence of flexible communicative signals. *Psychological Science*, *27*(12), 1550–1561.

Moll, H., & Tomasello, M. (2006). Level 1 perspective-taking at 24 months of age. *British Journal of Developmental Psychology*, *24*(3), 603–613.

Newell, A., Shaw, J. C., & Simon, H. A. (1957). Empirical explorations of the logic theory machine: A case study in heuristic. In *The Western Joint Computer Conference: Techniques for Reliability* (pp. 218–230). Association for Computing Machinery.

Olson, M. (1989). Collective action. In *The invisible hand* (pp. 61–69). Springer.

Parikh, P. (1991). Communication and strategic inference. *Linguistics and Philosophy*, *14*(5), 473–514.

Penn, D. C., & Povinelli, D. J. (2007). On the lack of evidence that nonhuman animals possess anything remotely resembling a 'theory of mind'. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1480), 731–744.

Piantadosi, S. T., Tily, H., & Gibson, E. (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291.

Pinker, S. (2003). *The language instinct: How the mind creates language*. Penguin UK.

Premack, D., & Woodruff, G. (1978). Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, *1*(4), 515–526.

Qing, C., & Franke, M. (2015). Variations on a Bayesian theme: Comparing Bayesian models of referential reasoning. In H. Zeevat & Schmitz, HC. (Eds.), *Bayesian Natural Language Semantics and Pragmatics*. Language, Cognition, and Mind (Vol. 2, pp. 201–220). Springer, Cham.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. Technical Report. OpenAI.

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., & Sutskever, I. (2021). Zero-shot text-to-image generation. In *International Conference on Machine Learning* (pp. 8821–8831). PMLR.

Richardson, D. C., Dale, R., & Kirkham, N. Z. (2007). The art of conversation is coordination. *Psychological Science*, *18*(5), 407–413.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10684–10695). Computer Vision Foundation.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Russell, S. J., Norvig, P., & Davis, E. (2010). *Artificial intelligence: A modern approach*. Prentice Hall.

Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E. L., Ghasemipour, K., Lopes, R. G., Ayan, B. K., Salimans, T., Ho, J., Fleet, D. J., & Norouzi, M. (2022). Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, *35*, 36479–36494.

Schelling, T. C. (1980). *The strategy of conflict: With a new preface by the author*. Harvard University Press.

Schweikard, D. P., & Schmid, H. B. (2021). Collective intentionality. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall, 2021 ed.). Metaphysics Research Lab, Stanford University.

Searle, J. R. (1990). Collective intentions and actions. *Intentions in Communication*, *195*, 220.

Searle, J. R., & Searle, J. R. (1969). *Speech acts: An essay in the philosophy of language* (Vol. 626). Cambridge University Press.

Sharma, H., & Jalal, A. S. (2021). A survey of methods, datasets and evaluation metrics for visual question answering. *Image and Vision Computing*, *116*, 104327.

Shockley, K., Richardson, D. C., & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science*, *1*(2), 305–319.

Shum, M., Kleiman-Weiner, M., Littman, M. L., & Tenenbaum, J. B. (2019). Theory of minds: Understanding behavior in groups through inverse planning. In *Proceedings of the AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence* (Vol. 33, pp. 6163–6170).

Sibicky, M. E., Schroeder, D. A., & Dovidio, J. F. (1995). Empathy and helping: Considering the consequences of intervention. *Basic and Applied Social Psychology*, *16*(4), 435–453.

Simon, H. A., & Newell, A. (1962). *Computer simulation of human thinking and problem solving* (pp. 137–150). Monographs of the Society for Research in Child Development.

Siposova, B., Tomasello, M., & Carpenter, M. (2018). Communicative eye contact signals a commitment to cooperate for young children. *Cognition*, *179*, 192–201.

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96.

Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Harvard University Press.

Stacy, S., Li, C., Zhao, M., Yun, Y., Zhao, Q., Kleiman-Weiner, M., & Gao, T. (2021). Modeling communication to coordinate perspectives in cooperation. In *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*. Cognitive Science Society.

Stacy, S., Parab, A., Kleiman-Weiner, M., & Gao, T. (2022). Overloaded communication as paternalistic helping. In *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society.

Sumers, T., Hawkins, R., Ho, M. K., & Griffiths, T. (2021). Extending rational models of communication from beliefs to actions. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. Cognitive Science Society.

Tan, H., & Bansal, M. (2019). Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 5100–5111). Association for Computational Linguistics.

Tang, N., Gong, S., Zhao, M., Chenya, G., Zhou, J., Shen, M., & Gao, T. (2022). Exploring an imagined "we" in human collective hunting: Joint commitment within shared intentionality. *Proceedings of the 44th Annual Meeting of the Cognitive Science Society*. Cognitive Science Society.

Tang, N., Stacy, S., Zhao, M. L., Marquez, G., & Gao, T. (2020). Bootstrapping an imagined we for cooperation. In Y. Xu, S. Denison, M. Mack, & B. C. Armstrong (Eds.), *Proceedings of the 42nd Annual Conference of the Cognitive Science Society* (pp. 2453–2458). Cognitive Science Society.

Tomasello, M. (2009). *Why we cooperate*. MIT Press.

