

This is an Accepted Manuscript for *Behavioral and Brain Sciences*. This Target Article has been accepted for publication but not yet copyedited or typeset, and so may be subject to change during the production process. The article is considered published and may be cited using its DOI 10.1017/S0140525X25103427

The Institutional Stance

Julian Jara-Ettinger¹ and Yarrow Dunham²

¹Yale University, USA, julian.jara-ettinger@yale.edu, <https://compdevlab.yale.edu/>

²Yale University, USA, yarrow.dunham@yale.edu, <https://www.socialcogdev.com/>

Short abstract

The power of human social cognition is typically associated with mental state reasoning, a *mentalistic stance*. Here we argue for the presence of a second equally powerful system, an *institutional stance*, that interprets social interactions in terms of structured sets of roles linked to rule-like behavioral expectations. Precursors of the institutional stance in other species may support social interaction in the absence of complex models of mental states. However, the institutional stance is unique in humans in its generative capacity, supporting the creation of a myriad range of institutional structures that shape our world and enable large scale social coordination.

Long abstract

Human success in navigating the social world is typically attributed to our capacity to represent other minds—a *mentalistic stance*. We argue that humans are endowed with a second equally powerful intuitive theory: an *institutional stance*. In contrast to the mentalistic stance, which helps us predict and explain unconstrained behavior via unobservable mental states, the institutional stance interprets social interactions in terms of role-based structures that constrain and regulate behavior via rule-like behavioral expectations. We argue that this stance is supported by a generative grammar that builds structured models of social collectives, enabling people to rapidly infer, track, and manipulate the social world. The institutional stance emerges early in development and its precursors can be traced across social species, but its full-fledged generative capacity is uniquely human. Once in place, the ability to reason about institutional structures takes on a causal role, allowing people to create and modify social structures,

supporting new forms of institutional life. Human social cognition is best understood as an interplay between a system for representing the unconstrained behavior of individuals in terms of minds and a system for representing the constrained behavior of social collectives in terms of institutional structures composed of interlocking sets of roles.

Keywords: Social cognition, Coordination, Institutional Stance, Theory of Mind, roles, scripts

1. Introduction

Humans are unparalleled in the ubiquity and complexity of their social lives. From passing interactions to life-long relationships, we not only effortlessly navigate the social world, but also shape it by building social relations, communities, and societies. How do humans reach levels of social complexity that no other species has achieved? Theories of human social intelligence in the cognitive sciences have been dominated by the idea that human social interactions are supported by a capacity to *mentalize*: We interpret other people's behavior by attributing mental states like preferences, knowledge, and desires. This allows us to explain and predict behavior, determine how to interact with others, and even plan for how to change minds. The complexity of this *mentalistic stance* is distinctively human (Martin & Santos, 2016): it emerges early in infancy (Sodian, 2011; Onishi & Baillargeon, 2005); is supported by specialized neural circuitry (Saxe & Kanwisher, 2003; Saxe, 2006); and permeates social thought, from language understanding to moral reasoning (Rubio-Fernandez et al., 2025; Sperber & Wilson, 1986; Young et al., 2007).

Here we argue that such a view is an incomplete characterization of human social thought. We propose that human social intelligence is best understood as the co-evolution of two systems of social understanding, the *mentalistic stance* just discussed, and an *institutional stance*. This institutional stance is on equal footing with the mentalistic stance in terms of representational complexity, evolutionary history, and everyday use, but it differs in that its core representational substrate does not involve representations of mental states. Instead, the institutional stance is an interpretative framework generated by representations of social structures that specify how agents interact with one another as a function of relative roles with rule-like normative entailments.

Consider two simple examples that illustrate the institutional stance at work. Imagine that you are at an office building when you see Kate step out of her office, wave at Eric at a desk outside, and rattle off “Can you feed my parking meter, pick up my lunch, and push my next appointment by 30 minutes?” which Eric jots down and rushes out to complete. In principle, we could interpret this event through a mentalistic stance: Kate *wants* lunch delivered, and she *believes* that she can fulfill her desires by uttering them to Eric. Conversely, Eric has a recursive *desire* to fulfill Kate’s desires, and *believes* that he can do so by completing the tasks that Kate described. Suppose now that, over the weekend, Kate goes to the movies and spots Eric sitting a few rows away, apparently on a date. Kate waves him over, asks him to get her some popcorn and a fountain drink, and turns back to continue watching the movie. Under the mentalistic analysis we just applied, this interaction should be wholly unsurprising. It provides nothing more than confirmatory evidence of the mental states we already attributed to Kate. And we should further expect Eric to excuse himself and happily go fulfill Kate’s desires. Why then does Kate’s behavior strike us as surprising and improper, and why do we expect Eric to be similarly taken aback? We propose that this surprise emerged because we interpreted Eric and Kate’s first interaction not through a mentalistic stance but through an institutional stance, where each person had different social affordances based on the roles they occupied at that moment (e.g., Eric is Kate’s assistant)—roles they do not presently occupy and that therefore no longer constrain or predict their behavior in the ways they did before.

Consider a second example. Imagine that Laurie lands in a new city and goes to a rental car agency where Paul greets her and helps her fill out the required paperwork. As Laurie is filling out the forms, she notices that the final fee is higher than she expected and starts arguing with Paul. After a minute, Paul briefly excuses himself, and a few minutes later, Alex shows up and picks up the discussion with Laurie right where Paul left off. Under the mentalistic stance, the transfer of Paul’s beliefs, desires, and intentions to Alex is strikingly odd, and we should reasonably expect Laurie to be surprised about the seamless swap of communicative partners. But the uninterrupted flow of the interaction is unsurprising under an institutional stance, where we recognize that a new person occupying the same role will take on the goals and behaviors attributed to that role (i.e., placating a customer). These inferences uniquely follow from an institutional stance. To see why, consider how different our expectations would be in a situation

where the institutional stance is *not* warranted. For example, imagine Laurie's surprise if, during a personal argument with a close friend, the friend excused themselves only to have a different friend step in and pick up the argument where it had been left off. In this case a seemingly similar transfer of goals would be surprising, if not off-putting.

These examples illustrate how analyzing social interactions by appealing to role-based representations allow us to understand and predict behavior, and they point to commonplace social situations where mentalistic representations alone struggle to provide accurate explanations. While it may be possible to provide a mentalistic interpretation of some of these cases, doing so requires assuming increasingly complex, auxiliary, context-sensitive, and intuitively implausible mental-state attributions. By contrast, we propose that these same social interactions are seamlessly understood via an *institutional stance*, an intuitive theory for analyzing behavior in terms of underlying role-based social structures.

A growing body of work across disciplines has converged onto the idea that role-based systems are central to understanding complex social behavior. This view spans philosophy (e.g. Maibom, 2007; Ritchie, 2020), computer science (Schank & Abelson, 1977), developmental psychology (Clément et al, 2011), and anthropology (e.g. Fiske, 1991), and it can be organized into three interrelated projects. The first is to characterize the forces that give rise to coordination and cooperation, from dyadic relations to large institutions that are built and refined over generations (e.g. Boyer & Petersen, 2012; Teraji, 2018). The second is to develop a social ontology characterizing what social structures are (i.e, how should we, as theorists, formalize and study institutions; Crawford & Ostrom, 1995; Guala, 2016; Searle, 2006; Hodgson, 2006). The third is to understand how social creatures, without formal training, intuitively represent, reason about, manipulate, and immerse themselves into social collectives (e.g., Eickers, 2024; Maibom, 2007; Noyes et al., 2020a, 2021; Ritchie, 2020). Our paper aligns with this third project, and our goal is to propose a specific theory of how humans do this, with a focus on the structure of these cognitive representations, their evolutionary origins, and everyday use. Thus, while there are deep connections across these projects, our goal is not to provide a comprehensive review of the many fields that have emphasized the importance of roles, but to develop a particular perspective derived from representational theories of human cognition.

Why is the institutional stance a *stance*? In our proposal, it is an interpretive framework for making sense of social behavior. Like the mentalistic stance, it can be adopted or applied as part of everyday sense-making. It is useful to the extent that it stands in principled enough relations to the real world that it allows prediction and explanation. Consider the representations driving the mentalistic stance (e.g., beliefs and desires). Clearly they are highly simplified as compared to the actual psychological causes of individual behavior, but they are useful because they capture important regularities. Similarly, the representations of social structures driving the institutional stance will often be considerably simpler than, and in some cases perhaps even divergent from, the actual social structures populating the world. Adopting the institutional stance is useful so long as its underlying representations succeed at capturing critical regularities. It is thus important to clearly distinguish between ground truth social structures actually present in the world and the mental representations of social structures populating individual minds. The institutional stance is adopted when those representations are used to place individuals into roles and reason about their actions in light of beliefs about those roles.

There are four central aspects to our proposal. First, people routinely interpret and explain behavior as caused by roles in latent institutional structures; we can apply this stance to nearly any social interaction, even when we do not necessarily believe that people are consciously or intentionally enacting roles; and the core representations driving this capacity cannot be adequately characterized as or reduced to *mentalizing*. Second, these representations are not the byproduct of general intelligence, but instead reflect a specialized and fully coherent intuitive theory of social behavior. Third, representations within the institutional stance go beyond dyadic role representations. Roles are interpreted in the context of a latent institutional representation that expresses how they fit into a broader system of social coordination (i.e., an institution). Fourth, the institutional stance is of comparable complexity to the mentalistic stance: It is evolutionarily ancient, emerges early in development, enjoys the support of specialized computations (and, perhaps, specialized neural circuitry), and it is uniquely human in important ways. The *mentalistic stance* and the *institutional stance* can be thought of as providing complementary solutions to some of the same problems posed by social life. The mentalistic stance provides a framework for understanding individual agents across a wide range of contexts, including novel ones, based on a causal theory of how mental states relate to behavior. The

institutional stance provides a framework for leveraging recurrent patterns of multi-party social interaction to understand how agents will engage with the world and one another in familiar contexts, including when mentalizing is costly or fallible. While they are independent systems, the two stances can also interact in ways that support increasingly complex forms of social reasoning (Figure 1).

To understand the institutional stance, it is helpful to see how precursors can be found across multiple social species. Mentalizing evolved to allow us to understand and predict each other's behavior across a wide range of contexts, in particular novel ones. However, most social species, both now and across evolutionary history, lack the ability to build rich interpretative models of other minds. This led to the emergence of systems that make behavior intelligible and predictable to solve coordination problems. Once in place, these structures are available to become explicit objects of cognition, driving the evolution of an interpretative stance that allows us to build mental models of social structures.

This representational system achieves its full power in humans. Non-human social species can represent a fixed and predetermined set of institutional structures (such as kinship relations, rank hierarchies, and troop patrols). By contrast, the human institutional stance is generative, akin to a grammar (Crawford & Ostrom, 1995): We can represent and create an unbounded number of arbitrary institutional structures, from soccer teams, knitting groups and book clubs, to home-owner associations, universities and governments. In other words, the institutional stance in humans is a full-fledged and generative system.

In the remainder of the paper, we present the case for mentalizing as the foundation of human social intelligence (section 2) and argue that it fails to explain socially-complex behavior across species. We propose that many forms of social coordination are enabled through role-based systems that form precursors to the institutional stance (section 3). We then formalize the institutional stance (section 4), and connect it to relevant frameworks (section 5). Finally, we discuss evidence for and implications of this framework (sections 6-7).

2. Mentalizing as a bedrock of social intelligence

Before presenting the institutional stance, it is helpful to briefly review the case for a mentalistic stance as a standalone cognitive system responsible for social intelligence. Taking a mentalistic stance is the activity of analyzing, representing, and interacting with social events via *Theory of Mind*—an intuitive theory positing unobservable mental states that cause behavior (Dennett, 1989; Gopnik & Meltzoff, 1997; Wellman, 2014). Evidence that the mentalistic stance is central to social cognition comes from four lines of research that broadly tell us: How the mentalistic stance emerges and develops; what mentalizing enables us to do that we couldn't do otherwise; how this capacity is instantiated in the brain; and how ubiquitous mentalizing is in human social life. We briefly review each of these lines below.

The first piece of evidence that mentalizing is central to social understanding comes from research showing that this system is functional from very early in life. Infants interpret other people's behavior in terms of unobservable mental states (Onishi & Baillargeon, 2005; Repacholi & Gopnik, 1997; Liszkowski et al., 2008) that rationally cause behavior (Gergely & Csibra, 2003; Liu et al., 2017; Woodward, 1998). This understanding structures early childhood cognition, including language understanding (Akhtar et al., 1996; Vaish et al., 2011; Rubio-Fernandez et al., 2025), causal learning (Gergely et al., 2002; Bonawitz et al., 2011), socio-moral evaluations (Hamlin et al., 2013; Sloane et al., 2012), and epistemic vigilance (Harris, 2012; Aboody et al., 2022).

In adults, many complex social behaviors appear to rely on mentalizing. Computational models of mentalizing capture with quantitative accuracy how people infer preferences (Baker et al., 2017; Jern et al., 2017), predict behavior (Jara-Ettinger et al., 2020a), cooperate (Wang et al., 2020), make social evaluations (Ullman et al., 2009), share knowledge (Shafto et al., 2014; Ho et al., 2016), and communicate (Goodman & Frank, 2016; Frank & Goodman, 2012).

Mentalizing is further supported by neural circuitry specialized for social computations (Saxe & Kanwisher, 2003; Saxe, 2006). These networks are directly linked to mental-state representations (Jamali et al., 2021); they are associated with social activities thought to require mentalizing, such as communication (Mi et al., 2021; Paunov et al., 2019) and moral reasoning

(Young et al., 2010, 2007); and they correlate with computational models of human mentalizing (Collette et al., 2017).

Finally, some aspects of mentalizing may be ubiquitous and automatic. The basic building blocks—detecting agents, reading their intentions, and inferring their goals—are hardwired into human perception (Heider & Simmel, 1944; Gao et al., 2010; McMahon & Isik, 2023), automatically triggering richer inferences about other people's mental states (Rubio-Fernández et al., 2019; Kamps & Southgate, 2020).

