Behavioral and Brain Sciences

cambridge.org/bbs

The institutional stance

Julian Jara-Ettinger and Yarrow Dunham

Department of Psychology, Yale University, New Haven, CT 06511, USA
julian.jara-ettinger@yale.edu
yarrow.dunham@yale.edu

Target Article

Cite this article: Jara-Ettinger J and Dunham Y. (2026) The institutional stance. *Behavioral and Brain Sciences* **49**, e367: 1–63. doi:10.1017/S0140525X25103427

Target Article Accepted: 30 September 2025
Target Article Manuscript Online: 20 November 2025
Commentaries Accepted: 17 April 2025

Keywords:
social cognition; coordination; institutional stance; theory of mind; roles; scripts

What is Open Peer Commentary? What follows on these pages is known as a Treatment, in which a significant and controversial Target Article is published along with Commentaries (p. 19) and an Authors' Response (p. 55). See bbsonline.org for more information.

Corresponding author: Julian Jara-Ettinger; Email: julian.jara-ettinger@yale.edu

Short Abstract

The power of human social cognition is typically associated with mental state reasoning—a mentalistic stance. Here, we argue humans have 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 reasoning 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.

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 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, Berke, & Jara-Ettinger, 2024; 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.

© The Author(s), 2025. Published by Cambridge University Press

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 allows 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 on the idea that role-based systems are central to understanding complex social behavior. This view spans philosophy (e.g., Maibom, 2007; Ritchie, 2020a), computer science (Schank & Abelson, 1977), developmental psychology (Clément, Bernard, & Kaufmann, 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 in social collectives (e.g., Eickers, 2024; Maibom, 2007; Noyes & Dunham, 2020; Noyes et al., 2021; Ritchie, 2020a). 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 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 in the institutional stance will often be considerably simpler than, and in some cases divergent from, the actual social structures populating the world. Adopting the institutional stance is useful so long as its underlying representations succeed in 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

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

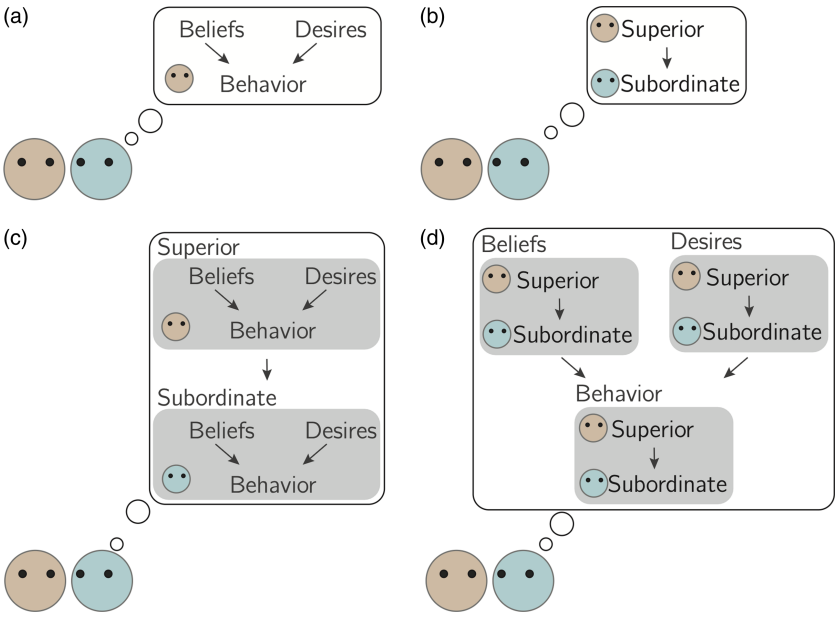


Figure 1. Schematic of an agent applying (a) a mentalistic stance (Section 2), (b) a simple institutional stance (Sections 3–4), (c) an institutional stance with mentalistic agents (Section 6.3), (d) a mentalistic stance to characterize and understand an institutional representation (Section 6.4).

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 homeowner 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 could not 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, Carpenter, & Tomasello 2008) that rationally cause behavior (Gergely & Csibra, 2003; Liu et al., 2017; Woodward, 1998). This understanding structures early childhood cognition, including language understanding (Akhtar, Carpenter, & Tomasello, 1996; Vaish, Demir, & Baldwin, 2011; Rubio-Fernandez, Berke, & Jara-Ettinger, 2024), causal learning (Gergely, Bekkering, & Király, 2002; Bonawitz et al., 2011), socio-moral evaluations (Hamlin et al., 2013; Sloane, Baillargeon, & Premack, 2012), and epistemic vigilance (Harris, 2012; Aboody, Huey, & Jara-Ettinger, 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, Lucas, & Kemp 2017), predict behavior (Jara-Ettinger, Schulz, & Tenenbaum 2020), cooperate (Wang et al., 2020), make social evaluations (Ullman et al., 2009), share knowledge (Shafto,

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

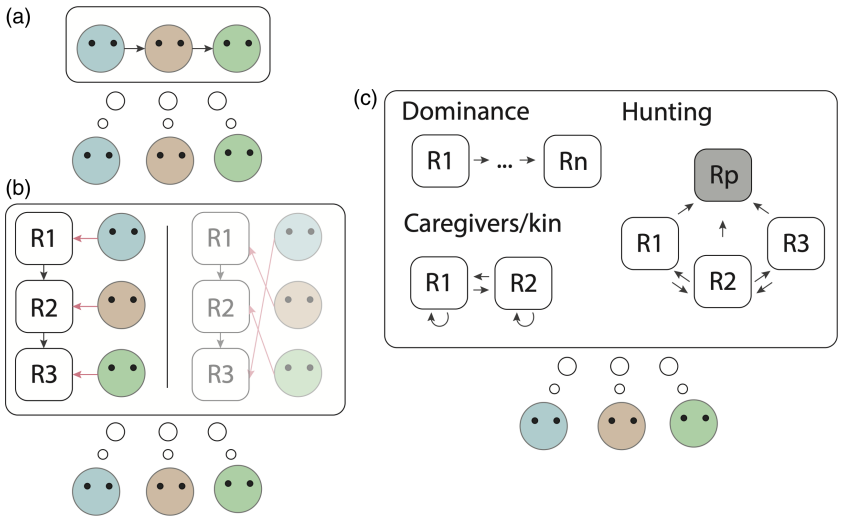


Figure 2. (a) A fixed proto-institutional structure where a group of agents act in coordination, showing here a dominance hierarchy. This type of implementation does not require users to differentiate roles from role-occupiers, or even to represent positions in the structure *as* roles at all. (b) Role-based 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.

Goodman, & Griffiths 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, Blank, & Fedorenko 2019) and moral reasoning (Young et al., 2007, 2010); 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, McCarthy, & Scholl, 2010; McMahon & Isik, 2023), automatically triggering richer inferences about other people's mental states (Rubio-Fernández et al., 2019; Kampis & 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; Berke et al., 2025). 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 determine 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

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

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's 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 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, Rossiter, & Faulkes 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 (Figure 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

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

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 a pressure to reason about how different social systems 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 interconnected 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 on 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 standalone 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 5 pm. 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 that 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 laboratory, 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.

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

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. At 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, Hamrick, & Tenenbaum 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 into social ontology and theories of institutions may be no exception. This suggests that theories of formal 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 for 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, Bernard, & Kaufmann 2011; Ritchie, 2020a; 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, 2020), via social agreement (Noyes & Dunham, 2017) and formal membership criteria (Noyes & Dunham, 2020). 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, Keil, & Dunham, 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, 2020a). 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, Gopnik, & Lombrozo 2018; Yang, Naas, & Dunham 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,

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

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, Young, & Barrett 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, Spaulding, & Westra 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, Reno, & Cialdini 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

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

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, 2024) 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, Gweon, & Sommerville 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, Baillargeon, & Surian 2018; Thomas et al., 2018), and they make transitive inferences from pairwise dominance displays (Gazes, Hampton, & Lourenco 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, Saxe, & Spelke 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, Sodian, & Pauen 2006; Tatone, Geraci, & Csibra 2015). Similarly, infants readily represent object transfer events in role-like terms: an actor *giving* an object to a recipient (Tatone, Hernik, & Csibra 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, Bernard, & Kaufmann 2011), 3-year-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-year-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., 2009) 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, Jara-Ettinger, & Dunham 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 who are securely attached (Ainsworth, 1979) have stronger abstractly coded expectations that a caregiver will comfort a distressed agent (Johnson, Dweck, & Chen 2007; Johnson et al., 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

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

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, Rakoczy, & Tomasello 2012), which they can infer purely through observation (Partington, Nichols, & Kushnir 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 Nichols, & Kushnir 2023; Schmidt et al., 2016; Rakoczy, Warneken, & Tomasello 2008; Rakoczy & Schmidt, 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, 1979, 1993), and children engage in these activities in part because they are unable to actually do them (Taggart, Heise, & Lillard 2018; Taggart et al., 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, Gopnik, & Goodman 2013; Ruffman, Taumoepeau, & Perkins 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, Jara-Ettinger, & Dunham 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, Chater, & Tenenbaum, 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 run 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 untoward 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, Dunham, & Jara-Ettinger 2024). This

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press



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.