Tomasello, M. (2010). *Origins of human communication*. MIT Press.

Tomasello, M. (2019). Becoming human. In *Becoming human*. Harvard University Press.

Tomasello, M., Call, J., & Gluckman, A. (1997). Comprehension of novel communicative signs by apes and human children. *Child Development*, *68*, 1067–1080.

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(5), 675–691.

Tomasello, M., Carpenter, M., & Hobson, R. P. (2005). The emergence of social cognition in three young chimpanzees. *Monographs of the Society for Research in Child Development*, *70*(1), 1–152.

Tomasello, M., & Haberl, K. (2003). Understanding attention: 12-and 18-month-olds know what is new for other persons. *Developmental Psychology*, *39*(5), 906–912.

Trott, S., Jones, C., Chang, T., Michaelov, J., & Bergen, B. (2022). *Do large language models know what humans know?* arXiv preprint arXiv: 2209.01515.

Tuomela, R. (2007). *The philosophy of sociality: The shared point of view*. Oxford University Press.

Turing, A. M. (2009 [1950]). Computing machinery and intelligence. In *Parsing the Turing test* (pp. 23–65). Springer.

Vaish, A., Carpenter, M., & Tomasello, M. (2016). The early emergence of guilt-motivated prosocial behavior. *Child Development*, *87*(6), 1772–1782.

Van Rooy, R. (2003). Questioning to resolve decision problems. *Linguistics and Philosophy*, *26*, 727–763.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*, 5998–6008.

Vygotsky, L. S., & Cole, M. (1978). *Mind in society: Development of higher psychological processes*. Harvard University Press.

Warneken, F., Chen, F., & Tomasello, M. (2006). Cooperative activities in young children and chimpanzees. *Child Development*, *77*(3), 640–663.

Warneken, F., Hare, B., Melis, A. P., Hanus, D., & Tomasello, M. (2007). Spontaneous altruism by chimpanzees and young children. *PLoS Biology*, *5*(7), e184.

Warneken, F., Lohse, K., Melis, A. P., & Tomasello, M. (2011). Young children share the spoils after collaboration. *Psychological Science*, *22*(2), 267–273.

Warneken, F., Steinwender, J., Hamann, K., & Tomasello, M. (2014). Young children's planning in a collaborative problem-solving task. *Cognitive Development*, *31*, 48–58.

Warneken, F., & Tomasello, M. (2006). Altruistic helping in human infants and young chimpanzees. *Science*, *311*(5765), 1301–1303.

Wellman, H. M. (1992). *The child's theory of mind*. The MIT Press.

Whorf, B. L. (2012). *Language, thought, and reality: Selected writings*. MIT Press.

Winkler, S. (2015). *Ambiguity: Language and communication*. Walter de Gruyter GmbH & Co KG.

Wittgenstein, L. (1953). *Philosophical investigations*. Blackwell.

Wobber, V., Herrmann, E., Hare, B., Wrangham, R., & Tomasello, M. (2014). Differences in the early cognitive development of children and great apes. *Developmental Psychobiology*, *56*(3), 547–573.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1–34.

Wu, S. A., Wang, R. E., Evans, J. A., Tenenbaum, J. B., Parkes, D. C., & Kleiman-Weiner, M. (2021). Too many cooks: Bayesian inference for coordinating multi-agent collaboration. *Topics in Cognitive Science*, *13*(2), 414–432.

Yamamoto, S., Humle, T., & Tanaka, M. (2012). Chimpanzees' flexible targeted helping based on an understanding of conspecifics' goals. *Proceedings of the National Academy of Sciences*, *109*(9), 3588–3592.

Zhao, M., Tang, N., Dahmani, A. L., Perry, R. R., Zhu, Y., Rossano, F., & Gao, T. (2021). Sharing is not needed: Modeling animal coordinated hunting with reinforcement learning. In *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. Cognitive Science Society.

Zhu, Y., Gao, T., Fan, L., Huang, S., Edmonds, M., Liu, H., Gao, F., Zhang, C., Qi, S., Wu, Y. N., Tenenbaum, J. B., & Zhu, S. C. (2020). Dark, beyond deep: A paradigm shift to cognitive ai with humanlike common sense. *Engineering*, *6*(3), 310–345.

AUTHOR BIOGRAPHY

Tao Gao investigates the visual roots of human social perception and cognition. He designs models of artificial social intelligence endowed with human-like visual commonsense. By merely sharing the same visual environment, these models can Interact and communicate with humans in ways that are intuitive, effective, and trustworthy. He received his PhD in cognitive psychology from Yale in 2011 and subsequently held a post-doctoral fellowship at MIT's Center for Brain, Mind, and Machine from 2011 to 2015. Later, he served as a computer vision scientist at GE Research from 2015 to 2017. Since 2017, Dr. Gao has been jointly appointed to the Departments of Communication, Statistics, and Psychology at UCLA.

How to cite this article: Stacy, S., Gong, S., Parab, A., Zhao, M., Jiang, K., & Gao, T. (2024). A Bayesian theory of mind approach to modeling cooperation and communication. *WIREs Computational Statistics*, *16*(1), e1631. https://doi.org/10.1002/wics.1631