3. Evolutionary origins of the institutional stance

While mentalizing is critical to human social intelligence, this capacity cannot be necessary for social behavior because complex social life is far more widespread across species than full-fledged mentalizing. Even in species with some form of Theory of Mind, such as non-human primates (Krupenye & Call, 2019; Royka & Santos, 2022), their mentalistic representations can be difficult to elicit, they are often brittle, and they lack the flexibility of human mentalizing (Martin & Santos, 2016; Horschler et al., 2023). How do social creatures that cannot build high-resolution predictive models of each other's minds coordinate and cooperate with one another?

The origins of the institutional stance reflect a pressure for living in social groups, combined with a limited capacity to understand and predict each other's behavior. Simply put, social species did not have the luxury of living in isolation until a sufficiently advanced form of mentalizing evolved (and mentalizing itself cannot emerge in full complexity in the absence of social life and its attendant pressures; Dunbar, 2009). We propose that, to solve this problem, species first evolve simple systems of social organization that make behavior structured and predictable, but do not require representing minds or roles. Once these forms of collective behavior are established, being able to manipulate and control the system becomes advantageous. A well-known principle in engineering is that one of the most powerful ways to control a system is by building a model of it (Conant & Ashby, 1970; Francis & Wonham, 1976; Wilterson & Graziano, 2021), and this led to the evolution of the capacity to represent systems of social coordination via role-based structures (see also Maibom, 2007 for a similar perspective on

building social models). We specifically propose that this can be understood in terms of three levels, which we describe below.

3.1 Level 1: Proto-institutional systems

Much research into the evolution of social behavior suggests that social coordination problems can be readily solved through what we will call *proto-institutional systems*: emergent systems of organization, such as kinship relations and biologically-determined divisions of labor, in which different individuals carry out different tasks. Importantly, while this might look role-like from the outside, it can occur without anyone in the system representing it in terms of roles (Figure 2a). In these systems, each member knows a set of rules and regularities that determines how to behave within a group and how to interact with others. These simple social structures can solve social interaction problems that would otherwise require mentalizing. For instance, a simple linear dominance hierarchy might encode a social affordance as basic as: in the presence of a conspecific, the higher-ranking member can do as they please while ignoring subordinates; subordinates must get out of the way and not interfere with the dominant conspecific. Such social structures enable creatures to co-exist in a group without access to each other's minds as long as they know how to behave within the social system. These representations are consistent with the notion of “minimal cooperation” (Maibom, 2007; Ritchie, 2020b), in which individuals playing roles in organized structures can, by simply doing their part, achieve relatively complex coordination as long as others do the same.

These basic forms of social organization appear to be recurrent solutions to coordination problems (Sumpter, 2010). They may have emerged many times independently across multiple social species, such as kinship relations, where conspecifics develop expectations about who will react and care for an infant in need (Dasser, 1988; Cheney & Seyfarth, 1980); primitive forms of in-group / out-group representations (Lorenz, 2021); or forms of coordinated hunting (such as with Aplomado falcons, where males and females perform different but fixed tasks; Hector, 1986). One of the best-known examples is the division of labor between workers and soldiers in ants (Hölldobler & Wilson, 1990), who act within biologically fixed social structures with clear roles that provide different social affordances. This is particularly apparent in ants in the case of *dulosis*—the process where parasitic ant species capture worker ants from a different species,

which then continue to carry out their duties, now in the service of the parasitic colonizers (D'Ettorre & Heinze, 2001).

We call these systems *proto-institutional* because members of the social group do not, or at least need not, have any representation of roles or of the social structure itself. That is, the role and role-occupier are not differentiated: associating different agents with different social affordances is enough (known as a lack of role-filler independence; Figure 2a). Despite this, knowing what behaviors are associated with each agent still supports some forms of inference. For instance, the types of transitive inferences in a dominance hierarchy (e.g., those seen among pinyon jays; Paz-y-Miño et al., 2004) can be instantiated by simply representing each conspecific as having a particular amount of power. The cognitive prerequisites of a *proto-institutional system* are also not trivial; an agent must know what social affordances each individual has, including themselves. This can be achieved in a variety of ways, including individual identification (e.g., recognizing your offspring) and consistent visual markers (e.g., in European paper wasps; Tibbetts & Lindsay, 2008). These behaviors might also require some basic form of action understanding (at a minimum, a capacity to process body movements and make simple predictions such as direction of motion, but they could also be grounded in richer representations such as a *teleological stance*; Gergely & Csibra, 2003). This suggests diversity in how proto-institutional systems work, but the critical point is that these systems do not yet require a flexible or complex representational understanding of roles or the social structures they are embedded within.

3.2 Level 2: Fixed institutional representations with role-filler independence

As social organization becomes more complex, groups benefit from the ability of conspecifics to take on different jobs depending on the context and the current group needs. This is the origin of role-based systems that allow members of a group to swap roles (Figure 2b)¹. The increased flexibility of such systems also imposes new representational and inferential demands via role-filler independence: Members must have an abstract representation of roles that is independent of the members occupying them; they must be able to infer which role a conspecific

¹ Here, we refer exclusively to role swapping happening *representationally* in conspecifics' minds. This is different from *physical* role swaps as in, e.g., physically changing one's sex in the case of sequential hermaphroditism (Gemmell, et al., 2019).

is occupying based on their appearance or behavior; they must track and remember role assignments and role changes throughout time; and they must know how to adjust their behavior as a function of the role that they themselves, as well as their interaction partners, occupy.

A candidate example is male dominance hierarchies in monkeys, which are dynamic and continuously renegotiated. Indeed, some primates like Baboons are thought to have specialized intelligence for tracking and remembering rapid and complex role changes (Cheney & Seyfarth, 2019; Bergman et al., 2003). The variety of social behaviors non-human primates engage in also increases the demands for context-specific systems of organization, and it is possible that they have additional institutional representations such as troop patrols or even multi-role hunting structures (Boesch, 1994, 2005; although such interpretation is disputed; see, e.g., Moll & Tomasello, 2007).

These types of abstract institutional representations may be present beyond primates, suggesting they evolved independently multiple times, or emerged a long time ago in our evolutionary history. In desert lions, for instance, female lions hunt through a coordinated structure that consists of two center hunters and two wings (Stander, 1992). While individual lions have preferred roles that correlate with successful hunts, they can nonetheless switch between roles, suggesting they represent the different roles and understand that all roles must be filled for the hunt to succeed. Naked mole rats also show role specialization (e.g., between defense, working to sustain the burrow, and caring for young; Faulkes & Bennett, 2021), and they switch to less preferred roles both when a preferred role is occupied (Gilbert et al., 2020) or when the animals previously fulfilling a role are removed (Mooney et al., 2015). Another striking example comes from honeybees, which flexibly shift from foraging to receiving nectar. When returning foragers can't find a bee to receive nectar they perform a tremble dance (distinct from the more famous waggle dance that indicates the source of nectar); observing foragers sometimes then switch roles to become receivers, maintaining an optimal balance between these two activities (Seeley, 1997).

These representations reflect the most basic capacity to build a mental model of an institutional structure—i.e., a set of interconnected roles occupied by and constraining the behavior of a set of agents. But they do not yet constitute a full institutional stance, for two

reasons. First, conspecifics only apply role representations to the actual roles that exist, rather than employing them to interpret any form of collective behavior, where roles can be a useful fiction through which to make sense of what is happening. Second, the set of institutional structures that conspecifics can represent is predetermined (Fig. 2c). There is no evidence of variability or cultural innovation across groups, even when they live in ecologically very different circumstances (di Fiore & Rendall, 1994). However, within these bounds, these representations are already minimally generative, in that the exact number of roles is mutable e.g., a dominance hierarchy can expand and contract based on group size.

3.3 Level 3: Generative institutional representations

Species that evolved fixed institutional representations have internally generative models of the social organizations they are part of (e.g., representing hierarchies of arbitrary size or multiple friend cliques). But as social life becomes more complex, social species have a growing repertoire of increasingly complex social organizations that govern behavior in different circumstances. Chimpanzees, for example, have separate male and female dominance hierarchies, friendship coalitions, caring relations, and troop patrols. At this point, we hypothesize that three pressures come together to create a fully-realized institutional stance. First, the growing repertoire of social organizations creates a pressure to efficiently place them in a common representational space. Second, the complexity of social life creates an increased pressure to reason about how different forms of social coordination get combined (e.g., how to predict the behavior of a conspecific who is simultaneously dominant in the hierarchy but also a family member; Bergman et al., 2003). Finally, the separation of role and role occupier allows increasing recognition that some social arrangements are more beneficial than others, entailing new advantages if one has the capacity to change those arrangements.

This leads to the full-fledged institutional stance, which we propose is uniquely human. There are two key differences in this final implementation. The first is that the representations over institutions are fully generative: they no longer rely on a fixed set of evolutionarily ancient role-based systems. Instead, we have a generative representational system to synthesize novel institutional structures, leading to an ability to understand any kind of role-based system, including ones we have never encountered in evolutionary history, from reading clubs and sports

teams to university committees and governments. And, as we explain in Section 6.4, this interpretative framework becomes a system that is itself reused for creating new institutional structures, leading to an explosion of role based systems in human society. This implies that we are equipped with a grammar for generating institutional structures, which can be built over complex representations of roles. In other words, we have an intuitive *Theory of Institutions*, akin to our *Theory of Mind*. The second key difference is that this ability to create arbitrary social structures allows us to apply this form of reasoning to any social system, making it a full-blown stance. This means that we can use it not only to reason about established systems of social organization, but also as a way to interpret social collectives even when we don't necessarily think the members think of themselves as implementing an institutional structure with roles (e.g., interpreting sibling dynamics or friend groups in terms of roles).

This is the institutional stance, which enables us to (1) understand complex social behavior in terms of structured systems of interconnected roles, (2) compositionally construct new such systems, and (3) use that knowledge to interact with others. Indeed, the ethologists' ability to watch animal behavior and infer the complex latent institutional representations at play is itself the institutional stance in use. Even if the cases we reviewed above turned out not to be precursors to the institutional stance, the propensity of humans to interpret animal behavior in this way illustrates our tendency to read complex social behavior through latent representations of roles and role relations. We now formalize this proposal, review supporting evidence, and discuss its implications.

4. The institutional stance

We now turn to a formal description of the institutional stance. In brief, to take the institutional stance is to represent an individual or a group's behavior in terms of a set of interconnected roles that regulate and guide how behavior unfolds over space and time. Roles are defined by internal representations of how they operate and how they relate to other roles. This network of interconnected roles constitutes the abstract representation of an institution. We describe the representational commitments entailed by each of these key components below.

4.1 What is an institution?

When deploying the institutional stance, an institution is a type of cognitive representation consisting of a set of inter-connected roles with a level of persistence and continuity. These representations don't need to be explicitly codified or mapped onto formal institutions, and they can also represent evolutionarily ancient social interactions like caring relations, as well as informal relations like friendships. Specifically, an institution can be thought of as a representation of a graph, where each node represents a role and each edge represents a relation.

Representations of roles are defined by the content of the role (i.e., how does one behave in that role?) and the conditions through which an agent can occupy or leave it. While the specifications of role content are likely flexible and can be culturally modified, most roles have three properties. First, they encode tasks or functions (i.e., the everyday meaning of *role*) that the role occupier prototypically enacts. Second, they encode socially-granted powers that allow particular affordances specific to the role. Third, they impose limitations to how one can wield the socially-granted power and how they complete their tasks or function. These components typically reference other roles (i.e., they are graph edges linking the role to other roles). For example, a barista has the task of fulfilling customer orders, the power to charge their credit cards, and the constraint that they must be friendly. But in some cases, they can also be stand-alone role features. For instance, we can represent a janitor without referencing other roles, as charged with keeping a building clean, and with access to all rooms after 5pm. Critically, roles do not always need to have all three components. We can represent roles that don't encode any task (e.g., care recipient), roles that don't grant any power (e.g., lowest ranking member in a hierarchy), and roles that don't impose any limitations (e.g., authoritarian dictator). However, at least one of these features must be present. A role with no purpose, power, or limitations has no behavioral or interpretative consequences.

In some cases, occupying a role is accomplished by simply enacting it, and one can leave a role by simply ceasing to behave as expected (e.g., one stops being a friend by no longer enacting the role of a friend). In other cases entry conditions are more formalized, such as needing to be approved by the occupant of another role. Thus, entry and departure conditions

frequently reference other roles who have the power to grant or coerce agents into other roles, as well as to revoke them. But the powers and limitations generated by the role are contingent on actively occupying it; for example the barista can be a customer while off duty.

Critically, an institutional representation does not need to be a complete representation of the graph's structure. More often than not, representing a subset of the graph is enough. For instance, our naive representation of universities only references a small set of roles that are most relevant to us (e.g., undergraduates, graduate students, department chairs), without representing the entire complex structure of actual roles. Each representation of a role can also be simplified relative to the true nature of a role, such as an undergraduate representing a professor role in terms of only teaching obligations. However, even when reasoning about a single role (e.g., the janitor), we are aware that the role is situated within a broader system of roles. This knowledge helps us navigate the institution effectively: e.g., if we wanted to change the time of day in which the janitor enters our lab we might speak with them directly, but we might also consider whether the right intervention involves a discussion with someone else, say an office manager. Rather remarkably, we can do this even if we had never previously thought about the janitor's position in the institutional structure at all; this is a hint as to the generative nature of institutional representations, something we return to below.

4.2 Intuitive institutions, theories of institutions, and real institutions: What's the connection?

The building blocks of our framework are similar to those used in formal theories of institutions developed in social sciences and philosophy (e.g., Crawford & Ostrom, 1995; Hodgson, 2006). We believe this is no coincidence. In the beginning of any formal endeavor, theoreticians' have no other way to evaluate the plausibility of a proposal other than comparing it against their intuitions. Inevitably then, seminal formalizations of different phenomena often share crucial similarities to our intuitive theories. To give just two examples, human intuitive physics is surprisingly similar to the first formalizations of physical laws (Battaglia et al., 2013) and, within mentalizing, intuitions about how others make decisions are also strikingly similar to principles of classical economics (Jara-Ettinger et al., 2016). The origins of a formal inquiry of social ontology and theories of institutions may be no exception. This suggests that looking into how theoreticians formalize institutions, particularly within social ontology, could hold valuable

clues as to the nature of our intuitive theories of institutions (e.g. Guala, 2016; Crawford & Ostrom, 1995; Searle, 2005).