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 of 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; 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, Moll, & Tomasello 2010; Jara-Ettinger et al., 2020), 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

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

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., 2016), 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 mind-reading 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 (Figure 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 that 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 (Figure 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.

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

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 is 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 (Ásta, 2018; Searle, 2005, 2006; Hindriks & Guala, 2015) and more recently by cognitive scientists (Noyes, Keil, & Dunham, 2018; Noyes & Dunham, 2020), 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 be involved in regulating the manufacture of hammers, the allowable forms of hammering, or who can perform the role of carpenter. But whatever those processes specify, we would still consider objects designed to deliver force to be hammers, 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, Keil, & Dunham 2018; Noyes, Keil, & Dunham, 2020; Noyes, Dunham, & Keil, 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 that 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

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

behavior is a dynamic reactive system. Whatever system a social group uses to understand each other's 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 literature 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, Harrington, & Jackson 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 eighteenth- and nineteenth-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

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

representations – agency, goals, and basic social relations (Abassi & Papeo, 2020; Isik et al., 2017; Papeo, Stein, & Soto-Faraco, 2017; Scholl & Tremoulet, 2000) – which can then be interpreted either through mentalizing or through the institutional stance; Figure 1). 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 & Scholl, 2011; 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 that 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.

Financial support. This work was supported by NSF awards BCS-2045778 and IIS-2106690.

Conflict of interests. The authors declare no conflicts of interest.

Note

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).

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.

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

Akhtar, N., Carpenter, M., & Tomasello, M. (1996). The role of discourse novelty in early word learning. *Child Development*, *67*(2), 635–645.

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*, *18*(2), 259–274. https://doi.org/10.1177/17456916221093593 (Original work published 2023).

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*. Oxford University Press.

Baker, A., Dunham, Y., & Jara-Ettinger, J. (2024). Roles guide rapid inferences about agent knowledge and behavior. In L. K. Samuelson, S. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th annual conference of the cognitive science society* (p. 6180). Austin, TX: Cognitive Science Society.

Baker, A., Jara-Ettinger, J., & Dunham, Y. (under review). Roles trigger agent-general normative explanations for agent behavior beginning in middle childhood.

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.

Baron-Cohen, S. (2000). Theory of mind and autism: A review. *International Review of Research in Mental Retardation*, *23*, 169–184.

Barrett, H. C., Bolyanatz, A., Crittenden, A. N., Fessler, D. M., Fitzpatrick, S., Gurven, M., Henrich, J., Kanovsky, M., Kushnick, G., Pisor, A., Scelza, B. A., Stich, S., von Rueden, C., Zhao, W., & Laurence, S. (2016). Small-Scale societies exhibit fundamental variation in the role of intentions in moral judgment. *Proceedings of the National Academy of Sciences*, *113*(17), 4688–4693.

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, F. (1932). *Remembering: A study in experimental and social psychology*. 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.

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.

Berke, M. D., Horschler, D. J., Royka, A., Santos, L. R., & Jara-Ettinger, J. (2025). What primates know about other minds and when they use it: A computational approach to comparative theory of mind. bioRxiv. https://doi.org/10.1101/2023.08.02.551487

Blasi, D. E., Henrich, J., Adamou, E., Kemmerer, D., & Majid, A. (2022). Over-reliance on English hinders cognitive science. *Trends in Cognitive Sciences*, *26*(12), 1153–1170.

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. In J. Fishman (Ed.), *Readings in the Sociology of Language* (pp. 252–275). De Gruyter Brill.

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*, *52*(1), 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.

D'Ettorre, P., & Heinze, J. (2001). Sociobiology of slave-making ants. *Acta Ethologica*, *3*, 67–82.

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*. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the annual conference of the cognitive science society* (p. 3720). Austin, TX: Cognitive Science Society.

Davis, I., Jara-Ettinger, J., & Dunham, Y., (2022). Inferring the internal structure of social collectives. In J. Culbertson, A. Perfors, H. Rabagliati, & V. Ramenzoni (Eds.), *Proceedings of the annual conference of the cognitive science society* (p. 1395). Austin, TX: Cognitive Science Society.

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, July 3). Jugements de valeur et jugements de réalité. In *Atti del IV Congresso Internazionale di Filosofia* (Vol. 1, pp. 99–114). Bologne, Italy.

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*. Routledge.

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. In R. Buffenstein, T. Park & M. Holmes (Eds), *The Extraordinary Biology of the Naked Mole-Rat* (pp. 1–33). Springer Nature Link.

Le Faye (2002). *Jane Austen: The world of her novels*. Harry N. Abrams, Inc.

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*, *55*(5), 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.

Gao, T., & Scholl, B. J. (2011). Chasing vs. stalking: Interrupting the perception of animacy. *Journal of Experimental Psychology: Human Perception and Performance*, *37*(3), 669.

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., Bekkering, H., & Király, I. (2002). Rational imitation in preverbal infants. *Nature*, *415*(6873), 755–755.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naıve theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Gershman, S. J., & Cikara, M. (2020). Social-structure learning. *Current Directions in Psychological Science*, *29*(5), 460–466.

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.

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

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.

Graham, J., Waytz, A., Meindl, P., Iyer, R., & Young, L. (2017). Centripetal and centrifugal forces in the moral circle: Competing constraints on moral learning. *Cognition*, *167*, 58–65.

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University 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. In D. Sperber, D. Premack, & A. J. Premack (Eds.), *Causal cognition: A multidisciplinary debate* (pp. 351–394). Clarendon Press/Oxford University Press.
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, 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.
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
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. In K. Jones, L. Schroeter, & F. Schroeter (Eds.) *Constructing Social Hierarchy*. 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*, *43*(1), 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. In M. R. Banaji, & S. A. Gelman (Eds.), *Navigating the social world: What infants, children, and other species can teach us* (pp. 101–106). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199890712.003.0019
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*. Springer-Verlag.
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., & Rubio-Fernandez, P. (2021). Quantitative mental state attributions in language understanding. *Science Advances*, *7*(47), eabj0970.
Jara-Ettinger, J., Schulz, L. E., & Tenenbaum, J. B. (2020). The naive utility calculus as a unified, quantitative framework for action understanding. *Cognitive Psychology*, *123*, 101334.
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.
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.
Kampis, D., & Southgate, V. (2020). Altercentric cognition: How others influence our cognitive processing. *Trends in Cognitive Sciences*, *24*(11), 945–959.
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.
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.
Lorenz, K. (2021). *On aggression*. Routledge.
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., & 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.
McGeer, V. (2021). Enculturating folk psychologists. *Synthese*, *199*(1), 1039–1063.

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

McGrew, W. C. (1972). *An ethological study of children's behavior*. New York: Academic Press.
McMahon, E., & Isik, L. (2023). Seeing social interactions. *Trends in Cognitive Sciences*, *27*(12), 1165–1179.
McNamara, R. A., Willard, A. K., Norenzayan, A., & Henrich, J. (2019). Weighing outcome vs. intent across societies: How cultural models of mind shape moral reasoning. *Cognition*, *182*, 95–108.
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, S. J., Filice, D. C. S., Douglas, N. R., & Holmes, M. M. (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.
Noyes, A., & Dunham, Y. (2017). Mutual intentions as a causal framework for social groups. *Cognition*, *162*, 133–142.
Noyes, A., & Dunham, Y. (2020). 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., 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., & Keil, F. C. (2020). 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. (2018). The emerging causal understanding of institutional objects. *Cognition*, *170*, 83–87. https://doi.org/10.1016/j.cognition.2017.09.008
Noyes, A., Keil, F. C., & Dunham, Y. (2020). 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.
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.
Papeo, L., Stein, T., & Soto-Faraco, S. (2017). The two-body inversion effect. *Psychological Science*, *28*(3), 369–379.
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. (2021). Characterizing the third visual pathway for social perception. *Trends in Cognitive Sciences*, *25*(7), 550–551.
Pitcher, D., & Ungerleider, L. G. (2021). Evidence for a third visual pathway specialized for social perception. *Trends in Cognitive Sciences*, *25*(2), 100–110.
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., & Schmidt, M. F. (2013). The early ontogeny of social norms. *Child Development Perspectives*, *7*(1), 17–21.
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.
Rehbein, B., & Söyler, T. (2020). The social ontology of caste. In B. Baumann, & D. Bultmann (Eds.), *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. In A. Fiebich (Eds.), *Minimal cooperation and shared agency* (pp. 93–109). Springer Nature Link.
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*, *29*(3), 269–281.
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.
Schank, R., & Abelson, R. (1977). *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*. Erlbaum.
Schank, R. C., & Abelson, R. P. (2013). *Scripts, plans, goals, and understanding: An inquiry into human knowledge structures*. Psychology Press.
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.
Schmidt, M. F., & Rakoczy, H. (2023). Children's acquisition and application of norms. *Annual Review of Developmental Psychology*, *5*, 193–215.
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.
Scholl, B. J., & Tremoulet, P. D. (2000). Perceptual causality and animacy. *Trends in Cognitive Sciences*, *4*(8), 299–309.
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. (2005). What is an institution? *Journal of Institutional Economics*, *1*(1), 1–22. https://doi.org/10.1017/S1744137405000020
Searle, J. R. (2006). Social ontology: Some basic principles. *Anthropological Theory*, *6*(1), 12–29.
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.
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.