Further, adopting the institutional stance allows analyzing social relationships in terms of institutional representations even when those representations do not clearly map onto the external institutional structure. This can occur when the external structure may or may not even be institutional (as in the friend group example mentioned above) as well as when the useful institutional representation departs from the formal institutional (e.g. when reasoning about a cafe, it is pragmatically useful to consider the role of customer in relation to the role of barista, even though the customer is not part of the formal institutional structure at all). This highlights the distinction between our focus on the institutional stance as an intuitive theory of person perception and the analysis of formal institutions that other social scientists might undertake.

5. Relation to competing perspectives on social reasoning

Our proposal builds on a broad family of approaches that aim to explain how people understand each other in ways that cannot be reduced to mentalizing. However, there are also alternative accounts that aim to explain similar phenomena without positing structured role representations as their foundation. These can be roughly organized into accounts for how people represent social collectives, how we represent socially-structured behavior, and how we represent social interactions.

5.1 Representing collectives

The idea that people not only represent other people's minds, but also represent social collectives has a long history in cognitive science and philosophy, grounded in the idea that humans have a *folk sociology*—an intuitive theory of how people organize around notions of groups that help us identify social types, evaluate them, and develop shared expectations for how group members behave (Rhodes & Baron, 2019; Hirschfeld, 2013; Clément et al., 2011; Ritchie, 2020; Shutts & Kalish, 2021).

Work in folk sociology has revealed that, from early childhood, people have a natural tendency to categorize people. This is most prominently in our tendency to divide the social world into “us” versus “them” (and positively evaluate the “us” in myriad ways; Dunham, 2018). But it also includes a large set of group categories that are essentialized—i.e., thought to be determined by patterns of deep internal, perhaps genetic or biological, similarity (Gelman, 2004; Rhodes & Mandalaywala, 2017), like race and gender (e.g. Prentice & Miller, 2007)—such that these categories are represented as stable and immutable. These types of social representations in folk sociology are not part of the institutional stance, in that they are intuitively represented as ‘natural’ categories intrinsically embedded in and stable to the person, rather than represented as interconnected roles that one can pass into and out of. At the same time, recent work in folk sociology has noted that not all social categories are essentialized. Many social groups are better defined via deontic or normative properties (i.e. via rights and obligations assigned to members; e.g. Foster-Hanson & Rhodes, 2019; Kalish & Lawson, 2008). Moreover, both children and adults believe that social and cultural forces can produce meaningful social kinds (Noyes & Keil, 2020b), via social agreement (Noyes et al., 2017) and formal membership criteria (Noyes et al., 2020a). Even the most frequently essentialized social categories such as race and gender can be reconceptualized in a role-like manner; indeed, we believe this is precisely what social ontologists have done in arguing that such categories are actually structured representations of hierarchical roles, such that to be conceived of as a woman or racial minority is to occupy a subordinate position (Ásta, 2018; Haslanger, 1995, 2000).

We believe that these second kinds of social categories emerge from the institutional stance, and they should be therefore understood as reflecting a separate cognitive system that is not supported by essentialized representations (Noyes et al., 2021). Thus, while we agree with this recent work rejecting essentialism as a universal representation of social categories, we diverge from it in two ways. First, past work has suggested that these socially-constructed categories are not a primitive type of representation in the human mind, and that they emerge rather late in development (approximately by age seven; Noyes et al., 2018, 2020), whereas we believe that institutional structures are evolutionarily ancient and readily available in infancy (although some components involve maturation and/or learning; see Section 6.1). Second, much of this past work has studied social groups as undifferentiated containers of group members in

which exchangeable individuals inherit properties (whether traits and behaviors or rights and obligations) directly from the group's attributes. In our account, these representations have rich internal structure specifying how roles are related to others in specific ways (see also Noyes et al., 2021; Ritchie, 2020). In this way, the institutional stance aligns with recent work taking seriously the structured relationships between groups and the distinct patterns of reasoning that emerge from them (Amemiya et al., 2022; Vasilyeva & Lombrozo, 2020 ; Vasilyeva et al., 2018; Yang et al., 2021).

5.2 Representing socially-structured behavior

In our account, institutional representations are grounded in representations of roles, each with membership conditions and content. While we are relatively agnostic about the exact internal representational nature of these components, norms are a promising candidate representation that could encode the duties, power, and limitations in a role. It is thus worth considering whether this view, in its strongest form, means that we don't even need to appeal to roles, and our proposal can be reduced to the idea that people encode agent-relative norms that affect behavior. In what way does the institutional stance go beyond this idea?

Normative systems are critical for social coordination, in three ways. First, norms reduce coordination costs by making behavior predictable; if I know what you are likely to do I can easily adapt my behavior to be complementary (and you can do the same). Second, since norm violations undermine cooperation, violators may face reputational costs or retaliation. Knowing this incentivizes compliance and produces a "sense of should", i.e. an internalized belief that one should comply (Theriault et al., 2021). Third, given potential costs associated with negotiating ad-hoc norms on the fly, norms tend to stabilize within groups, allowing coordination to proceed more easily.

In particular, some recent work has suggested that some integration of folk psychology with norms can provide a powerful framework for reasoning about others (e.g., Eickers, 2024; McGeer, 2021, Andrews et al., 2021). Our perspective also asserts that norms are critical to maintaining cooperative social life and extends this claim in two ways. First, even if the institutional stance is composed from basic representation of norms, the way in which norms are structured is not trivial, requiring mechanisms for representing norms as bundled into roles and

related to other roles in normatively laden ways. Bundling norms into roles requires the machinery we elaborated on above, i.e. specifying membership conditions, limitations, and/or socially-granted powers. Further, the cognitive machinery necessary to structure and productively build complex social structures over roles is not trivial, and is the key to turning the social world's vast set of norms into a useful interpretive system.

The second and related point of departure is that, in grounding the institutional stance in norms, it suggests the need for a special type of *institutional norm* that has largely been neglected in the literature. Classical accounts divide norms into two types: *injunctive* (enforced through social pressure, e.g., being quiet in a library) and *descriptive* (emergent behavioral patterns without enforcement; e.g., clapping after a play (Kallgren, et al., 2000; Lapinski & Rimal, 2005)). Institutional norms have three features that do not fit this dichotomy: (1) they are role-dependent, with their scope defined by a role and its position in the underlying institutional structure (rather than scope being defined by context, statistical prevalence, or degree of social enforcement); (2) adherence to a norm is represented as a causal consequence of occupying a role (i.e., we reason that norm adherence can be explained by the role, a la formal explanation), rather than a merely descriptive pattern or a prudential one generated by fear of retaliation; and (3) failing to adhere to a norm is often seen as evidence that the agent is no longer occupying a role, rather than as a punishable offense or as evidence that the norm is less descriptive of the population (e.g., spotting this morning's barista sitting at a table at a coffee shop and eating lunch would not lead us to revise our beliefs about barista roles, and we wouldn't expect them to be subject to punishment for non-compliance, nor would we walk up to them and ask for a coffee). Many norms within the institutional stance may well be special cases of descriptive or injunctive norms. But our point is that reasoning about when a norm becomes relevant or active in social settings requires a latent representation of the institutional structure, such that an account that simply enumerates injunctive and descriptive norms will be insufficient to produce the type of reasoning the institutional stance provides.

Moving up to the representation of a role as a unit, some related work has argued that these representations are instantiated by a capacity to share intentions (in psychology: Tomasello, et al., 2005; Tomasello & Carpenter, 2007; in philosophy: Searle, 2006; Gilbert, 1992). Consider the task of collaboratively opening a cardboard box, in which one person holds

the box while the other cuts it open (Tomasello et al., 2005). Each agent has a set of focal goals they must keep in mind (hold the box steady; cut the tape with the box cutter) but they must also keep in mind the goal of their partner, and adjust their behavior to comply with the higher and lower-level implementation their partner adopts. Once this entire structure is represented, role switching is unproblematic; the holder can become the cutter and vice versa.

Proposers of shared intentionality have argued that this capacity is unique to humans (Tomasello & Carpenter, 2007; Tomasello & Herrmann, 2010; Tomasello & Moll, 2010) and that it is a cognitive prerequisite for role representations that produce cooperative behavior (Moll & Tomasello, 2007; Tomasello, 2022). Therefore, this view might imply that the institutional stance ultimately depends on mentalistic representations. We do not believe this is the case. Shared intentions are undoubtedly part of how people represent roles in the institutional stance, but a core claim of our perspective is that this needn't, and often isn't, the case. The institutional stance allows us to analyze any behavior as if it were governed by latent roles, even when the agents are performing role-like behavior without a representation of roles or joint intentionality. In this sense, the roles within the institutional stance are not *Tomasellian roles* (Tomasello, 2020). Thus, we can think of emergent social coordination in terms of roles without committing ourselves to the view that the social creatures are engaging in joint intentionality.

5.3 Representing interactions

Finally, at the lowest level, the institutional stance represents social interactions. One prominent set of alternatives comes from schema and script approaches to social cognition (Bartlett, 1932; Eickers, 2023, 2024; Mandler, 2014; Taylor & Crocker, 2022). Schemas refer to knowledge that structures and organizes a particular behavioral domain, which is sometimes analogized to a dramatic script. For example, a supermarket script might describe the types and sequences of activities that take place in a supermarket. This knowledge is organized via the roles that frequently occur in that setting as well as the activities that characterize those roles. For instance, it might specify the role of *shopper* in terms of activities like getting a shopping cart, gathering groceries, going to the registrar, paying, and leaving with groceries; a role of cashier in terms of things like scanning each item, placing items in bags, accepting payment, and so on. Critically, script approaches provide a domain general account of how information governing

familiar domains of behavior is organized. One can in this sense have a script for solitary and non-social activities such as building a campfire.

In our view, script representations are a broad capacity that reflects general intelligence and can be applied to any domain. Indeed, these can be thought of as model-free representations of sequential events, social or not. The institutional representations proposed in our account go beyond simple sequences of how actions typically unfold and so are able to explain behavior not merely as the next line in a script, but rather as the causal consequence of institutional representations. This can be seen by considering the limitations of a social representation that has nothing but scripts. Such a representation would only be useful as long as events unfold as unexpected, but would be of little use when an unexpected event happens. By contrast, the institutional stance also provides guidance as to what to do when an expected social interaction breaks down. For instance, if a barista deviates from the expected behavior by being rude or unhelpful, we are not merely at a loss about what to do. Rather, because we represent the cashier as linked to other roles such as a manager, we can flexibly consider how to intervene (at least, if we decide it's worth the time). Thus, while script-like representations are a basic form of efficient storage that high-level cognition makes use of in a variety of ways, they become dynamic and generative when embedded within a mental model of role-based behaviors and the relation between them.

Another important framework for thinking about social relationships arises out of the Relational Models Theory (Fiske, 1991), and a recent proposal that Fiske-style relational models are at the heart of an innate “core knowledge” system for thinking about relationships (Thomas, 2025) in which authority, communal sharing, and equality matching are the initial building blocks of more complex understandings of relationships. In our view, these relational models are simple institutional structures capturing a particular configuration of roles and relations. They emerge early and are culturally ubiquitous because the simplicity of their structure and role contents makes them useful tools for understanding real-world phenomena such as ingroup-outgroup structures and dominance hierarchies. Thus, rather than proposing a small number of innate relationship templates, we propose that even infants are able to flexibly construct all of these institutional models, and likely many more, via a particular type of generative grammar operating over roles.

6. Institutional stance in use

Having defined the institutional stance, we now turn our discussion to questions that arise about this framework: How does the institutional stance emerge and develop (Section 6.1)? How is the institutional stance deployed in everyday life (Section 6.2)? How do the institutional stance and the mentalistic stance interact (Section 6.3)? And what kinds of social behaviors does the institutional stance uniquely explain (Section 6.4)?

6.1 How are institutional representations acquired?

6.1.2 The emergence and development of role-based representations

According to our view, the core capacity to represent institutional structures is not culturally acquired but evolutionarily endowed (although the capacity to represent large and complex multi-role institutions will naturally depend on domain-general capacities like memory and attention that develop substantially over early childhood). This predicts that even infants should show a basic capacity to represent behavior in terms of roles. Consistent with this, infants can readily represent social interactions in terms of stable dominance representations, which they use to make predictions (Enright et al., 2017; Mascaro & Csibra, 2012; Thomsen, et al., 2011; Stavans & Baillargeon, 2019; Mascaro et al, 2023). Critically, they also distinguish dominance through coercion (which does not require a role representation) from dominance that has been socially granted (i.e., role-like; Margoni, et al., 2018; Thomas, et al., 2018), and they make transitive inferences from pairwise dominance displays (Gazes et al., 2017; Mascaro et al., 2023), revealing they are representing the roles as part of a larger structure. Similarly, infants can also readily represent social interactions in terms of caring relations (Spokes & Spelke, 2017), which can be particularly difficult to understand under a mentalistic stance (Gopnik, 2023).

Can infants also represent more abstract types of roles (i.e., detecting roles that are not necessarily tied to evolutionarily ancient ones like dominance or caring)? Some evidence suggests this is the case: infants can encode simple social interactions in the context of latent social structures (Powell & Spelke, 2013; Thomas et al., 2022), and find it surprising when an agent and patient swap behaviors in an interaction, suggesting that they detect role reversals (Golinkoff & Kerr, 1978; Schöppner, et al., 2006; Tatone, et al., 2015). Similarly, infants readily

represent object transfer events in role-like terms: an actor *giving* an object to a recipient (Tatone, et al., 2019).

Research with older children also decisively shows that their patterns of inference and prediction show different signatures than mentalizing. For example, in a series of studies (Clément et al., 2011), 3-yr-old children show the classic failure to reason about false beliefs by predicting that a person will look for an object where it actually is rather than where, based on the available information, the person ought to believe it is. But when this same dilemma is framed as the application of a rule (e.g. dolls are always in the toy box) these same 3-yr-olds expect the agent to look in the toy box even though the children themselves know it is no longer there. Similarly, classic inferences associated with the mentalizing system, like inferring that a hinderer has bad intentions (Ullman et al., 2010) or that an agent who incurs a higher cost has a strong preference (e.g., Jara-Ettinger et al., 2016), change markedly when the agent is presented as performing a role. In such cases, a hinderer is seen as performing a role-linked requirement (e.g., preventing someone from entering a restricted area), and incurring higher costs does not reveal agent's personal preferences but rather the normative entailments of the role (Baker, et al., *under review*).