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

Sluckin A. M., & Smith P. K. (1977). Two approaches to the concept of dominance in preschool children. *Child Development*, *48*, 917.
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). Harvard University Press.
Spokes, A. C., & Spelke, E. S. (2017). The cradle of social knowledge: Infants' reasoning about caregiving and affiliation. *Cognition*, *159*, 102–116.
Stander, P. E. (1992). Cooperative hunting in lions: the role of the individual. *Behavioral Ecology and Sociobiology*, *29*, 445–454.
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. (2024). Cognitive representations of social relationships and their developmental origins. *Behavioral and Brain Sciences*, 1–53. https://doi.org/10.1017/S0140525X24001328
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., 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., 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., Frankenhuis, 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. (2007). Shared intentionality. *Developmental Science*, *10*(1), 121–125.
Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, *28*(5), 675–691. https://doi.org/10.1017/S0140525X05000129
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* (pp. 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*, *54*(9), 1735–1744.
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*. Mullen Books.
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*, *25*, e13169. https://doi.org/10.1111/desc.13169
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.
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.
Zawidzki, T. W. (2013). *Mindshaping: A new framework for understanding human social cognition*. MIT Press.

Open Peer Commentary

Institutions as adaptive regulators of social stability

Abby Jo Adams*

Western Michigan University, USA
abby.j.pastor@wmich.edu

*Corresponding author.

doi:10.1017/S0140525X26105494, e368

Abstract

Jara-Ettinger and Dunham propose that human social cognition relies on an institutional stance grounded in generative, role-structured representations. I extend this account by suggesting that institutional reasoning functions as a constraint-regulating

cognitive system, balancing stability and flexibility under changing social demands and helping explain how role-based coordination remains robust without producing rigidity, overload, or breakdown.

Jara-Ettinger and Dunham propose that human social cognition relies not only on mentalistic reasoning but also on an institutional stance that interprets behavior through role-structured systems of constraint and coordination. Their account provides a compelling framework for explaining how humans infer social structure, track role-based expectations, and generate novel institutional forms. By framing institutions as generative cognitive representations, the authors expand models of social reasoning beyond individual mental states toward structured collective dynamics. Here, I suggest that institutional representations may function as constraint-regulating cognitive systems that balance stability and flexibility during social coordination.

While the institutional stance is described as a generative cognitive system that structures social understanding through roles and rule-like expectations, less attention is given to how these representations remain adaptive as social environments change. Institutional reasoning must not only constrain behavior but also accommodate uncertainty, variability, and shifting contextual demands. Viewing institutional representations as constraint-regulating cognitive processes raises important questions about how stability is maintained without producing rigidity and how flexibility emerges without undermining predictability. If institutional representations function as generative grammars, as the authors propose, constraint-regulation may describe the dynamic mechanism that allows such grammars to remain viable under changing social load. Addressing these dynamics may clarify both the strengths and boundary conditions of institutional reasoning in real-world coordination.

One way to address these questions is to view institutional representations not only as generative models of social structure but also as constraint-regulating cognitive mechanisms that balance competing demands. Roles and rule-like expectations may help manage the tension between predictability and flexibility, enabling coordinated action without fully determining behavior. From this perspective, institutional reasoning does not simply encode static structures; it supports a dynamic equilibrium in which social coordination remains stable while retaining adaptive capacity. Framing institutional cognition in this way helps clarify how role-based systems persist, evolve, or break down when demands exceed regulatory limits.

Framing institutional representations as constraint-regulating cognitive processes also helps illuminate several phenomena that remain difficult to explain through structural description alone. Role-based expectations can support efficient coordination, yet excessive stabilization may produce rigidity, moral disengagement, or the dehumanization of role occupants when constraints become over-specified. Conversely, insufficient constraint can reduce predictability, leading to fragmentation or coordination failure. These dynamics suggest that individuals may shift between institutional and mentalistic interpretations not merely as alternative representational strategies but as adaptive responses to changing social uncertainty and cognitive load. This perspective highlights how the same cognitive mechanisms that enable large-scale coordination may also generate instability when regulatory balance is lost.

Taken together, this perspective complements Jara-Ettinger and Dunham's proposal by situating the institutional stance not only as a generative cognitive system for representing roles and social structure but also as a constraint-regulating mechanism that helps balance stability and adaptability during social coordination. Framing institutional reasoning in this way may help bridge cognitive, developmental, and social-level accounts of coordination, highlighting how role-based representations support both persistence and change in collective behavior. Rather than revising the authors' framework, this view extends it by clarifying how institutional cognition remains viable across changing contexts while preserving its generative power.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. The author declares no competing interests.

Is the institutional stance about prediction or normative regulation?

Christopher Joseph An*

School of Philosophy, Psychology and Language Sciences, The University of Edinburgh, UK
cjan.phil@gmail.com

*Corresponding author.

doi:10.1017/S0140525X26105718, e369

Abstract

I argue that, in light of grounding the institutional stance in role-based norms, this stance is not a causal or predictive model of social cognition. Instead, it is fundamentally normative in that it facilitates practices of holding responsible that mere predictive models would have difficulty explaining. The institutional stance, as such, facilitates robust ought-based claims and demands directed towards role-holders.

"[T]he café waiter cannot immediately be a café waiter in the sense in which this inkwell is an inkwell . . . " (Sartre, 2022).

The institutional stance is a welcome and important addition to theories of social cognition, providing a rich account of human social life insofar as it is embedded within rule-bound institutions. I argue, however, that the grammar supporting the institutional stance is not a causal one but a fundamentally normative or ought-based grammar scaffolded by deontic commitments of role-holders. This distinction is important because this normative grammar facilitates practices of holding responsible that mere predictive models would have trouble explaining (I also think that the mentalistic stance is also normative but I set this aside).

As an "intuitive theory" of social structures that "place individuals into roles and reason about their actions in light of beliefs about those roles," Jara-Ettinger and Dunham suggest that we

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

reason about other people through causal and predictive judgments. This is consistent with what philosophers would call an "objective stance" (Strawson, 1962).

By contrast, I argue that the institutional stance is continuous with reasoning in the "participant stance," where beliefs about roles are subject to "ought-based" responsibility judgments rather than "evidence-based" predictive judgments (Holton, 1994; Marušić, 2017). This is because institution-based roles, rather than merely about people prediction or social coordination, are likely more about identifying and tracking people's duties and obligations associated with role-based statuses and commitments, and also reasoning about role-holders' actions precisely by holding them responsible for those obligations (Searle, 2010). The expectations created by the institutional stance on this reading are not *merely* one of predictive expectation enabling reliable coordination, but rather a normative trust-based expectation that role-holding agents will fulfil obligations entrusted upon them.

This distinction matters because we treat other people not as mere objects to be managed or controlled based on predictive expectations but as fellow responsible agents in a shared norm-governed community. Instead of an objective stance, we direct normative expectations and judgments on others as responsible, role-holding agents. Using Jara-Ettinger and Dunham's example, our interaction with a barista is not merely "broadly similar" to a vending machine, which we treat in a purely functional, instrumental manner. We treat (or better, ought to treat) the barista instead as a role-holding agent whom we trust will follow through on their given duties.

The distinction can be roughly encapsulated by a predictive (or explanatory) "is" and a normative "ought." The "ought" here is also not reducible to injunctive norms explained in terms of prudential motivations (i.e., fear of punishment), but is instead about robust duties and obligations based on role-based statuses and commitments.

For instance, a person elected to public office who violates ethical standards might suggest a failure of role prediction or a shift towards a mentalistic stance to identify their motives. But we do not *merely* predict what people holding public office *will* do nor *merely* identify what their motives and intentions are. We primarily hold them accountable to role-based standards expected of people holding public office. Even where their past behavior gives us reliable evidence that is highly predictive of nefarious motives and ill intentions, we still think that they *should* (even if this goes against the evidence) have acted based on the obligations demanded of the role. Here, the institutional stance is less about explaining people's behavior and us adjusting to predicted expectations. Instead, it is about participating in a practice of holding one another responsible for roles we recognize and commit to.

The institutional stance can, of course, be "evidentialized" based on predictive explanations gleaned from the role. But this takes the observational or objective perspective as operative in the intuitive theory. The likely content of the intuitive theory is the participant's perspective of taking on and tracking normative role-based commitments. A politician who takes an oath of office is subject to normative standards of honesty, integrity, etc. These normative standards and expectations usually track predictability because we hold role-holders like politicians accountable in fulfilling their duties and obligations. Our robust practices of holding responsibility as such are the main engine that drives what, from an external observational standpoint, looks like reliably predictable behavior. Deviations from role obligations and standards are not merely treated as behavioral evidence warranting further explanation. We also primarily treat them as actions that demand accountability. Prediction is thus likely only an epiphenomenon to the main engine of the intuitive theory, normative regulation driven by responsibility judgments.

Nonetheless, we might also decide to take the objective stance, perhaps in trying to manage or anticipate a situation where efforts at accountability have failed. Evidentializing can perhaps be a proper response to such circumstances. But this, I suggest, takes on a detached, non-participant view of the institutional structure. On this evidentialist view, we treat others not as responsible agents but as variables towards whom we make causal, predictive adjustments (for instance applying "nudge" strategies).

Jara-Ettinger and Dunham also point out that our social cognition is a "dynamic reactive system" which is reflexive in the sense that our conceptualizations of it affects the very dynamics of the system itself. Perhaps taking an objective stance is just one way of conceptualizing our role-based institutional structures. Perhaps this mirrors different approaches to moral reasoning, with the objective stance roughly mirroring utilitarian reasoning and the participant stance mirroring duty-based deontic reasoning. But insofar as the institutional structure gives us an "interpretive framework" that specifies roles with "rule-like normative entailments," deontic reasoning grounded on a role's specification of duties looks more apt.

Moreover, there are ways to conceptualize rule-governed institutional structures not as functional mechanisms for coordination but as co-regulated practices organized to scaffold and realize responsible agency (An, 2026; Zabel, 2015). Our conceptions of institutions thus matter insofar as those who inhabit institutional roles are perceived either as responsible agents or perhaps as something lesser. So how we conceptualize institutions, as *mere* predictive mechanisms for behavior coordination or as normative spaces for relating with and regulating expectations between responsible agents, is also of moral importance.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

References

An, C. J. (2026). Against cooperation: Or why Telic/functional explanations do not account for the evolution of rule-following. *Topoi*, *45*(1), 189–205. https://doi.org/10.1007/s11245-025-10173-9

Holton, R. (1994). Deciding to trust, coming to believe. *Australasian Journal of Philosophy*, *72*(1), 63–76. https://doi.org/10.1080/00048409412345881

Marušić, B. (2017). Trust, reliance and the participant stance. *Philosopher's Imprint*, *17*(17), 1–10. https://hdl.handle.net/2027/spo.3521354.0017.017

Sartre, J.-P. (2022). *Being and nothingness: An essay in phenomenological ontology*. Routledge. https://doi.org/10.4324/9780429434013

Searle, J. (2010). *Making the social world: The structure of human civilization*. Oxford University Press.

Strawson, P. (1962). Freedom and resentment. *Proceedings of the British Academy*, *48*, 187–211.

Zabel, B. (2015). The institutional turn in Hegel's philosophy of right: Towards a conception of freedom beyond individualism and collectivism. *Hegel Bulletin*, *36*(1), 80–104. https://doi.org/10.1017/hgl.2015.5

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

The institutional stance, self-fulfilling prophecies, and hidden structures

Lieke Asma*

Munich School of Philosophy, Germany
liekeasma@gmail.com

*Corresponding author.

doi:10.1017/S0140525X26105512, e370

Abstract

If Jara-Ettinger and Dunham are right about the pervasive role the institutional stance plays, we have to seriously consider that this stance perpetuates injustice, because it tends to reinforce, normalize, and hide problematic structures. To address the problems that come with this stance, we need to take into account history and context and ask ourselves what is morally right.

Jara-Ettinger and Dunham claim that "[t]he 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." But is the institutional stance itself cut out to provide these tools, or is it rather more likely to perpetuate existing injustices?

Of course, a first step to criticizing social roles and their associated institutes and systems is to recognize the pervasive role they play in our lives, in the sense that we often use a person's social role to understand and predict their behavior, and base our expectations on it. Critically reflecting on this, however, involves going beyond the institutional stance. For one thing, it involves recognizing how people end up in, for example, dominant or subordinate social roles, and considering the extent to which this is often outside of the person's control. Furthermore, it requires us to recognize how certain expectations that come with social roles are unfair or have unfair consequences. Social change, then, will not come from the institutional stance itself, but from reflecting on social roles and expectations from historical, contextual, and moral perspectives. And importantly, we do not often seem to take up this kind of critical reflection. In this commentary, I will address three reasons for why the institutional stance may prevent us from doing so. Together, these reasons show how the institutional stance in fact explains why social change is hard and injustices tend to persist.

First of all, we can take a variety of stances to understand human behavior and, as Jara-Ettinger and Dunham address as well, this is not neutral: stances lead to looping effects, and they affect social behavior. This is especially clear in case of social roles, since they are strongly tied to the expectations we have of other persons. We do not just predict or try to understand, we judge and criticize when persons do not meet these expectations. But if that is true, there is not such a clear distinction between the stance and reality, not (just) because the institutional stance provides us with such a good picture of reality, but because the institutional stance makes its own expectations true: it is self-fulfilling (see, e.g., Peters, 2022). Expecting women to take care of family members in need, for example, may make it so that women in fact do often take care of others. As a result, adopting the institutional stance tends to prevent social change, because it tends to keep persons in the positions they are expected to be in.

This brings me to my second and related point: Jara-Ettinger and Dunham make a rather sharp distinction between behavior that is "constrained" by social roles and "unconstrained" behavior that is the result of personal desires. But things are not as simple as that. We are born into an existing world with established social roles as a woman or minority, for example, and these social roles and expectations affect who we are and how we perceive ourselves. Our desires have been shaped by how others perceive and treat us, and it may take time and effort to come to recognize our so-called adaptive preferences, that in fact perpetuate our own oppression (e.g., Khader, 2012). Understanding persons from the institutional stance, then, may make it more likely that they adapt their preferences to what is expected of them in virtue of their social role. Even if their behavior may seem unconstrained and a matter of personal desire, social roles may still unintentionally and unconsciously play a role in what persons want and decide. This, in turn, may suggest that there is no problem that requires addressing; persons just do what they want or end up in a social role that suits them.

Third and finally, at some points, Jara-Ettinger and Dunham talk about hidden institutional structures and role-based systems, but it remains unclear what they exactly mean by this. That is, they do not elaborate on the sense in which these systems and structures are hidden, especially given the fact that the institutional stance at the same time is taken to be quite pervasive. My proposal is that we may be using social roles to understand and expect certain behavior, but we do not always recognize that this is influenced by *a* social role or by *this* social role. For example, it may turn out that we expect of mothers to drop everything to pick up their sick child, think we expect this of all parents, but in reality we do not have the same expectations of fathers. An explanation for this might be misattribution: we think that it is the person's *parent*hood because of which we have this expectation of them, but in reality, it is their *mother*hood (see Asma, 2025 for further analysis). In that sense, the institutional stance can be pervasive and hidden at the same time, because we fail to realize where our expectations exactly come from and what they imply. As such, again, we may fail to recognize that our expectations are unfair and do not see any reason to critically reflect.

To conclude, if Jara-Ettinger and Dunham are right about the pervasive role the institutional stance plays, and I think they are, we have to seriously consider that this stance may perpetuate injustice because it tends to reinforce, normalize, and hide problematic structures. To address the problems that come with this stance, the fact that, but also why certain persons end up in social positions that are subordinate and out of line with their well-being, we should take into account history and context, and ask ourselves what is morally right. A better picture thus is: the institutional stance exists and plays a pervasive role in how we perceive and understand others, but it tends to keep social structures in place and explains why unfair systems are hiding in plain sight.

Financial support. Lieke Asma's research for this commentary was funded by the Deutsche Forschungsgemeinschaft-AS 667/1–2.

Competing interests. None.