6.1.3 Aligning institutional representations to the social world

While infants can already represent social interactions in terms of roles, the exact nature of how they represent the content of roles develops, including for relations that are evolutionarily ancient. In the case of caregiving, infants' first-person experience shapes how they represent caring relations: infants that are securely attached (Ainsworth, 1979) have stronger abstractly coded expectations that a caregiver will comfort a distressed agent (Johnson et al., 2007, 2010). In the case of hierarchies, infants readily attend to the existence of culturally-specific social categories, and can learn the roles through passive language learning—evidenced by Tseltal infants in southern Mexico learning their languages' honorific terms from indirect passive exposure (Foushee & Srinivasan, 2023).

At the same time, being able to represent the institutional structures that align with complex real-world social dynamics is a cognitively complex activity that we suspect must take years to develop, and this may be an underappreciated dimension of cognitive development in early childhood. Consistent with this, young children selectively devote time, attention, and memory to learn the social structures in their world. Children have better knowledge and memory of role-based events than specific events (Fivush, 1984), suggesting they are motivated to learn these patterns. Moreover, their ability to make inferences from role-based social interactions (such as whether the shoppers paid the cashier) surpasses their ability to make simple logical inferences or inferences from general world knowledge (Hudson & Slackman, 1990), suggesting that they are learning to encode the role-based social information into an institutional structure that they can use to reason about the social world. At these ages, children are also strongly motivated to learn norms (Schmidt & Rakoczy, 2023), and they represent them as having scope that is restricted to only a set of individuals (Schmidt et al., 2012), which they can infer purely through observation (Partington et al., 2023), revealing that they have the capacity to begin to detect, infer, and organize norms into role-based representations.

Critically, children do not just learn how to reason about institutional representations via observation alone, but also by enacting them. Children in social groups are notoriously eager to enforce the social norms they've learned (Partington et al., 2023; Schmidt et al., 2016; Rakoczy et al., 2008, 2013), and they spontaneously create basic social structures in their group interactions (including hierarchies similar to those seen in non-human primates; McGrew, 1972; Sluckin & Smith, 1977). More strikingly, children also practice and reinforce their understanding of role-based behavior through pretend play. While children are often characterized in terms of their rich imaginations, much of what they actually imagine is relatively mundane, frequently focusing on the enactment of role scripts (Harris, 2021): much of children's pretend play is *sociodramatic*—pretend play that enacts real-life roles such as caregiver, shopkeeper, doctor, or schoolteacher, as opposed to fictional ones (Christie, 1982; Corsaro, 1993; 1979), and children engage in these activities in part because they are unable to actually do them (Taggart, et al., 2018; 2020). These behaviors provide a powerful way for children to acquire, refine, and solidify their causal models of latent institutional structures.

6.1.4 Inferring institutional structures as adults

To what extent are the learning algorithms that humans use in this process specific to the institutional stance? While this is an open question, research into adults suggests that this capacity might be a combination of domain-general and domain-specific algorithms.

First the capacity to learn institutional representations is likely supported by the general ability to detect the covariation between behaviors and situational contexts (Kelley, 1980; Seiver et al., 2013; Ruffman et al., 2012). This work has largely explored how people infer the extent to which people's behavior might be due to a situational context (rather than an intrinsic and stable mental state), but these same principles should have the power to help detect roles and infer their content. For instance, imagine walking into a building and having a person greet you. You might initially assume they were just being polite, but if you continue visiting the building throughout the year, and notice that there is always a person there waiting to greet you, this pattern of covariation reveals the presence and content of a role.

Similarly, adults can use Bayesian inference to infer key aspects of institutional structures such as group boundaries (Lau et al., 2018; Gershman & Cikara, 2020) and to reconstruct the institutional graph that best explains how agents are interacting with one another (Davis, et al., 2022). According to Bayesian frameworks, doing so requires integrating domain-specific capacities—in this case, a causal model of institutions that specifies the likelihood function (i.e., the likelihood through which different hypotheses about institutions would produce the observed behaviors)—with domain-general ones that integrate priors and likelihoods to produce posterior beliefs (see Griffiths et al., 2024). Here, research has also found that knowledge of roles also helps people quickly generalize behavior, even when they don't know the full set of interconnected roles (Hackel & Kalkstein, 2023).

Finally, the ability to mentalize likely makes learning institutional structures easier, particularly for adults. When we realize that an agent should be interpreted through an institutional stance, we can apply our Theory of Mind to their behavior and attribute the underlying inferences to the role rather than to the individual. For instance, imagine watching an unfamiliar sport, like the pre-columbian game *pok-ta-pok*. Applying standard Theory of Mind, one would infer that the players are trying to hit a rubber ball with their hips to reach the

opposite wall or pass it through a stone ring, which we can then interpret as the rules of the game, rather than idiosyncratic mental states of the individuals (Whittington & Bradley, 2001).

6.2 When do we deploy the institutional stance?

Given that people can analyze behavior in terms of mental states, when do we choose to apply the institutional stance instead? This can be a challenging problem, because many behaviors can be explained *either* as the expression of a role *or* as a combination of an appropriate belief and desire.

At one extreme, it is possible that the institutional stance is pervasively and automatically operating in all or nearly all social settings. This would follow if its evolutionary history and centrality to social cognition runs as deep as mentalizing, which has often been described as automatically deployed (Kampis & Southgate, 2020). Even in the cases where we are richly mentalizing about others, latent role representations—such as ‘friend’, ‘co-worker’, or even ‘stranger’—might nonetheless provide a set of expectations that could greatly simplify social settings (e.g., they might help us assign traits that simplify mentalizing; Westra, 2018, 2020). With this in mind, return to our paper’s opening example, where Kate asks Eric to fulfill certain roles when they run into each other at the cinema. The surprise of the untowards request could come not only from knowing that the previous roles are no longer operative, but also because we know and are automatically tracking new roles— ‘acquaintance’ or ‘friend’—that have very different entailments.

Interestingly, some evidence about the automaticity of these processes comes from classical work that is usually thought to showcase the primacy and automaticity of mentalizing. Consider Heider & Simmel’s (1944) classic study showing that people spontaneously interpret the behavior of geometrical shapes as agentive. While some of those interpretations evoked mental state language, many of the original participants also used role-based language, e.g. talking about villains and heroes or parents and children. Moreover, the basic units of analysis often invoke primitive role relations like chasing and fighting, which are detected from low-level statistical movement cues that might be culture-independent (Barrett et al., 2005). Some more recent evidence has also found that, in self-paced reading tasks, the mention of a role triggers immediate expectations about how the agent is likely to act, whether they are replaceable (i.e., a

role can be replaced by another agent), and even what kind of knowledge they might possess (Baker et al., 2024). This suggests that some aspects of the institutional stance might be pervasive and automatic.

The breadth of the institutional stance can also be illustrated by cases in which agents take action to encourage others to apply the stance to them. This is particularly the case when their behavior is the result of a role and where it is critical for observers to recognize that fact (e.g., police officers directing traffic). In other words, to ensure that our own behavior is appropriately interpreted through the institutional stance when relevant, we broadcast the roles we occupy to help resolve potential ambiguity. The efforts to do so have been dubbed “status indicators” (Searle, 2006). Visible markers like uniforms—which are pervasive in the social world, such as those worn by police officers, priests, lifeguards, doctors, and baristas—are notable examples. Consider Figure 3, which shows two identical agents taking identical actions in identical contexts. The simple presence of an apron leads to a radical revision to how we interpret the behavior and mental states of the agent on the right. In the first case, we might infer that the person likes olive oil, intends to get some, and is likely to cook a meal with it that week. In the second case, the same behavior no longer reveals any personal attitude towards olive oil, and we instead read the behavior as revealing a role-based task, such as ensuring that the oil has not yet passed the sell-by date.

In related cases agents signal their institutional status in other ways, most notably verbally. Emails may begin with a disclosure such as “I’m writing in my capacity as Chair of the Search Committee” to indicate that the action one is taking (perhaps rejecting a candidate) should be attributed to the carrying out of an institutional role rather than the agent’s personal beliefs and desires. In some such cases speakers even explicitly disavow the mental states that might normally be inferred from the behavior (“it gives me no pleasure to have to write this letter...”), a form of speech that can be interpreted as urging the recipient to consider their behavior the enactment of a role demand rather than a personal desire. These institutional disclosures are even deployed in more informal roles: consider someone in a heated debate declaring that they are simply “playing the devil’s advocate”.

Beyond physical institutional indicators (e.g., uniforms) and linguistic constructions (e.g., explicitly disclosing an institutional role), markers of institutionality might sometimes even be embedded in language itself. For instance, many languages use different pronouns to encode levels of formality (Brown & Gilman, 1960; Goffman & Giglioli, 1972) that coincide with institutional roles. Examples of this include the *T-V distinction* in Indo-European languages (e.g., *tú/usted* in Spanish), honorifics in Japanese in which relative status is encoded in verb endings (McCready, 2019), or Nahuatl honorifics (a four-level morphological system; Hill & Hill, 1978). It is possible that these linguistic markers are used not only for deferential purposes but to signal when an engagement ought to be read through an institutional stance rather than a mentalistic one.

Despite such efforts to encourage or discourage the adoption of the institutional stance, there will remain many cases in which people must determine whether or not to adopt it in the absence of clear status indicators. In these cases, the ability to detect automatic behavior can help us detect scripted behavior that could be caused by a role (Ulman & Bass, 2024). Knowledge of the institutional structure that one is operating within also likely helps us infer the relevant roles. Imagine, for instance, walking into a coffee shop and having a person in plain clothes greet you and ask you what you want. In this case, the low likelihood that a stranger would be motivated to spontaneously inquire about your general desires, combined with our knowledge of the roles typically found in coffee shops and our awareness that they are frequently not visibly marked, is enough for us to identify a role agent and so adopt the institutional stance with respect to them. In cases like these, it may be that the institutional structure is assumed to be well known that leads to a decreased use of institutional markers. While these strategies might be particularly powerful for adults, they might also be available to children in some capacity. For instance, young children have expectations about what types of costs are reasonable to incur for others (e.g., ask for help only when the helper can accomplish the goal more easily than you could alone; Grosse et al., 2010; Jara-Ettinger et al., 2020b), and violations of such expectations might help reveal roles and the tasks they perform.

Nonetheless, the strategies described above, both identifying institutional indicators and inferring roles from knowledge of institutional settings even in the absence of those indicators, are more likely to work for enculturated adults with rich knowledge of the institutions they

encounter. It cannot account for actions within highly unfamiliar or entirely novel institutional structures (a domain which, even if modest in adulthood, will be vast earlier in development). This makes the mechanisms for inferring the hidden institutional structure reviewed in 6.1 critical, a process likely boosted by explicit cultural transmission concerning institutions and their central roles.

6.3 Interplay with mentalizing

6.3.1 Institutional stance without mentalizing

Given that humans have a mentalistic stance and an institutional stance, how do the two interact? To start, the institutional stance can reduce the cognitive costs and uncertainty associated with mentalizing: Structured role representations allow us to effortlessly engage in an otherwise complex interaction without invoking representations of beliefs or desires. For instance, if you order a coffee at a coffee shop and interpret the barista under the institutional stance, you might then simply assume that the goal will be fulfilled, and that a coffee will materialize within a few minutes, without the need to actively track and infer the barista's mental states to determine whether the barista indeed has the goal of preparing the coffee for you or not. This could be broadly similar to how we can interact with a vending machine expecting an outcome, such as (an admittedly awful) coffee to appear, without attributing a rich mental life to the vending machine. In both cases, this is possible because of a structured understanding of particular contexts defined at least in part via how our actions in one role will tend to produce reactions in other roles.

There are a few lines of research suggesting that the institutional stance could effectively reduce the online computational burden of mentalizing. First, the development of scripts as a tool for social reasoning came from early practitioners of artificial intelligence (Schank & Abelson, 1977) developing socially interactive agents that explicitly lacked true mentalizing abilities. Independently, research in game theory has found that problems that are difficult to solve via mental state coordination (e.g. who will lead or follow the dance, and do we all share that common knowledge?) can be simplified by transforming them into non-mentalistic rule following, simply by conditioning behavior on a rule invoked via some feature of the context initially external to the game (e.g., men lead and women follow; Guala, 2016; O'Connor, 2019).

Further, when an observer knows the structure of possible actions an agent might take in a situation, action understanding becomes computationally simple, no longer requiring complex Theory of Mind (Davis & Jara-Ettinger, 2022). Clinical research also suggests that people with Autism Spectrum Disorder—who often have difficulties in mental state reasoning (Baron-Cohen, 2000)—benefit from being taught detailed script knowledge for a range of social interactions (reviewed in Akers et al., 2106), again suggesting that the institutional stance is distinct from mental state reasoning. Finally, some empirical work does suggest that norm-based systems for predicting behavior need not rest on mental state reasoning (Kaufmann & Clément, 2014). Together, this work suggests that role-based reasoning within the institutional stance does not directly require mental state reasoning, and can be handled without deploying a mentalistic stance at all (although, naturally, the social representations provided by perception—notions of agent, goal, and intention—ought to be readily available to the institutional stance; McMahon & Isik, 2023).

However, the efficiency of the institutional stance in these cases depends on a particular learning history in which that institutional structure became familiar. Thus, the system's cognitive efficiency is the result of having front-loaded learning costs at an earlier point. In this way it serves as a powerful complement to the mentalistic stance, which can be flexibly deployed when no prior knowledge is available but likely requires a greater real-time commitment of cognitive resources.

6.3.2 Institutional and mentalistic stance in competition

Because many behaviors can be explained either through a mentalistic or an institutional stance, it is natural that these two frameworks might often compete. In these cases, an explanatory or predictive failure of one system can trigger inferences in the other. That is, if an institutional prediction fails we try to reanalyze behavior in terms of mental states, and if a mentalistic explanation fails we are more likely to apply the institutional stance. Returning to the coffee shop example, suppose now that you've been waiting for ten minutes and your coffee is still not ready. You might then start attending to the barista to try to figure out what's going on. If you see them working hectically you might infer that they just haven't gotten to your order yet, whereas if you see them relaxedly talking to a coworker, you might infer that they either lost or

forgot your order. In other words, deviations from role-based expectations (like deviations from other normative expectations) likely trigger a shift towards greater mental state inference, deployed in an effort to make sense of the deviation. Indeed, several analyses of mindreading have posited that its use might become active precisely when non-mentalistic expectations fail (Bohl, 2015; Zawidzki, 2013).

Conversely—although perhaps less commonly—violations of a mentalistic expectation might trigger institutional expectations. Imagine, for instance, that you are friends with the barista, but notice that they are acting unusually formal towards you. Such a deviation might trigger us to consider whether they are encouraging you to take an institutional stance (e.g., perhaps their boss is there for the day). These switches can happen frequently within a single interaction. If the barista were your friend, you might quickly alternate between catching up and placing an order, which would require rapidly knowing which aspects of the interaction are institutional and which are mentalistic. Our ability to do this suggests that we are able to have active parallel representations of agents' mental lives and their institutional roles.

6.3.3 Combining institutional and mentalistic representations

Finally, the ability to represent other people's mental states enables us to create and represent richer types of institutional structures than would be otherwise possible. This is because we can encode expectations about mental states within roles and their relations (Fig. 1c). For instance, we can express interactions where one agent's role is to satisfy another agent's general desires (such as in the office example in the introduction), or where occupying a certain role requires having certain mental states (such as expecting a stock person at a supermarket to know where items are located). In many of these cases, even if the roles require us to represent mental states, these are tagged accordingly. For instance, in the supermarket stock person example, we might locally represent them as *wanting* to stock the shelves quickly to complete their shift early, but we are simultaneously aware this is not because they find this activity intrinsically rewarding and indeed would probably rather be doing something else. Similarly, their detailed knowledge of where things are stocked would not lead us to infer that they must be passionate about supermarkets—an inference that might be warranted if this knowledge were acquired in their free time. That said, we suspect that there is widespread variability in how

readily we transfer behaviors or goals from a role to a role-occupying agent. For example, in baseball, the pitcher's goal is to successfully throw the ball to the catcher without the batter getting the chance to hit the ball. In this case, our institutional representation of the role of pitcher does seem to intuitively transfer a goal of striking out a batter to the agent in that role. That is, the pitcher is much more likely to truly want to accomplish their role's goals than is the stock person (Nguyen, 2020). Mentalizing and the institutional stance also combine when we attribute mental states to institutional representations (Fig. 1d). The propensity to ascribe mental states to groups—institutional or not—is common (e.g. Russia *wants* to annex Ukraine, and Ukraine *intends* to resist; Jenkins, et al., 2014; Tang & Gray, 2023), and this might reflect a general capacity to anthropomorphize groups and roles by ascribing mental states to them.

6.4 Social Constitution

We have argued that the institutional stance is a tool for reasoning about social interactions across a wide range of settings, from the evolutionarily ancient (e.g., hierarchies), through emergent forms of coordination where users don't think of themselves as performing roles (e.g., a family or group of friends), to formally codified institutions (e.g., universities). In doing so we have emphasized that the institutional stance is a generative system that, at least in humans, allows us to represent and manipulate arbitrary institutional structures. This ability to treat institutional representations as objects of thought helps explain two uniquely-human aspects of social intelligence.

First, the capacity to read complex multi-agent behavior through the institutional stance means that we can apply this capacity as an analytic tool, including to analyze phenomena that we might not initially think of as being institutional. This is an important engine for social change that enables us to reveal hidden role-based systems that were previously unobservable to us or that were initially conceptualized in other ways. For example, the folk psychological conception of race, gender, and caste-based systems are often essentialized, i.e., thought of as natural kinds. Yet many theorists now analyze them in role-based structural terms arguing that they are hierarchical systems of interlocking social roles in which, for example, *woman* as a structural position defined in subordination to *man* (Haslanger, 2000; Rehbein & Söyler, 2020). Analyzing complex patterns of social behavior in this way can reveal hidden structures, allowing

us to then evaluate them and consider how to modify or intentionally subvert them (Hesni, in press). This gives us the ability to discover systems of role-based behavior that are hiding in plain sight, reason about what kinds of consequences the system produces and, if we determine it to be unfair or oppressive, consider how to improve or erase it. In other words, the generative aspect of human institutional reasoning allows us to imagine and pursue genuine institutional alternatives.

Second, the generative capacity to create arbitrary institutional representations to understand complex forms of social behavior gives a uniquely-human capacity, namely the ability to use the institutional stance to actively invent new institutions. Simply put, if we can imagine a new type of role-based social organization and communicate it to others, people may decide to actually enact it, thereby making it real. In this sense, the generative capacity of the institutional stance catalyzed the explosion of complex forms of social organization, such as universities, sports leagues, fraternities/sororities, study groups, and charity groups.

This productive dimension of the institutional stance is on full display when it generates patterns of social coordination that bring fundamentally new kinds of entities into existence. Take for example a *ballot*, the act of *voting*, and a *citizen*. These kinds are not merely *socially constructed* in a narrow sense, such as the sense in which we might say a tool is socially constructed if it required several people to come together to participate in its fabrication. Instead, each of these kinds is *socially constituted* in a stronger sense in which defining them requires invoking social forces, including social roles and the relations between them (what we here refer to as institutions). As discussed extensively by philosophers in the social ontology tradition (Asta, 2018; Searle, 2005, 2007; Hindriks & Guala, 2015) and more recently by cognitive scientists (Noyes et al., 2018, 2020a), these entities have a unique ontological status in that their existence is not just brought about by but actively sustained by social forces. An institutional process can render a piece of paper a ballot, a particular act of marking that paper voting, and a particular individual a citizen; a further change in institutional structure can reverse those changes even without affecting any intrinsic or even any physical facts about the object, action, or person. This contrast becomes clear when we consider institutional kinds in comparison to objects like *hammers*, acts like *hammering*, and people who are *carpenters*. The factors which make these things what they are can be defined without directly making reference to social

forces (even if, as we alluded to above, social forces may have been involved in initially bringing them about, say by manufacturing them or communally developing a set of practices). Institutional processes might well be involved in regulating the use or manufacture of hammers, the allowable forms of hammering, or the people allowed to perform the role of carpenter, but however such processes play out we would still consider a certain class of objects designed to deliver force a hammer and the act of driving in a nail hammering. On the other hand, a ballot can really cease to be a ballot, and an action can really cease to be voting, as the direct result of institutional processes.

Existing work assessing whether these philosophical distinctions map onto everyday psychology suggests that adults and older children assert that changes in institutional practice (such as what most people do and believe) can make an object like a ballot no longer a ballot, or an action like voting no longer voting. By contrast, they assert that this is not the case for objects like hammers, or actions like hammering (Noyes et al. 2018, 2020c, 2023). In other words, for a special class of institutional kinds, people explain their causal powers not via internal features of the objects or agents, but rather because of the way in which those objects or agents relate to patterns of human intention that are themselves positioned within social structures. This work suggests that people come to see certain classes of objects, actions, and human kinds as *institutional* in the relevant sense, raising the possibility that the institutional stance is the cognitive basis of this form of *social reality*, a vital component of the generativity of human social structures.

7. Conclusions and Looking Forward

When comparing human social life to that of other animals, two differences quickly become apparent. First, face-to-face social interactions are qualitatively distinct from those observed in any other animal. This is often thought to reflect people's ability to build high-resolution models of each other's minds, which are then used to support complex forms of collaboration and communication. Second, human societies have experienced an explosion of social systems that we must navigate: Coffee shops and restaurants where we interact with strangers, special interest groups and neighborhood associations where we interact with friends and acquaintances, and large-scale formal systems of education and government which structure

social life as a whole. If we are correct, these institutions are not a mere byproduct of mentalizing, nor an offshoot of more general non-social forms of intelligence, but rather a direct consequence of a complex framework we have for interpreting—and then creating—social behavior through the institutional stance.

Do we need both a mentalistic and an institutional stance? Most domains we evolved to understand are non-reactive: they do not change in response to the cognitive system analyzing them. The laws of physics do not change as a function of how our mind models physics, and faces are not shaped by how we analyze them (at least within shorter evolutionary timescales). By contrast, social behavior is a dynamic reactive system. Whatever system a social group uses to understand each others' behavior also affects social behavior itself (e.g. “looping effects” in which individuals and groups react to how they are conceptualized by others, in turn affecting those conceptualizations; Hacking, 1995). This means that our social cognition is not attempting to model a passive system, but rather an active one that changes as our social cognition evolves, a state of affairs which necessitates interpretive systems of considerable flexibility. This uniquely dynamic nature of social cognition produced pressure to co-evolve two complementary systems of analysis: A general mentalistic stance for understanding unconstrained behavior in any context, and an institutional stance for representing coordinated action when mentalizing fails or when the context is sufficiently familiar. And while these might have originally emerged as competing solutions to the same problem, the intersection of the two leads us to create complex forms of social intelligence that are unprecedented in any other species.

We have also argued that these two stances should be thought of as comprising distinct and coequal cognitive systems, even if they can be integrated (as when we represent roles with mental states). However, there are alternatives. One possibility is that the mentalistic stance is in some way more primary. The intuition can be motivated by considering how often we explain why someone is in a particular role by appealing to their mental states. A cashier, for instance, is likely a cashier because they want to earn money, and believe this is a good way of doing so given their current circumstances. From this perspective, institutional representations might be grounded in or composed of mental states. While we agree that it is sometimes possible to represent causes behind roles in terms of mental states, a primary empirical claim motivating this

paper is that we very often do not. That is, much of everyday social cognition instead relies solely on rule-like entailments stemming from agent positions in social structures.

A different possibility is that the institutional stance and the mentalistic stance are a single, non-separable system composed of mentalistic representations embedded in roles. The intuition here is that all interactions are with particular social partners that can be conceived of as roles (e.g. friend, stranger, partner), and both the mental states and the roles are necessary to understand the interaction. As discussed in Section 6.3.3, we agree that many interactions involve a combination of the two systems. But it remains fundamental to our proposal that these systems are fully separable. To illustrate why we believe this, consider the minimal representations that allow us to predict behavior from either stance. For the mentalistic stance, you must represent both beliefs and desires to predict action. For instance, you can predict how someone would behave if you learned that they want to get coffee and believe there is a cheap and highly rated coffee shop around the corner. Knowing only the person's desires but not their belief would not be enough to know exactly what they'll do. The upshot is that isolated mental states are insufficient to predict behavior, while combinations of mental states are sufficient, a hallmark of a single predictive system that, in this case, needn't make reference to roles and norms. Moving to the institutional stance, I can predict the behavior of a barista in response to a request for coffee if I have knowledge of both roles and role entailments; I needn't represent anything about the person's mental states. This again suggests that the institutional stance is a coherent predictive system that needn't rely on mentalistic representations.

7.1 Implications and Open Questions

So far, we have reviewed theoretical positions and empirical literatures that, to varying degrees, support our contention that the institutional stance is a core aspect of human social cognition. In the remainder of this paper, we highlight potential implications of the institutional stance for social cognition more broadly, as well as open questions in which new data would be particularly illuminating.

First, the relative reliance on mentalistic versus institutional understanding is likely shaped by culture. For instance, some cultures might have a stronger belief that institutional roles ought to be played out even when they conflict with our own mental states (and hence prioritize

taking the institutional stance), while other cultures may have an expectation that personal values override institutional roles and failure to do so is morally reprehensible (hence reading mental states even when agents declare to be enacting roles), a phenomenon that we could imagine is linked to aspects of individualism versus collectivism (analogous to the notion of tight versus loose societies; Gelfand et al., 2017). Similarly, there is interesting variation in the extent to which members of different cultural groups invoke mentalistic explanation, for example when engaging in moral reasoning (Barrett et al., 2016; McNamara et al., 2019). An open question is whether these cultures put correspondingly greater weight on institutional explanation.

Relatedly, the degree to which people rely on the institutional stance might also vary across historical periods. A notable example is the 18th- and 19th-century regency era of British history—best known in modern times from Jane Austen’s novels—that was characterized by a heavy focus on social class, with the upper-classes regulated by a complex set of social roles, each with tight expectations for social behavior across different social events (Erickson, 1986; Le Faye, 2002). Interestingly, Jane Austen’s writing itself also reveals how taking an institutional stance can trigger extremely rich mentalistic inferences from behaviors that would generally be unremarkable but become highly meaningful when they violate role-based expectations.

It is also possible that the applicability of an institutional vs mentalistic stance does not depend solely on explanatory power but also on cognitive efficiency. In particular, rich mentalizing about agents is resource intensive and capacity limited. To name just one example, it would be daunting to attempt to track all the minds in a large crowd. Yet, people do readily track large-scale role-based complex behavior. This is particularly evident when watching sports like basketball or football, where what is (to a newcomer) an unintelligible confusion of running bodies becomes (to an expert) an intricate yet intelligible ballet. This happens not via a better leveraging of mental state reasoning but rather via fluency with the internal role structure, which renders each agent’s behavior at least in large part predictable and efficiently directed towards higher order goals (a touchdown, say). We suggest that similar dynamics allow us to reason efficiently about other forms of complex multi-party interaction.

Another open question is whether the institutional stance operates purely within high-level cognition, or whether some aspects could be implemented within perception itself. Recent

work has argued that the visual stream has a pathway specialized for social computation, which is both functionally and anatomically different from the ventral and dorsal streams (Pitcher & Ungerleider, 2021; Pitcher, 2021; Grossman, 2021). It is thus possible that visual perception forms a first set of basic representations—agency, goals, and basic social relations (Abassi & Papeo, 2020; Isik et al., 2017; Papeo et al., 2017; Scholl & Tremoulet, 2000)—which can then be interpreted either through mentalizing or through the institutional stance; Fig. 4). But it is also possible that some social computations within vision serve as bottom-up cues for knowing when to deploy the institutional stance. For instance, vision might rapidly extract certain role-like relations from visual input, such as *chaser* and *chased* (Gao et al., 2012; Barret et al., 2005). It is even possible that people exploit principles of how perception works (which did not evolve for the institutional stance) to trigger the institutional stance: For instance, visual synchrony is an important cue within perception for segmenting visual input (Usher & Donnelly, 1998), and groups that should be interpreted through the institutional stance often enact forms of tight synchrony (such as an army marching in unison), which might be a cultural strategy to create a strong visual percept that a set of people ought to be read as a large coordinated institutional group.