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

References

Asma, L. J. F. (2025). Explaining unconscious discrimination: Misattribution and rationalization. *Phenomenology and the Cognitive Sciences*. https://doi.org/10.1007/s11097-025-10061-y

Khader, S. J. (2012). Must theorising about adaptive preferences deny women's agency?. *Journal of Applied Philosophy*, *29*(4), 302–317. https://doi.org/10.1111/j.1468-5930.2012.00575.x

Peters, U. (2022). What is the function of confirmation bias?. *Erkenntnis*, *87*(3), 1351–1376. https://doi.org/10.1007/s10670-020-00252-1

The institutional stance enables key insights into problems of selfhood and meaningfulness

Roy Baumeister[a]*, Dianne Tice[b] and Yukun Zhao[c]

[a]BIGSSS, Beijing Normal University, Beijing , China; [b]Brigham Young University, USA and [c]The Chinese University of Hong Kong, Hong Kong

r.baumeister@uq.edu.au

diannetice@byu.edu

zhaoyukun@gmail.com

*Corresponding author.

doi:10.1017/S0140525X2610555X, e371

Abstract

The notion of institutional stance clarifies existential problems. The institutional stance enabled the ancients to view society as stably organized and to rather easily find their place in it. Societal changes gradually complicated the pursuit of self-knowledge in this fashion, and the meaning of life emerged as a problem in the 1800s as Western civilization transformed its institutions.

Two of the major existential dilemmas in modern life involve the quest for self-understanding and the quest for a meaningful life. People have sought self-knowledge in diverse ways, ranging from psychoanalysis to astrology to psychedelic drugs. Meanwhile, the quest for meaning has also become an increasing focus of individuals. Both selfhood and meaningfulness have become major research areas in psychology as well.

We find Jara-Ettinger and Dunham's theory of the institutional stance offers exciting possibilities for advancing psychology's theories about selfhood and meaningfulness. By way of context, our view of the human mind is that it was largely shaped by evolution that selected for traits that facilitate participation in culture (e.g., Baumeister, 2005). The distinctively human traits are therefore the result of biological adaptations that improve the organism's ability to participate in and benefit from culture. Definitions of culture vary considerably, but we understand it as shared beliefs and practices that support a social system – presumably one that uses cooperation and communication to prolong life. Haidt (2012) characterized humans as the "world champions of cooperation," given that humans cooperate far more extensively and productively than other species. Such productive cooperation has contributed to the biological flourishing of the human species.

Central to that success is the creation of complex social systems. These include economic marketplaces, military organizations, governments, religious groups, corporations, and more. Unlike psychology's emphasis on the individual and single mind, sociology has sought to understand society from the top down, and in our experience, such analyses often emphasize the various institutions that populate and organize a society. Although many animals have societies (e.g., Moffett, 2018), they do not have anywhere close to the diversity of institutions found in a modern human society: schools, companies, banks, laboratories, mass media, churches, and more. Some animals, such as ants, do have a kind of army, but they lack the command structures that human armies have. Indeed, Moffett et al. (2021) pointed out that ant society, despite its role differentiation and other characteristics, does not have bosses or leaders. To be sure, the aggressive egalitarianism of hunter-gatherer societies (e.g., Boehm, 1999) precluded bosses as well, but the egalitarian innovation fostered collective decision-making based on group discussion, which again is minimal to nonexistent in other species.

The institutional stance is an extremely useful notion for elaborating the view of humans as essentially cultural animals. It may be a vital evolutionary innovation to enable and encourage the mind to search the social world for organized systems and fit oneself into these. The developmental phenomena of overimitation (i.e., mimicking nonessential behaviors alongside causally essential actions) show that, unlike other apes, humans observe others not only to learn about the physical environment but also to learn *how things are done in our social group* (Tomasello, 2014).

The historical backgrounds of how the self became problematic and how life's meaning became problematic are quite different. The self became a problem gradually, over many centuries and in multiple steps (Baumeister, 1987). Although "know thyself" was engraved in stone at Delphi and has been attributed to multiple different Greek philosophers, it apparently did not have its modern meaning of exploring one's inner self in the hope of having life's goals and mysteries revealed from one's unconscious mind. Rather, it may have been closer to "know your place," as in knowing one's role in society and what its duties are. The notion of an institutional stance makes that interpretation highly plausible: People assume there are social structures and seek to fit in to them, and to do that, they must understand their own role(s) along with attendant obligations and expectations. As society became more complex and people took on responsibility for shaping their identities amid multiple opportunities, the essence of one's self became less a fixed or given property but something more changeable. A major advance was that people began to think of the self as an entity that took on roles, rather than equating the self with its roles (Baumeister, 1987).

The problematization of life's meaning was more historically concentrated than the problematization of selfhood. As found by Baumeister and Zhao (2026), life's meaning became a problem during the 1800s, consistently across six major Western societies. Why then?

To be sure, there have been turmoil and uncertainty throughout history. Nevertheless, the 1800s arguably posed extraordinary challenges for the institutional stance. For centuries, Western civilization had operated with the framework of Christian faith.

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

Society was assumed to be largely stable (Nisbet, 1973) based on God's plan. (As Jara-Ettinger and Dunham make clear, the proto-institutional social structures of animal societies are likewise quite stable, even as different animals take their places in it.) Life's meaning was understood in that context. But during the early modern period (1500–1800), many long-standing certainties came into question. Secularization, driven in part by religious conflict and scientific advances, undermined the assumption that all humans lay under the obligation to live their life as God had supposedly ordained. Radical political changes undermined the notion that the social system was essentially stable, and indeed, the divine right of kings to rule was vividly undermined when England and France executed their kings, and the upstart United States dispensed with kingship altogether. Communism and anarchism offered radically different views of how society's institutions could and should function. Industrialization changed the nature and meaning of daily work. Urbanization challenged moral values (as did secularization). Social mobility enabled people to change their position in society, for better or worse.

Thus, the institutional stance directed how people sought to understand themselves and to find meaning in life. As institutions became more complex, allowing more scope for self-determination, selfhood became increasingly problematic. And as institutions shifted and changed, it became more difficult for the institutionally attuned mind to establish the meaning of its life.

Financial support. No grant funds were used.

Competing interests. The authors have no competing interests to declare.

References

Baumeister, R. F. (1987). How the self became a problem: A psychological review of historical research. *Journal of Personality and Social Psychology*, *52*, 163–176.

Baumeister, R. F. (2005). *The cultural animal: Human nature, meaning, and social life*. Oxford University Press.

Baumeister, R. F., & Zhao, Y. (2026). *How life's meaning became a problem*. Unpublished manuscript. Chinese University of Hong Kong.

Boehm, C. (1999). *Hierarchy in the forest: The evolution of egalitarian behavior*. Harvard University Press.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Pantheon.

Moffett, M. W. (2018). *The human swarm*. Basic Books.

Moffett, M.W., Garnier, S., Eisenhardt, K.M., Furr, N.R., Warglien, M., Sartoris, C., Ocasio, W., Knudsen, T., Bach, L.A., & Offenberg, J. (2021). Ant colonies: Building complex organizations with miniscule brains and no leaders. *Journal of Organizational Design*, *10*, 55–74.

Nisbet, R. (1973). *The social philosophers: Community and conflict in Western thought*. Crowell.

Tomasello, M. (2014). *A natural history of human thinking*. Harvard University Press.

When markets meet minds: The institutional stance resolves decades of theoretical ambiguity in behavioural economics

Michał Białek[a]*, Katarzyna Pypno-Blajda[b] and Kacper Szymański[a]

[a]Institute of Psychology, University of Wroclaw, Wroclaw, Poland and [b]Department of Psychology, University College of Professional Education, Poland

michal.bialek3@uwr.edu.pl

kpypno@gmail.com

kacper.szymanski@amu.edu.pl

*Corresponding author.

doi:10.1017/S0140525X26105548, e372

Abstract

Behavioural economics knows that market norms displace social norms, but has never explained why the substitution is cognitively abrupt or why some findings built on it replicate while others do not. Jara-Ettinger and Dunham's institutional stance supplies both answers: stance adoption makes the two logics genuinely incompatible, and embeddedness in real role relationships determines whether an effect survives replication.

This puzzle bothered behavioural economists: You offer a friend $10 to help you move a sofa, and he is offended. You offer nothing and simply ask for help, and he shows up on Saturday with coffee and good humour. Everyone intuitively senses what went wrong in the first scenario, but the field has never had a fully satisfying account of the mechanism. Jara-Ettinger and Dunham (henceforth JED) finally provide the connection: behavioural economics and social psychology have been studying stance deployment for decades without realizing it. The JED framework gives this scattered empirical landscape the theoretical backbone it has been missing and, as we argue, explains why parts of that landscape failed to replicate.

The deepest connection lies in Margaret Clark and Judson Mills's (1993) research programme on communal and exchange relationships: when you help a friend, you give because they need something; when you help a business partner, you give because they gave first, or because you expect them to give later. Communal and exchange logics generate very different downstream

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

consequences when they get mixed up. Monitoring contributions in a friendship – Clark and Mills's term for tracking who gave what to whom – feels like a betrayal rather than just awkward social behaviour, because it signals that you have stopped reading your friend's helpfulness as an expression of genuine care and started reading it as the discharge of a role obligation.

The JED framework explains exactly why this reclassification carries such weight. Exchange relationship behaviour follows naturally from the institutional stance: behaviour gets attributed to role-based entitlements, tracking is appropriate because role compliance is what matters, and violations signal role-departure. Communal relationship behaviour follows from the mentalistic stance: helpfulness gets attributed to accurate perception of the other's needs and genuine desire to meet them, and tracking is harmful because it overwrites a personal attribution with an institutional one. When your friend starts keeping score, the offence runs deeper than a mere norm violation because an entire attributional framework has shifted. The two stances generate fundamentally different answers to the question of what the same behaviour reveals about the person performing it. JED supply Clark and Mills's architecture, explaining why the two modes feel so qualitatively different and why crossing between them carries such costs.

Market mindset research adds another layer. People turn toward market relationships specifically when they feel a threat to personal control and need predictability in their social environment (Gasiorowska & Zaleskiewicz, 2021), and more anxiously attached individuals are especially likely to do so under threat (Gasiorowska & Zaleskiewicz, 2023). The structured, rule-governed quality of transactional interactions becomes psychologically attractive precisely when communal relationships feel too uncertain or demanding, a finding that maps directly onto JED's argument that the institutional stance reduces the cognitive demands of full mentalistic inference.

When both stances are simultaneously active, JED predict interference rather than a clean blend. This pattern has a vivid real-world signature: the moral suspicion surrounding philanthropic entertainment – the genre of YouTube creators who fund surgeries and give away cars while filming everything for millions of viewers – where identical charitable outcomes attract criticism simply because the transactional structure is too explicit to suppress (Białek, 2026). Newman and Cain (2014) demonstrated the laboratory analog: charitable acts that also benefit the donor receive less moral credit even when the net good is greater. The same logic that explains co-activation interference also predicts which behavioural economics findings will hold under replication and which will not.

Stance-based effects tend to be most robust when a manipulation genuinely places participants inside real role structures and real relational stakes, but most fragile when it tries to activate a mindset through incidental semantic exposure without changing anything about the actual relational context. Vohs and colleagues' money priming paradigm exemplifies the fragile end: participants count Monopoly money or unscramble sentences containing financial words, manipulations that alter semantic context without embedding participants in any actual institutional relationship, and large pre-registered replications failed to reproduce the originally reported behavioural effects (Rohrer, Pashler, & Harris, 2015; Vohs, Mead, & Goode, 2006, 2008).

Gneezy and Rustichini's (2000) daycare study is more instructive because it looks like genuine contextual manipulation: introducing a fine into an ongoing parent–caregiver relationship explicitly reframes the interaction in transactional terms, turning late pickup from a breach of trust into a paid service. The original field study produced the famous counterintuitive result, with late pickups increasing after the fine was introduced and not recovering after its removal. But Metcalf, Powell, and Zhong (2020) found the opposite pattern in a survey-based replication, with fines reducing non-compliance and effects reversing cleanly. We suggest that the JED framework explains this divergence. Vignette participants evaluate a hypothetical scenario involving strangers rather than navigate a real relationship with caregivers they see daily, and we hypothesize that this strips away the embeddedness in an actual role relationship that the institutional stance requires to engage. If so, the methodology was rigorous but theoretically misspecified, and what a study needs to replicate a field finding is not just methodological rigour but genuine embeddedness in an actual role relationship. This tells us not just why some studies fail to replicate, but what a study needs to contain if it genuinely deserves to succeed.

One might ask whether mapping behavioural economics onto the JED framework is merely relabelling – attaching new vocabulary to patterns researchers already understood. We think not. The JED framework specifies the *mechanism* rather than the pattern. Prior accounts described that market norms displace social norms; the institutional stance explains why the substitution is so cognitively abrupt, why the two logics cannot simply coexist, and why embeddedness in a real role relationship is a prerequisite rather than a convenience. This clarifies why some findings have proven robust for thirty years while others evaporated under replication, and why disparate phenomena share a common cognitive source. The JED framework explains this uneven landscape and tells us what to do differently.

Acknowledgements. None.

Financial support. This research was funded by the National Science Centre, Poland, grant no. 2023/51/B/HS6/02522 assigned to Michał Białek.

Competing interests. None.

References

Białek, M. (2026). Breaking the fourth wall of philanthropy. *Journal of Medical Ethics Practice & Bioethics*, *1*(2), e000092. https://doi.org/10.1136/jmepb-2025-000092

Clark, M. S., & Mills, J. R. (1993). The difference between communal and exchange relationships: What it is and is not. *Personality and Social Psychology Bulletin*, *19*(6), 684–691. https://doi.org/10.1177/0146167293196003

Gasiorowska, A., & Zaleskiewicz, T. (2021). Desired and real control in market and non-market social exchanges. *Journal of Personality and Social Psychology*, *120*(4), 965–988. https://doi.org/10.1037/pspi0000299

Gasiorowska, A., & Zaleskiewicz, T. (2023). Money, attachment, and the need for cognitive closure: How market mindset protects anxiously attached people from uncertainty. *Journal of Social and Personal Relationships*, *40*(1), 291–314. https://doi.org/10.1177/02654075221108791

Gneezy, U., & Rustichini, A. (2000). A fine is a price. *Journal of Legal Studies*, *29*(1), 1–17. https://doi.org/10.1086/468061

Metcalf, L., Powell, D., & Zhong, S. (2020). Replication of Gneezy and Rustichini (2000b). *Journal of Economic Psychology*, *81*, 102306. https://doi.org/10.1016/j.joep.2020.102306

Newman, G. E., & Cain, D. M. (2014). Tainted altruism: When doing some good is evaluated as worse than doing no good at all. *Psychological Science*, *25*(3), 648–655. https://doi.org/10.1177/0956797613504785

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

Rohrer, D., Pashler, H., & Harris, C. R. (2015). Do subtle reminders of money change people's political views? *Journal of Experimental Psychology: General, 144*(4), e73–e85. https://doi.org/10.1037/xge0000058

Vohs, K. D., Mead, N. L., & Goode, M. R. (2006). The psychological consequences of money. *Science*, *314*(5802), 1154–1156. https://doi.org/10.1126/science.1132491

Vohs, K. D., Mead, N. L., & Goode, M. R. (2008). Merely activating the concept of money changes personal and interpersonal behaviour. *Current Directions in Psychological Science*, *17*(3), 208–212. https://doi.org/10.1111/j.1467-8721.2008.00576.x

Where IS the ought? The Institutional Stance lacks normative force

Lucas Payne Butler[a]* and Marco F. H. Schmidt[b]

[a]University of Maryland, College Park, USA and [b]Universität Konstanz, Germany
lpbutler@umd.edu
marco.schmidt@uni-konstanz.de

*Corresponding author.

doi:10.1017/S0140525X26105597, e373

Abstract

The Institutional Stance (IS) offers a promising account of how humans navigate institutional structures, but it suffers from a problem inherent in purely third-person frameworks: a detached observer can predict behavior aligned with institutional expectations, but cannot feel the binding grip of normative force. We propose reconceptualizing the IS as an important aspect of social cognition rooted in shared intentionality.

Jara-Ettinger and Dunham's Institutional Stance (IS) is a valuable framework for understanding how humans navigate complex socio-cultural structures. Drawing an analogy with Dennett's (1987) intentional stance, the IS holds that we explain, predict, and reason about social behavior by making inferences about roles and behavior governed by institutional norms and expectations (Searle, 1995). We share the authors' aim of accounting for human capacities to understand and represent actions within institutional structures, a capacity central to human social cognition. However, we argue that the proposed IS framework cannot fully account for the normative dimension of institutional life.

The core issue is that on any standard definition of institutional phenomena (cf., Giddens, 1984; Hodgson, 2006; Searle, 1995, 2010; Turner, 1997), the IS account must go beyond merely predicting and explaining human actions. It must account for the binding nature of such phenomena, explaining how prescriptions, obligations, entitlements, and justifications come to carry normative force (Schmidt & Rakoczy, 2023). As formulated, the IS acknowledges the importance of normativity, but fails to fully capture the central distinction between descriptive and normative expectations. Institutional cognition requires more than predicting how someone *will* act (e.g., by attributing individual mental states). It requires a representation of how someone *ought* to act according to shared norms governing the social role they are enacting (i.e., based on collective mental states).

For example, consider a simple, everyday example: your local barista. Our expectations for how the barista enacts their role are not merely predictive. They also have normative force and what Searle (1983) termed a world-to-mind direction of fit. If the barista fails to perform the role as normatively expected – say, serving the tallest customers first, acting unfriendly, or simply failing to say when our drink is ready – we expect and may enforce some consequence ourselves. This would likely be true, all things considered, even if we travel to a café thousands of miles away.