We have focused primarily on the power and utility of the institutional stance as a lens through which to do social cognition, but applying it to others likely has downsides. Consider the frequency of attempting to justify morally unjustifiable acts via appeal to the mere enactment of institutional roles, as perhaps most famously in the trial of Nazi officer Adolf Eichmann, in which he described himself as merely executing a role with little further reflection (Arendt, 2006), as well as the moral complexities involved with allocating blame to agents embedded within larger systems (e.g. Gantman et al., 2020). More generally, applying the institutional stance might sometimes reduce the extent to which we represent the mental and emotional life of others, thereby reducing empathy and stripping others of their humanity (Haslam & Loughnan, 2014; Haslam, 2006; Harris & Fiske, 2006). If so, links between the institutional stance and dehumanization might run deep, from mundane cases of “everyday” dehumanization such as engaging with your barista instrumentally to more severe and consequential forms that spur serious conflict (i.e., leading us to contract our moral circle; Graham et al., 2017). If so,

evolution's attempt to help us understand each other might, in some cases, prevent us from seeing who we really are.

But the institutional stance is also a source of great human potential. The cognitive tools we have outlined here seem to be a necessary precursor to questioning existing categories and subverting long-standing exploitative patterns. Thus, the same psychological forces which can reduce people to cogs in institutional structures also offer the possibility of envisioning and enacting new and more just futures. The institutional stance gives us the tools to analyze rather than blindly accept institutional structures, to recognize unjust and exploitative systems, and to imagine possibilities for positive change. In these cases our humanity most shines when we are able to consider—and critique—the roles we play.

Acknowledgments

We thank Paul Bloom, Annya Dahmani, Tania Lombrozo, Eric Mandelbaum, Alexander Noyes, Alexander Paunov, Katie McAuliffe, and Kate Richie for helpful conversations. This work was supported by NSF awards BCS-2045778 and IIS-2106690.

Funding Statement

This work was supported by NSF award BCS-2045778.

Conflict of Interest Statement

The authors declare no conflicts of interest.

References

- Abassi, E., & Papeo, L. (2020). The representation of two-body shapes in the human visual cortex. *Journal of Neuroscience*, 40(4), 852-863.
- Aboody, R., Huey, H., & Jara-Ettinger, J. (2022). Preschoolers decide who is knowledgeable, who to inform, and who to trust via a causal understanding of how knowledge relates to action. *Cognition*, 228, 105212.
- Akhtar, N., Carpenter, M., & Tomasello, M. (1996). The role of discourse novelty in early word learning. *Child development*, 67(2), 635-645.
- Ainsworth, M. S. (1979). Infant–mother attachment. *American psychologist*, 34(10), 932.
- Akers, J. S., Pyle, N., Higbee, T. S., Pyle, D., & Gerencser, K. R. (2016). A Synthesis of Script Fading Effects With Individuals With Autism Spectrum Disorder: A 20-Year Review. *Review Journal of Autism and Developmental Disorders*, 3(1), 1–17. <https://doi.org/10.1007/s40489-015-0062-9>
- Amemiya, J., Mortenson, E., Heyman, G. D., & Walker, C. M. (2022). Thinking Structurally: A Cognitive Framework for Understanding How People Attribute Inequality to Structural Causes. *Perspectives on Psychological Science*, <https://doi.org/10.1177/17456916221093593>.
- Andrews, K., Spaulding, S., & Westra, E. (2021). Introduction to folk psychology: Pluralistic approaches. *Synthese*, 199(1), 1685-1700.
- Arendt, H. (2006). *Eichmann in Jerusalem: A report on the banality of evil*. Penguin.
- Ásta, Á. (2018). Categories we live by: The construction of sex, gender, race, and other social categories.
- Baker, A., Jara-Ettinger, J., & Dunham, Y. (*under review*). Roles trigger agent-general normative explanations for agent behavior beginning in middle childhood

- Baker, A., Dunham, Y., & Jara-Ettinger, J. (2024, May). Roles guide rapid inferences about agent knowledge and behavior. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 46).
- 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(4), 0064.
- Bergman, T. J., Beehner, J. C., Cheney, D. L., & Seyfarth, R. M. (2003). Hierarchical classification by rank and kinship in baboons. *Science*, 302(5648), 1234-1236.
- Baron-Cohen, S. (2000). Theory of mind and autism: A review. *International review of research in mental retardation*, 23, 169-184.
- Barrett, H. C., Todd, P. M., Miller, G. F., & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: A cross-cultural study. *Evolution and Human Behavior*, 26(4), 313-331.
- Bartlett, (1932). *Remembering: A study in experimental and social psychology*. Cambridge, UK: Cambridge university press.
- Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, 110(45), 18327-18332.
- Bergman, T. J., Beehner, J. C., Cheney, D. L., & Seyfarth, R. M. (2003). Hierarchical classification by rank and kinship in baboons. *Science*, 302(5648), 1234-1236.
- Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in cognitive sciences*.
- Boesch, C. (1994). Cooperative hunting in wild chimpanzees. *Animal Behaviour*, 48(3), 653-667.

- Boesch, C. (2005). Joint cooperative hunting among wild chimpanzees: Taking natural observations seriously. *Behavioral and Brain Sciences*, 28(5), 692-693.
- Bohl, V. (2015). We read minds to shape relationships. *Philosophical Psychology*, 28(5), 674-694.
- Bonawitz, E., Shafto, P., Gweon, H., Goodman, N. D., Spelke, E., & Schulz, L. (2011). The double-edged sword of pedagogy: Instruction limits spontaneous exploration and discovery. *Cognition*, 120(3), 322-330.
- Boyer, P., & Petersen, M. B. (2012). The naturalness of (many) social institutions: Evolved cognition as their foundation. *Journal of Institutional Economics*, 8(1), 1–25. <https://doi.org/10.1017/S1744137411000300>
- Brown, R., & Gilman, A. (1960). The pronouns of power and solidarity. *Style in language*, 252-281.
- Cheney, D. L., & Seyfarth, R. M. (1980). Vocal recognition in free-ranging vervet monkeys. *Animal behaviour*, 28(2), 362-367.
- Cheney, D. L., & Seyfarth, R. M. (2019). *Baboon metaphysics: The evolution of a social mind*. University of Chicago Press.
- Christie, J. F. (1982). Sociodramatic play training. *Young children*, 37(4), 25-32.
- Clément, F., Bernard, S., & Kaufmann, L. (2011). Social cognition is not reducible to theory of mind: When children use deontic rules to predict the behaviour of others. *British Journal of Developmental Psychology*, 29(4), 910-928.
- Collette, S., Pauli, W. M., Bossaerts, P., & O'Doherty, J. (2017). Neural computations underlying inverse reinforcement learning in the human brain. *Elife*, 6, e29718.
- Conant, R. C., & Ross Ashby, W. (1970). Every good regulator of a system must be a model of that system. *International journal of systems science*, 1(2), 89-97.

- Corsaro, W. A. (1979). Young children's conception of status and role. *Sociology of Education*, 46-59.
- Corsaro, W. A. (1993). Interpretive reproduction in children's role play. *Childhood*, 1(2), 64-74.
- Crawford, S. E., & Ostrom, E. (1995). A grammar of institutions. *American political science review*, 89(3), 582-600.
- Dasser, V. (1988). A social concept in Java monkeys. *Animal Behaviour*, 36(1), 225-230.
- Davis, I., & Jara-Ettinger, J. (2022). Hierarchical task knowledge constrains and simplifies action understanding. *Proceedings of the annual meeting of the cognitive science society*
- Davis, I., Jara-Ettinger, J., & Dunham, Y., (2022). Inferring the internal structure of social collectives.
- D'Ettorre, P., & Heinze, J. (2001). Sociobiology of slave-making ants. *Acta ethologica*, 3, 67-82.
- Dennett, D. C. (1989). *The intentional stance*. MIT press.
- Di Fiore, A., & Rendall, D. (1994). Evolution of social organization: a reappraisal for primates by using phylogenetic methods. *Proceedings of the National Academy of Sciences*, 91(21), 9941-9945.
- Dunbar, R. I. (2009). The social brain hypothesis and its implications for social evolution. *Annals of human biology*, 36(5), 562-572.
- Dunham, Y. (2018). Mere Membership. *Trends in Cognitive Sciences*, 22(9), 780–793.
<https://doi.org/10.1016/j.tics.2018.06.004>
- Durkheim, É. (1911, March). Jugements de valeur et jugements de réalité. In *Atti del IV Congresso Internazionale di Filosofia* (Vol. 1, pp. 99-114).
- Eickers, G. (2023). Coordinating Behaviors: Is social interaction scripted?. *Journal for the Theory of Social Behaviour*, 53(1), 85-99.

- Eickers, G. (2024). Scripts and Social Cognition. *Ergo an Open Access Journal of Philosophy*, 10(0). <https://doi.org/10.3998/ergo.5191>
- Enright, E. A., Gweon, H., & Sommerville, J. A. (2017). 'To the victor go the spoils': Infants expect resources to align with dominance structures. *Cognition*, 164, 8-21.
- Erickson, C. (1986). *Our tempestuous day: a history of Regency England*. William Morrow & Company.
- Faulkes, C. G., & Bennett, N. C. (2021). Social evolution in African mole-rats—a comparative overview. *The Extraordinary Biology of the Naked Mole-Rat*, 1-33.
- Fiske, A. P. (1991). *Structures of social life: The four elementary forms of human relations: Communal sharing, authority ranking, equality matching, market pricing*. Free Press.
- Fivush, R. (1984). Learning about school: The development of kindergartners' school scripts. *Child development*, 1697-1709.
- Foster-Hanson, E., & Rhodes, M. (2019). Normative social role concepts in early childhood. *Cognitive Science*, 43(8), e12782.
- Foushee, R., & Srinivasan, M. (2023). Infants who are rarely spoken to nevertheless understand many words.
- Francis, B. A., & Wonham, W. M. (1976). The internal model principle of control theory. *Automatica*, 12(5), 457-465.
- Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, 336(6084), 998-998.
- Gantman, A. P., Sternisko, A., Gollwitzer, P. M., Oettingen, G., & Van Bavel, J. J. (2020). Allocating moral responsibility to multiple agents. *Journal of Experimental Social Psychology*, 91, 104027.
- Gao, T., McCarthy, G., & Scholl, B. J. (2010). The wolfpack effect: Perception of animacy irresistibly influences interactive behavior. *Psychological science*, 21(12), 1845-1853.

- Gazes, R. P., Hampton, R. R., & Lourenco, S. F. (2017). Transitive inference of social dominance by human infants. *Developmental science*, 20(2), e12367.
- Gelfand, M. J., Harrington, J. R., & Jackson, J. C. (2017). The Strength of Social Norms Across Human Groups. *Perspectives on Psychological Science*, 12(5), 800–809. <https://doi.org/10.1177/1745691617708631>
- Gelman, S. A. (2004). Psychological essentialism in children. *Trends in Cognitive Sciences*, 8(9), 404–409. <https://doi.org/10.1016/j.tics.2004.07.001>
- Gemmell, N. J., Todd, E. V., Goikoetxea, A., Ortega-Recalde, O., & Hore, T. A. (2019). Natural sex change in fish. *Current topics in developmental biology*, 134, 71-117.
- Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naïve theory of rational action. *Trends in cognitive sciences*, 7(7), 287-292.
- Gergely, G., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, 415(6873), 755-755.
- Gershman, S. J., & Cikara, M. (2020). Social-structure learning. *Current Directions in Psychological Science*, 29(5), 460-466.
- Gilbert, M. (1992). *On social facts*. Princeton University Press.
- Gilbert, J. D., Rossiter, S. J., & Faulkes, C. G. (2020). The relationship between individual phenotype and the division of labour in naked mole-rats: it's complicated. *PeerJ*, 8, e9891.
- Goffman, E., & Giglioli, P. (1972). *Language and social context: Selected readings*.
- Golinkoff, R. M., & Kerr, J. L. (1978). Infants' perception of semantically defined action roles changes in filmed events. *Merrill-Palmer Quarterly of Behavior and Development*, 24(1), 53-61.
- Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, 20(11), 818-829.

- Gopnik, A. (2023). Caregiving in Philosophy, Biology & Political Economy. *Dædalus*, 152(1), 58-69.
- Gopnik, A., & Meltzoff, A. N. (1997). Words, thoughts, and theories. Mit Press.
- Griffiths, T. L., Chater, N., & Tenenbaum, J. B. (Eds.). (2024). Bayesian models of cognition: reverse engineering the mind. MIT Press.
- Grosse, G., Moll, H., & Tomasello, M. (2010). 21-month-olds understand the cooperative logic of requests. *Journal of Pragmatics*, 42(12), 3377-3383.
- Grossmann, T. (2021). Developmental origins of the pathway for social perception. *Trends in Cognitive Sciences*, 25(7), 546-547.
- Guala, F. (2016). *Understanding institutions*. Princeton University Press.
- Hackel, L. M., & Kalkstein, D. A. (2023). Social concepts simplify complex reinforcement learning. *Psychological Science*, 34(9), 968-983.
- Hacking, I. (1995). The looping effects of human kinds. *Causal Cognition: A Multidisciplinary Debate*.
- Hamlin, K., Ullman, T., Tenenbaum, J., Goodman, N., & Baker, C. (2013). The mentalistic basis of core social cognition: Experiments in preverbal infants and a computational model. *Developmental science*, 16(2), 209-226.
- Harris, P. L. (2021). Early Constraints on the Imagination: The Realism of Young Children. *Child Development*, 92(2), 466–483. <https://doi.org/10.1111/cdev.13487>
- Harris, L. T., & Fiske, S. T. (2006). Dehumanizing the lowest of the low: Neuroimaging responses to extreme out-groups. *Psychological science*, 17(10), 847-853.
- Harris, P. L. (2012). Trusting what you're told: How children learn from others. Harvard University Press.