At its core, this limitation in the current formulation of the IS stems from an inherent problem with purely third-person (here, theory-theory) approaches to normativity (Sellars, 1963; Strawson, 1962). From day one, humans are "thrown in" to a normatively prestructured world full of rules, traditions, roles, and meanings (Heidegger, 1927/1984; Vygotsky, 1978). Correspondingly, humans have early social motivations to participate in social interactions and to be part of a group or collective, learning step by step how "we" do things. Notably, to feel the binding force of norms, being embedded/participating in early (and later) social interactions is not optional, but necessary. A detached observer, as an individual adopting the IS would be, might predict and explain others' behavior in institutional contexts. But they would have little reason to hold either themselves or others to social norms, because expectations within a purely causal-explanatory framework lack normative force. A solitary modeler computing institutional regularities from the outside might be analogous to an observer from an intelligent alien species lacking social norms. They can make predictions from patterns of behavior, but cannot feel the deontic "oughtness" of norms – the expectations, entitlements, and obligations that make institutions and the roles they govern what they are.

Young children's understanding of human action is not merely predictive, but represents socio-affective and normative dimensions from early on. Eighteen-month-olds display concern in response to moral, but not conventional, transgressions (Kassecker, Verschoor, & Schmidt, 2023), and correct non-conforming actions by others (Schmidt, Rakoczy, & Tomasello, 2019). By age three, children protest when others violate norms within institutional structures, including game rules, ownership conventions, and entitlements (Rakoczy, Warneken, & Tomasello, 2008; Schmidt, Rakoczy, & Tomasello, 2013), even without explicit institutional framing (Schmidt et al., 2016). This early norm enforcement goes well beyond mere prediction. When children see a puppet break a rule, they are not merely surprised. They protest, correct, and sanction, clearly understanding the action *should* not have occurred. Preschoolers even extend this to individual epistemic states and speech acts (Butler et al., 2018; Fedra & Schmidt, 2019; Rakoczy & Tomasello, 2009), something the IS has difficulty in capturing and delegates to the non-normative institutional stance. Overall, the tendency to rapidly infer normative structure from minimal input (Schmidt et al., 2016; Wolf et al., 2026) cannot be wholly explained by a purely predictive IS framework.

We propose that the IS would be better conceptualized not as a separate stance analogous to the intentional stance, but rather as a sophisticated aspect, another dimension of social cognition enabled by and built upon shared intentionality (Schmidt & Tomasello, 2012; Tomasello et al., 2005; Tomasello, 2019). Understanding institutions means understanding norms and roles, and thus grasping that members of a group collectively accept that roles carry certain obligations and entitlements. This

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

shared or "*we*-intentional" perspective helps supply the normative force absent in the proposed framework. To return to our previous example, a barista's obligations (and entitlements) are not directly observable from behavioral regularities. Rather, participants in the institution (customers, baristas, other employees) jointly recognize and hold one another accountable to their collective obligations and entitlements (Searle, 1995, 2010). From a developmental perspective, we suggest that children do not only layer on normative understanding after having observed sufficient behavior in a given context. Rather, they engage in shared practices from early in life, providing the basis and giving rise to both institutional understanding and normative cognition simultaneously (Schmidt & Rakoczy, 2023).

This reconceptualization of the IS framework has multiple advantages. It is more parsimonious, connecting the IS to extensive literature on norm psychology and its development without positing an independent framework that must be coherently integrated with other stances. It also resolves concerns stemming from a purely third-person formulation of the IS. Shared intentionality supports that perspective, but also provides both the first-person plural and second-person perspectives necessary for normative force. We encourage the authors to integrate this participatory, normative dimension into their otherwise promising account of institutional cognition.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. None.

References

Butler, L. P., Schmidt, M. F. H., Tavassolie, N. S., & Gibbs, H. M. (2018). Children's evaluation of verified and unverified claims. *Journal of Experimental Child Psychology*, *176*, 73–83. https://doi.org/10.1016/j.jecp.2018.07.007

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

Fedra, E., & Schmidt, M. F. H. (2019). Older (but not younger) preschoolers reject incorrect knowledge claims. *British Journal of Developmental Psychology*, *37*(1), 130–145. https://doi.org/10.1111/bjdp.12264

Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Polity Press.

Heidegger, M. (1927/1984). *Sein und Zeit* (15th ed.). Max Niemeyer Verlag.

Hodgson, G. M. (2006). What are institutions? *Journal of Economic Issues*, *40*(1), 1–25. https://doi.org/10.1080/00213624.2006.11506879

Kassecker, A., Verschoor, S. A., & Schmidt, M. F. H. (2023). Human infants are aroused and concerned by moral transgressions. *Proceedings of the National Academy of Sciences*, *120*(31), e2306344120. https://doi.org/10.1073/pnas.2306344120

Rakoczy, H., & Tomasello, M. (2009). Done wrong or said wrong? Young children understand the normative directions of fit of different speech acts. *Cognition*, *113*(2), 205–212. https://doi.org/10.1016/j.cognition.2009.07.013

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), 875–881. https://doi.org/10.1037/0012-1649.44.3.875

Schmidt, M. F. H., 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. https://doi.org/10.1177/0956797616661182

Schmidt, M. F. H., & Rakoczy, H. (2023). Children's acquisition and application of norms. *Annual Review of Developmental Psychology*, *5*(1), 193–215. https://doi.org/10.1146/annurev-devpsych-120621-034731

Schmidt, M. F. H., Rakoczy, H., & Tomasello, M. (2013). Young children understand and defend the entitlements of others. *Journal of Experimental Child Psychology*, *116*(4), 930–944. https://doi.org/10.1016/j.jecp.2013.06.013

Schmidt, M. F. H., Rakoczy, H., & Tomasello, M. (2019). Eighteen-month-old infants correct non-conforming actions by others. *Infancy*, *24*(4), 614–635. https://doi.org/10.1111/infa.12292

Schmidt, M. F. H., & Tomasello, M. (2012). Young children enforce social norms. *Current Directions in Psychological Science*, *21*(4), 232–236. https://doi.org/10.1177/0963721412448659

Searle, J. R. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.

Searle, J. R. (1995). *The construction of social reality*. Free Press.

Searle, J. R. (2010). *Making the social world*. Oxford University Press.

Sellars, W. (1963). *Science, perception and reality*. Routledge & Kegan Paul.

Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, *48*, 187–211.

Tomasello, M. (2019). *Becoming human: A theory of ontogeny*. Harvard University 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*(5), 675–691. https://doi.org/10.1017/S0140525X05000129

Turner, J. (1997). *The institutional order*. Longman.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wolf, W., Iva, V., Larsen, I., & Tomasello, M. (2026). Young children enforce self-created norms promiscuously. *Journal of Experimental Child Psychology*, *262*, 106396. https://doi.org/10.1016/j.jecp.2025.106396

Dehumanizing those closest to us

Mina Cikara*

Harvard University, USA
mcikara@fas.harvard.edu

*Corresponding author.

doi:10.1017/S0140525X26105676, e374

Abstract

I extend the scope of the institutional stance analysis to a potentially surprising domain: marriage. Long-term romantic relationships, like many other institutions, often give rise to strongly prescribed roles for both parties. Over time, the resulting models and expectations of our counterparts may become sufficiently well articulated that they supplant the individual mind in front of us.

The authors introduce the institutional stance as the mental process of representing other people with respect to their roles rather than their individual minds (as unfolds in the mentalistic stance). In this framework, an institution is "a type of cognitive representation consisting of a set of inter-connected roles with a level of persistence and continuity," which may or may not be formally codified. Here I extend the application of the institutional stance analysis to marriage, which may not be an obvious target given that many of the vivid examples in the target article are hierarchical (including parent–child structures) or occupation-bound. My argument is that some institutions – here, marriage, or monogamous long-term romantic partnerships, more broadly –

support the creation of such well-articulated models and role-bound expectations of our counterparts, that the roles come to replace the person in front of us, and in so doing, strip them of their full humanity.

At least two features of long-term partnerships could contribute to this phenomenon. First, it may be that over time, the role-based models we develop of our partners are sufficient to support coordinated action in the context of this institution (e.g., division of labor in the home). In the absence of any major system failures, the motivation to devote continued attention and processing resources – as is required by theory of mind (Saxe, 2006) – may decline. Second, the demands of life beyond the partnership may constrain the cognitive resources we have available to engage in the costly process of continuous mentalizing. Thus, even when partners are motivated, they may not have sufficient bandwidth to follow through on adopting a mentalistic stance. The challenge is that circumstances and people change; the models we develop for our partners-as-roles may degrade in their fitness over time as models of our partners-as-minds. In this way, we cease to see clearly those ostensibly closest to us.