- Haslam, N. (2006). Dehumanization: An integrative review. *Personality and social psychology review*, 10(3), 252-264.
- Haslam, N., & Loughnan, S. (2014). Dehumanization and infrahumanization. *Annual review of psychology*, 65, 399-423.
- Haslanger, S. (1995). Ontology and Social Construction. *Philosophical Topics*, 23(2), 95–125.
- Haslanger, S. (2000). Gender and race:(What) are they?(What) do we want them to be?. *Noûs*, 34(1), 31-55.
- Hector, D. P. (1986). Cooperative hunting and its relationship to foraging success and prey size in an avian predator. *Ethology*, 73(3), 247-257.
- Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American journal of psychology*, 57(2), 243-259.
- Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and brain sciences*, 33(2-3), 61-83.
- Hesni, Samia. (in press). Script disruption and social change. *Constructing Social Hierarchy*. eds. Karen Jones, Laura Schroeter, Francois Schroeter, Oxford University Press.
- Hill, J. H., & Hill, K. C. (1978). Honorific usage in Modern Nahuatl: The expression of social distance and respect in the Nahuatl of the Malinche Volcano area. *Language*, 123-155.
- Hindriks, F., & Guala, F. (2015). Institutions, rules, and equilibria: A unified theory*. *Journal of Institutional Economics*, 11(3), 459–480. <https://doi.org/10.1017/S1744137414000496>
- Hirschfeld, L. A. (2013). The myth of mentalizing and the primacy of folk sociology. *Navigating the social world: What infants, children, and other species can teach us*, 101-106.
- 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.

Hodgson, G. M. (2006). What are institutions?. *Journal of economic issues*, 40(1), 1-25.

Hölldobler, B. & Wilson, E. O. 1990. *The Ants*. Berlin: Springer-Verlag.

Horschler, D., Berke, M., Santos, L., & Jara-Ettinger, J., (2023). Differences Between Human and Non-Human Primate Theory of Mind: Evidence from Computational Modeling. *bioRxiv*, 2023-08.

Hudson, J. A., & Slackman, E. A. (1990). Children's use of scripts in inferential text processing. *Discourse Processes*, 13(4), 375-385.

Isik, L., Koldewyn, K., Beeler, D., & Kanwisher, N. (2017). Perceiving social interactions in the posterior superior temporal sulcus. *Proceedings of the National Academy of Sciences*, 114(43), E9145-E9152.

Jamali, M., Grannan, B. L., Fedorenko, E., Saxe, R., Báez-Mendoza, R., & Williams, Z. M. (2021). Single-neuronal predictions of others' beliefs in humans. *Nature*, 591(7851), 610-614.

Jara-Ettinger, J., Floyd, S., Huey, H., Tenenbaum, J. B., & Schulz, L. E. (2020). Social pragmatics: Preschoolers rely on commonsense psychology to resolve referential underspecification. *Child development*, 91(4), 1135-1149.

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(8), 589-604.

Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020a). The naïve utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, 123, 101334.

Jara-Ettinger, J., Floyd, S., Huey, H., Tenenbaum, J. B., & Schulz, L. E. (2020b). Social pragmatics: Preschoolers rely on commonsense psychology to resolve referential underspecification. *Child development*, 91(4), 1135-1149.

Jara-Ettinger, J., & Rubio-Fernandez, P. (2021). Quantitative mental state attributions in language understanding. *Science advances*, 7(47), eabj0970.

- Jenkins, A. C., Dodell-Feder, D., Saxe, R., & Knobe, J. (2014). The neural bases of directed and spontaneous mental state attributions to group agents. *PLoS One*, 9(8), e105341.
- Jern, A., Lucas, C. G., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, 168, 46-64.
- Johnson, S. C., Dweck, C. S., & Chen, F. S. (2007). Evidence for infants' internal working models of attachment. *Psychological Science*, 18(6), 501-502.
- Johnson, S. C., Dweck, C. S., Chen, F. S., Stern, H. L., Ok, S. J., & Barth, M. (2010). At the intersection of social and cognitive development: Internal working models of attachment in infancy. *Cognitive science*, 34(5), 807-825.
- Kalish, C. W., & Lawson, C. A. (2008). Development of social category representations: Early appreciation of roles and deontic relations. *Child Development*, 79(3), 577-593.
- Kampis, D., & Southgate, V. (2020). Altercentric cognition: how others influence our cognitive processing. *Trends in Cognitive Sciences*, 24(11), 945-959.
- Kallgren, C. A., Reno, R. R., & Cialdini, R. B. (2000). A focus theory of normative conduct: When norms do and do not affect behavior. *Personality and social psychology bulletin*, 26(8), 1002-1012.
- Kaufmann, L., & Clément, F. (2014). Wired for society: Cognizing pathways to society and culture. *Topoi*, 33, 459-475.
- Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual review of psychology*, 31(1), 457-501.
- Krupenye, C., & Call, J. (2019). Theory of mind in animals: Current and future directions. *Wiley Interdisciplinary Reviews: Cognitive Science*, 10(6), e1503.
- Lapinski, M. K., & Rimal, R. N. (2005). An explication of social norms. *Communication theory*, 15(2), 127-147.

- Lau, T., Pouncy, H. T., Gershman, S. J., & Cikara, M. (2018). Discovering social groups via latent structure learning. *Journal of Experimental Psychology: General*, 147(12), 1881.
- Le Faye (2002). *Jane Austen The World of Her Novels*. HARRY N. ABRAMS, INC. New York.
- Lorenz, K. (2021). *On aggression*. Routledge.
- Liszkowski, U., Carpenter, M., & Tomasello, M. (2008). Twelve-month-olds communicate helpfully and appropriately for knowledgeable and ignorant partners. *Cognition*, 108(3), 732-739.
- 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.
- Lucas, C. G., Griffiths, T. L., Xu, F., Fawcett, C., Gopnik, A., Kushnir, T., ... & Hu, J. (2014). The child as econometrician: A rational model of preference understanding in children. *PloS one*, 9(3), e92160.
- Maibom, H. L. (2007). Social Systems. *Philosophical Psychology*, 20(5), 557–578. <https://doi.org/10.1080/09515080701545981>
- Mandler, J. M. (2014). *Stories, scripts, and scenes: Aspects of schema theory*. Psychology Press.
- Margoni, F., Baillargeon, R., & Surian, L. (2018). Infants distinguish between leaders and bullies. *Proceedings of the National Academy of Sciences*, 115(38), E8835-E8843.
- Martin, A., & Santos, L. R. (2016). What cognitive representations support primate theory of mind?. *Trends in cognitive sciences*, 20(5), 375-382.
- Mascaro, O., Goupil, N., Pantecouteau, H., Depierreux, A., Van der Henst, J. B., & Claidière, N. (2023). Human and animal dominance hierarchies show a pyramidal structure guiding adult and infant social inferences. *Nature Human Behaviour*, 7(8), 1294-1306.
- Mascaro, O., & Csibra, G. (2012). Representation of stable social dominance relations by human infants. *Proceedings of the National Academy of Sciences*, 109(18), 6862-6867.

- Mascaro, O., Goupil, N., Pantecouteau, H., Depierreux, A., Van der Henst, J. B., & Claidière, N. (2023). Human and animal dominance hierarchies show a pyramidal structure guiding adult and infant social inferences. *Nature Human Behaviour*, 7(8), 1294-1306.
- McCready, E. (2019). The semantics and pragmatics of honorification: Register and social meaning (Vol. 11). Oxford University Press, USA.
- McGeer, V. (2021). Enculturating folk psychologists. *Synthese*, 199(1), 1039-1063.
- McGrew, W. C. (1972). An ethological study of children's behavior.
- McMahon, E., & Isik, L. (2023). Seeing social interactions. *Trends in Cognitive Sciences*.
- Mi, Q., Wang, C., Camerer, C. F., & Zhu, L. (2021). Reading between the lines: Listener's vmPFC simulates speaker cooperative choices in communication games. *Science Advances*, 7(10), eabe6276.
- Moll, H., & Tomasello, M. (2007). Cooperation and human cognition: the Vygotskian intelligence hypothesis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 362(1480), 639-648.
- Mooney SJ, Filice DCS, Douglas NR, Holmes MM (2015) Task specialization and task-switching in eusocial mammals. *Animal Behavior* 109:227–233
- Nguyen, C. T. (2020). *Games: Agency as art*. Oxford University Press, USA.
- Noyes, A., Keil, F. C., & Dunham, Y. (2018). The emerging causal understanding of institutional objects. *Cognition*, 170, 83–87. <https://doi.org/10.1016/j.cognition.2017.09.008>
- Noyes, A., Dunham, Y., & Keil, F. C. (2023). Subjectivity and social constitution: Contrasting conceptions of institutions, artifacts, and animals. *Developmental Psychology*, 59(2), 377–389. <https://doi.org/10.1037/dev0001499>
- Noyes, A., Dunham, Y., Keil, F. C., & Ritchie, K. (2021). Evidence for multiple sources of inductive potential: Occupations and their relations to social institutions. *Cognitive Psychology*, 130, 101422. <https://doi.org/10.1016/j.cogpsych.2021.101422>

- Noyes, A., & Dunham, Y. (2020a). Groups as institutions: The use of constitutive rules to attribute group membership. *Cognition*, 196, 104143. <https://doi.org/10.1016/j.cognition.2019.104143>
- Noyes, A., & Keil, F. C. (2020b). There is no privileged link between kinds and essences early in development. *Proceedings of the National Academy of Sciences*, 117(20), 10633–10635. <https://doi.org/10.1073/pnas.2003627117>
- Noyes, A., Keil, F. C., & Dunham, Y. (2020c). Institutional actors: Children's emerging beliefs about the causal structure of social roles. *Developmental Psychology*, 56(1), 70–80. <https://doi.org/10.1037/dev0000847>
- O'Connor, C. (2019). *The origins of unfairness: Social categories and cultural evolution*. Oxford University Press, USA.
- O'Laughlin, M. J., & Malle, B. F. (2002). How people explain actions performed by groups and individuals. *Journal of personality and social psychology*, 82(1), 33.
- Onishi, K. H., & Baillargeon, R. (2005). Do 15-month-old infants understand false beliefs?. *science*, 308(5719), 255-258.
- Partington, S., Nichols, S., & Kushnir, T. (2023). Rational learners and parochial norms. *Cognition*, 233, 105366.
- Paunov, A. M., Blank, I. A., & Fedorenko, E. (2019). Functionally distinct language and Theory of Mind networks are synchronized at rest and during language comprehension. *Journal of neurophysiology*, 121(4), 1244-1265.
- Paz-y-Miño C, G., Bond, A. B., Kamil, A. C., & Balda, R. P. (2004). Pinyon jays use transitive inference to predict social dominance. *Nature*, 430(7001), 778-781.
- Pearl, J. (2009). *Causality*. Cambridge university press.
- Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences*, 25(2), 100-110.

- Pitcher, D. (2021). Characterizing the third visual pathway for social perception. *Trends in Cognitive Sciences*, 25(7), 550-551.
- Powell, L. J., & Spelke, E. S. (2013). Preverbal infants expect members of social groups to act alike. *Proceedings of the National Academy of Sciences*, 110(41), E3965-E3972.
- Prentice, D. A., & Miller, D. T. (2007). Psychological Essentialism of Human Categories. *Current Directions in Psychological Science*, 16(4), 202–206. <https://doi.org/10.1111/j.1467-8721.2007.00504.x>
- Rakoczy, H., Warneken, F., & Tomasello, M. (2008). The sources of normativity: young children's awareness of the normative structure of games. *Developmental psychology*, 44(3), 875.
- Rakoczy, H., & Schmidt, M. F. (2013). The early ontogeny of social norms. *Child Development Perspectives*, 7(1), 17-21.
- Rehbein, B., & Söyler, T. (2020). The social ontology of caste. In *Social Ontology, Sociocultures, and Inequality in the Global South* (pp. 85-99). Routledge.
- Repacholi, B. M., & Gopnik, A. (1997). Early reasoning about desires: evidence from 14-and 18-month-olds. *Developmental psychology*, 33(1), 12.
- Rhodes, M., & Baron, A. (2019). The Development of Social Categorization. *Annual Review of Developmental Psychology*, 1(1), 359–386.
- Rhodes, M., & Mandalaywala, T. M. (2017). The development and developmental consequences of social essentialism. *Wiley Interdisciplinary Reviews: Cognitive Science*, 8(4), e1437-18.
- Ritchie, K. (2020a). Social Structures and the Ontology of Social Groups. *Philosophy and Phenomenological Research*, 100(2), 402–424. <https://doi.org/10.1111/phpr.12555>
- Ritchie, K. (2020b). Minimal cooperation and group roles. *Minimal Cooperation and Shared Agency*, 93-109.

- Royka, A., & Santos, L. R. (2022). Theory of Mind in the Wild. *Current Opinion in Behavioral Sciences*, 45, 101137.
- Rubio-Fernandez, P., Berke, M. D., & Jara-Ettinger, J. (2024). Tracking minds in communication. *Trends in Cognitive Sciences*.
- Rubio-Fernández, P., Mollica, F., Ali, M. O., & Gibson, E. (2019). How do you know that? Automatic belief inferences in passing conversation. *Cognition*, 193, 104011.
- Ruffman, T., Taumoepeau, M., & Perkins, C. (2012). Statistical learning as a basis for social understanding in children. *British Journal of Developmental Psychology*, 30(1), 87-104.
- Saxe, R. (2006). Uniquely human social cognition. *Current opinion in neurobiology*, 16(2), 235-239.
- Saxe, R., & Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in “theory of mind”. *Neuroimage*, 19(4), 1835-1842.
- Schmidt, M. F., Rakoczy, H., & Tomasello, M. (2012). Young children enforce social norms selectively depending on the violator’s group affiliation. *Cognition*, 124(3), 325-333.
- Schmidt, M. F., & Rakoczy, H. (2023). Children's Acquisition and Application of Norms. *Annual Review of Developmental Psychology*, 5.
- Schöppner, B., Sodian, B., & Pauen, S. (2006). Encoding action roles in meaningful social interaction in the first year of life. *Infancy*, 9(3), 289-311
- Searle, J. R. (2006). Social ontology: Some basic principles. *Anthropological theory*, 6(1), 12-29.
- Searle, J. R. (2005). What is an institution? *Journal of Institutional Economics*, 1(1), 1–22.
<https://doi.org/10.1017/S1744137405000020>
- Seeley, T. D. (1997). Honey bee colonies are group-level adaptive units. *The American Naturalist*, 150(S1), s22-S41.

- Seiver, E., Gopnik, A., & Goodman, N. D. (2013). Did she jump because she was the big sister or because the trampoline was safe? Causal inference and the development of social attribution. *Child development*, 84(2), 443-454.
- Schank, R. C., & Abelson, R. P. (2013). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.
- Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in cognitive sciences*, 4(8), 299-309.
- Schmidt, M. F., Butler, L. P., Heinz, J., & Tomasello, M. (2016). Young children see a single action and infer a social norm: Promiscuous normativity in 3-year-olds. *Psychological Science*, 27(10), 1360-1370.
- Shafto, P., Goodman, N. D., & Griffiths, T. L. (2014). A rational account of pedagogical reasoning: Teaching by, and learning from, examples. *Cognitive psychology*, 71, 55-89.
- Shutts, K., & Kalish, C. W. (2021). Intuitive sociology. In *Advances in Child Development and Behavior* (Vol. 61, pp. 335–374). Elsevier. <https://doi.org/10.1016/bs.acdb.2021.05.004>
- Sloane, S., Baillargeon, R., & Premack, D. (2012). Do infants have a sense of fairness?. *Psychological science*, 23(2), 196-204.
- Sluckin A. M., Smith P. K., Two approaches to the concept of dominance in preschool children. *Child Dev.* 48, 917 (1977).
- Sodian, B. (2011). Theory of mind in infancy. *Child Development Perspectives*, 5(1), 39-43.
- Sperber, D., & Wilson, D. (1986). *Relevance: Communication and cognition* (Vol. 142). Cambridge, MA: Harvard University Press.
- Stander, P. E. (1992). Cooperative hunting in lions: the role of the individual. *Behavioral ecology and sociobiology*, 29, 445-454.
- Spokes, A. C., & Spelke, E. S. (2017). The cradle of social knowledge: Infants' reasoning about caregiving and affiliation. *Cognition*, 159, 102-116.

- Stavans, M., & Baillargeon, R. (2019). Infants expect leaders to right wrongs. *Proceedings of the National Academy of Sciences*, 116(33), 16292-16301.
- Sumpter, D. J. (2010). *Collective animal behavior*. In *Collective animal behavior*. Princeton University Press.
- Taggart, J., Becker, I., Rauen, J., Al Kallas, H., & Lillard, A. S. (2020). What shall we do: Pretend or real? Preschoolers' choices and parents' perceptions. *Journal of Cognition and Development*, 21(2), 261-281.
- Taggart, J., Heise, M. J., & Lillard, A. S. (2018). The real thing: Preschoolers prefer actual activities to pretend ones. *Developmental Science*, 21(3), e12582.
- Tang, S., & Gray, K. (2023). Do companies think and feel? Mind perception of organizations. *Cognitive Science*, 47(8), e13320.
- Tatone, D., Geraci, A., & Csibra, G. (2015). Giving and taking: Representational building blocks of active resource-transfer events in human infants. *Cognition*, 137, 47-62.
- Tatone, D., Hernik, M., & Csibra, G. (2019). Minimal cues of possession transfer compel infants to ascribe the goal of giving. *Open Mind*, 3, 31-40.
- Taylor, S. E., & Crocker, J. (2022). Schematic bases of social information processing. In *Social cognition* (pp. 89-134). Routledge.
- Teraji, S. (2018). *The Cognitive Basis of Institutions*. Elsevier.
- Theriault, J. E., Young, L., & Barrett, L. F. (2021). The sense of should: A biologically-based framework for modeling social pressure. *Physics of Life Reviews*, 36, 100-136.
- Thomas A.J. (in press). Cognitive Representations of Social Relationships and their Developmental Origins. *Behavioral and Brain Sciences*. Published online 2024:1-53. doi:10.1017/S0140525X24001328

- Thomas, A. J., Thomsen, L., Lukowski, A. F., Abramyan, M., & Sarnecka, B. W. (2018). Toddlers prefer those who win but not when they win by force. *Nature Human Behaviour*, 2(9), 662-669.
- Thomas, A. J., Saxe, R., & Spelke, E. S. (2022). Infants infer potential social partners by observing the interactions of their parent with unknown others. *Proceedings of the National Academy of Sciences*, 119(32), e2121390119.
- Thomas, A. J., Woo, B., Nettle, D., Spelke, E., & Saxe, R. (2022). Early concepts of intimacy: Young humans use saliva sharing to infer close relationships. *Science*, 375(6578), 311-315.
- Thomsen, L., Frankenhuys, W. E., Ingold-Smith, M., & Carey, S. (2011). Big and mighty: Preverbal infants mentally represent social dominance. *Science*, 331(6016), 477-480.
- Tibbetts, E. A., & Lindsay, R. (2008). Visual signals of status and rival assessment in *Polistes dominulus* paper wasps. *Biology Letters*, 4(3), 237-239.
- Tomasello, M. (2020). The role of roles in uniquely human cognition and sociality. *Journal for the Theory of Social Behaviour*, 50(1), 2-19.
- Tomasello, M. (2022). *The Evolution of Agency: Behavioral Organization from Lizards to Humans*. MIT Press.
- 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(05). <https://doi.org/10.1017/S0140525X05000129>
- Tomasello, M., & Carpenter, M. (2007). Shared intentionality. *Developmental science*, 10(1), 121-125.
- Tomasello, M., & Herrmann, E. (2010). Ape and human cognition: What's the difference?. *Current Directions in Psychological Science*, 19(1), 3-8.
- Tomasello, M., & Moll, H. (2010). The gap is social: Human shared intentionality and culture. *Mind the gap: Tracing the origins of human universals*, 331-349.

- Ullman, T., Baker, C., Macindoe, O., Evans, O., Goodman, N., & Tenenbaum, J. (2009). Help or hinder: Bayesian models of social goal inference. *Advances in neural information processing systems*, 22.
- Ullman, T. D., & Bass, I. (2024). The detection of automatic behavior in other people. *American Psychologist*, 79(9), 1322.
- Usher, M., & Donnelly, N. (1998). Visual synchrony affects binding and segmentation in perception. *Nature*, 394(6689), 179-182.
- Vaish, A., Demir, Ö. E., & Baldwin, D. (2011). Thirteen-and 18-month-old infants recognize when they need referential information. *Social Development*, 20(3), 431-449.
- Vasilyeva, N., Gopnik, A., & Lombrozo, T. (2018). The Development of Structural Thinking About Social Categories. *Developmental Psychology*, 10.
- Vasilyeva, N., & Lombrozo, T. (2020). Structural thinking about social categories: Evidence from formal explanations, generics, and generalization. *Cognition*, 204, 104383. <https://doi.org/10.1016/j.cognition.2020.104383>
- Wang, R. E., Wu, S. A., Evans, J. A., Tenenbaum, J. B., Parkes, D. C., & Kleiman-Weiner, M. (2020). Too many cooks: Coordinating multi-agent collaboration through inverse planning.
- Weber, M. (2004). *The vocation lectures*. Hackett Publishing.
- Wellman, H. M. (2014). *Making minds: How theory of mind develops*. Oxford University Press.
- Westra, E. (2018). Character and theory of mind: An integrative approach. *Philosophical Studies*, 175(5), 1217-1241.
- Westra, E. (2020). Getting to know you: Accuracy and error in judgments of character. *Mind & Language*, 35(5), 583-600.
- Whittington, E. M., & Bradley, D. E. (2001). The sport of life and death: The Mesoamerican ballgame.

- Wilterson, A. I., & Graziano, M. S. (2021). The attention schema theory in a neural network agent: Controlling visuospatial attention using a descriptive model of attention. *Proceedings of the National Academy of Sciences*, 118(33), e2102421118.
- Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1), 1-34.
- Yang, X., Naas, R., & Dunham, Y. (2021). Testing the limits of structural thinking about gender. *Developmental Science*. <https://doi.org/10.1111/desc.13169>
- Young, L., Cushman, F., Hauser, M., & Saxe, R. (2007). The neural basis of the interaction between theory of mind and moral judgment. *Proceedings of the National Academy of Sciences*, 104(20), 8235-8240.
- Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., & Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, 107(15), 6753-6758.
- Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. MIT Press.

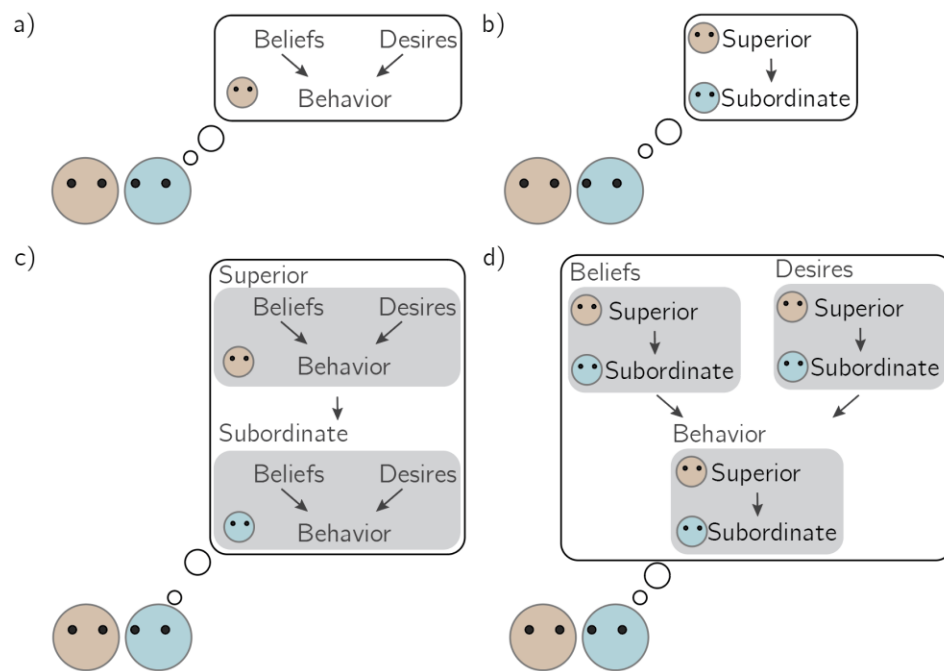


Figure 1. Schematic of an agent applying (a) a mentalistic stance (Section 2), (b) a simple institutional stance (Sections 3-4), (c) an institutional stance with mentalistic agents (Section 6.3), (d) a mentalistic stance to characterize and understand an institutional representation (Section 6.4).

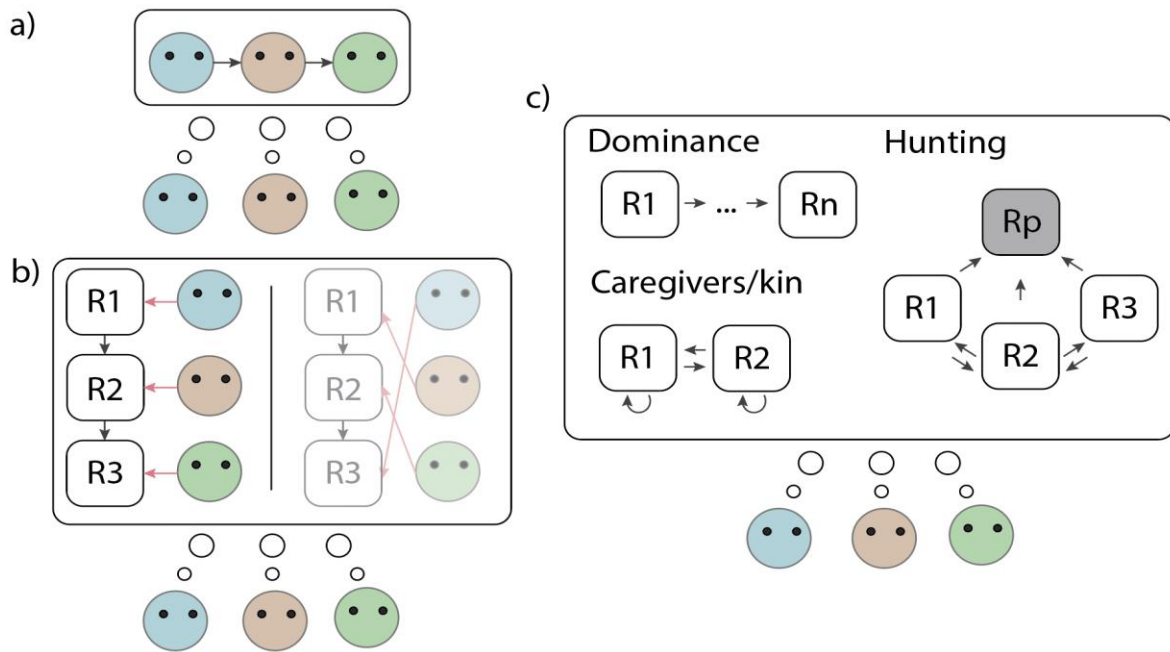


Figure 2. (a) A fixed proto-institutional structure where a group of agents act in coordination, showing here a dominance hierarchy. This type of implementation does not require users to differentiate roles from role-occupiers, or even to represent positions in the structure *as* roles at all. (b) Role-based structure where a group has an abstract representation of the social structure, and a mapping from agents to roles. The ability to represent social roles as detached from individuals enables members to consider alternate arrangements such as role switches. (c) Group with multiple role-based representations to structure a more complex social life. Here, the dominance representation is a linear structure that can be expanded or compressed based on group size; the caregiver/kin structure represents a system which includes expectations for how agents in the same role act towards one another (e.g. mutual expectations among siblings or among caregivers) as well as how agents in different roles act towards one another (e.g. expectations of care and deference). The hunting structure represents the prey (Rp) as a role in the system, despite the prey not being aware of this, as well as relationships between different hunter roles. As social complexity increases, conspecifics must represent a larger number of role-based systems, and consider how they might combine productively (e.g., how caregiver/kin relations affect conflict in or have critical entailments within dominance hierarchies). The increased number of institutional structures, and the need to reason about them compositionally, may have led to the generative institutional grammar in humans, which enables us to rapidly synthesize and infer institutional structures.



Figure 3. Two identical agents taking identical actions in the same context. The single presence of an apron invites observers to take an institutional stance that triggers different inferences, predictions, and expectations. Rather than reading the behavior as revealing an agent’s desire for olive oil, we are more likely to construe it as revealing a role obligation, such as restocking or checking the sell-by date..