This idea is inspired by work on the Continuum Model of impression formation (Fiske & Neuberg, 1990) and other related dual-process models of social cognition. According to the Continuum Model, the default representation we have when we meet a new person is largely based on the stereotypes we associate with them (e.g., on the basis of visual features, such as skin tone or size). However, with sufficient motivation and capacity, we can push past the stereotype to build more elaborate, individualized representations: filling the representations in with the person's specific traits and attributes. Examples of conditions which activate the *motive* to elaborate include needing to build an accurate representation of the other in support of some aim (e.g., finding a study group partner) and personal relevance of the target for one's own outcomes (e.g., determining whether someone might make a good roommate). Cognitive *capacity* simply refers to having sufficient mental resources to devote to the task of individuating. Accordingly, situational constraints like time pressure and stress undermine individuation, resulting in more automatic, stereotype-based judgments of others.

Critically, creating outcome dependency between people is one of the most robust interventions for activating individuation in lab experiments (Fiske, Lin, & Neuberg, 1999). Are not long-term relationships the paradigmatic instance of outcome dependency? While romantic relationships may be incredibly motivating with respect to engaging individuating processes, including theory of mind, early on, these more effortful processes may give way over time to habits, rules, and routines within the context of a long-term relationship. Add to this the challenges and stresses of daily life – work and domestic demands, raising children, taking care of sick parents and pets – and the likelihood of relying on role-based models and the institutional stance (not unlike stereotypes) only increases.

There are two important implications of the Continuum Model and other related dual-process models of social cognition, more generally (Chaiken & Trope, 1999). First, because individuation requires both motivation and capacity, category-based cognitive processing tends to prevail. Second, people attend more to category-consistent than inconsistent information. Therefore, within the marriage case study, once role-based expectations are well established, the institutional stance may dominate mentalizing processes, making it particularly hard to detect when and why our partner's minds believe, desire, or prefer something different than the role-based model suggests.

This case study raises several interesting questions for the nature of the relationship between the institutional and mentalistic stances. First, is it the case that the institutional stance dominates the mentalistic stance in long-term relationships (in the sense that it is adopted more readily as a default)? Second, are institutional and mentalizing stances mutually inhibiting? And related, if there is a bistability in the representation of our partners along these lines, what triggers top-down control over which representation is appropriate or active in a given context? And finally, are there any relationships with a minimal level of persistence and continuity that do *not* eventually become institutions by this definition?

Acknowledgments. I thank Jara-Ettinger for discussing this idea with me several years ago when he first shared the idea of the institutional stance.

Financial support. Not applicable.

Competing interests. I have no competing interests.

References

Chaiken, S., & Trope, Y. (Eds.). (1999). *Dual-process theories in social psychology*. Guilford Press.

Fiske, S. T., Lin, M., & Neuberg, S. L. (1999). The continuum model: Ten years later. In S. Chaiken & Y. Trope (Eds.), *Dual-process theories in social psychology*, (pp. 231–254). The Guilford Press.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. *23*, pp. 1–74). Academic Press.

Saxe, R. (2006). Uniquely human social cognition. *Current Opinion in Neurobiology*, *16*(2), 235–239.

Before institutions: The social and cognitive foundations of institutions

Fabrice Clément[a]* and Laurence Kaufmann[b]

[a]Cognitive Science Center, Université de Neuchâtel, Neuchâtel, Switzerland and [b]Université de Lausanne, Faculté des sciences sociales et politiques, Switzerland
fabrice.clement@unine.ch
laurence.kaufmann@unil.ch

*Corresponding author.

doi:10.1017/S0140525X2610569X, e375

Abstract

We agree with J. Jara-Ettinger and Y. Dunham that social understanding cannot be reduced to mental-state attribution, but we challenge their view that role-based expectations are

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

inherently institutional. Institutions require collective symbolic functions, unlike expectations that arise from recurring relational patterns. These patterns – what we call naïve sociology – underlie an elementary social stance that evolutionarily predates institutional reasoning and provides a more parsimonious basis for interpreting social behavior.

We welcome Julian Jara-Ettinger and Yarrow Dunham's argument that the understanding and prediction of others' behavior cannot be reduced to mental-state reasoning. Their article contributes to our long-standing efforts to demonstrate that social cognition is not reducible to theory of mind (Clément et al., 2011; Kaufmann & Clément, 2007, 2014). Yet, although we are very sympathetic to their position, we have some reservations about their "institutional stance."

The foundation(s) of institutions

In our view, Jara-Ettinger and Dunham's main hypothesis is not controversial: in addition to navigating the social world by representing other minds, humans and, to some extent, other animals are able to "interpret social interactions in terms of role-based structures that constrain and regulate behavior via rule-like behavioral expectations" (p. 1). What is more problematic, however, is the claim that this ability is inherently institutional. The word *institution* appeared in Middle English before c. 1400, where it referred to "the action of establishing or founding (a system of government, a religious order, etc.)" (Etymonline). Put differently, institutions require that a social group collectively endows a physical substrate (an object, gesture, or action) with a specific meaning or function. To account for the constitutive backbones of institutional facts, and the collective recognition of "status function" they involve, the philosopher John Searle uses the formula "X counts as Y in context C" (Searle, 1995): a piece of paper X (such as a ten-dollar bill), for instance, counts as a bearer of a certain value Y in the context of economic exchange C. Institutions are therefore complex phenomena. From a *psychological* perspective, they require the capacity to recognize that objects, persons, or events can possess properties that are not intrinsic to them but are instead *added* through a process of collective attribution. From a *sociological* perspective, institutions are sustained by intricate cultural devices (juridical rules, administrative organizations, police forces, etc.) that "objectively" constrain and limit agents' own wills. Given the impersonal and external dimension of these constraints, it seems odd to restrict institutions to their "mentalistic" counterpart by defining them as a "type of cognitive representation consisting of a set of interconnected roles with a level of persistence and continuity" (sect. 4.1, para. 1).

Once they have defined their institutional stance, the authors attempt to imagine an evolutionary scenario that could explain how non-human animals might also use it to make sense of their conspecifics' behavior. To do so, they must assume that social animals live within "proto-institutional" systems. One of their examples is ants, which live in complex social structures with clearly defined roles; such a system could be considered 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" (sect 3.1, para. 3). Yet comparing insects' social systems to institutions falls into what we could call *retrospective determinism*: it reads the traits of our institutional systems back into their supposed evolutionary precursors, as though these systems had been moving toward our own all along.

From institutions to social forms

Avoiding this teleological trap while preserving the idea that social cognition cannot be reduced to mentalizing requires a cognitively and evolutionarily parsimonious understanding of what *social* means. If we follow the sociologist Georg Simmel (1910), the ontological building blocks of social life are elementary forms of reciprocal relationships: exchange, hierarchy, alliance, competition, nurturance, and so on. Those relationships are primarily recognizable through the observable behavior of the individuals embedded in them. For instance, the low posture and shifty gaze of a chimpanzee approaching a dominant member indicate – almost in a perceptual manner – its social position within the group. This position comes with a set of expectations about what is likely to happen next. Observers will, for example, be highly surprised when witnessing a sudden attack by a subordinate on a dominant individual (Slocombe et al., 2010). Children likewise develop early expectations about how individuals behave within hierarchical relationships (Charafeddine et al., 2015; Thomsen et al., 2011). The fact that such anticipations appear both in non-human primates and in young children suggests that the underlying cognitive mechanisms do not require a high degree of complexity. Interpreting these mechanisms as a rudimentary layer of a fully fledged "institutional stance" – one that presupposes representations of social structures built on internal models of "interconnected roles that regulate and guide how behavior unfolds over space and time" (sect. 4, para. 1) – does not seem evolutionarily plausible.

From the institutional stance to the social stance

To avoid moving backwards along the evolutionary path, it is better to follow the chronological order in which social structures – and the cognitive processes that support them – emerged (Kaufmann & Clément, 2014, 2024). This also preserves the arguments made by Jara-Ettinger and Dunham, provided that elementary forms of social life are not regarded as *proto-institutional*. Rather, they consist of reciprocal relations that social agents learn to process, progressively acquiring the ability to identify their essential properties – what we have called "naïve sociology."

We are therefore more optimistic than they are about the role of scripts in how we use our "social stance" to make sense of others' behaviors (Taylor et al., 2023). Before becoming institutional experts, social agents observe the interactions around them and extract regular patterns that structure their social world. This external reality can be considered as "objective" as physical reality and cannot be reduced to internal representations alone.

This view has at least two advantages. From an *epistemological* standpoint, it avoids "retrospective determinism" while capturing both evolutionary continuity and the novelty of human sociality. From a *cognitive* standpoint, it grounds the institutional stance in a more elementary social stance, such as naïve sociology, that enables social agents to navigate a shared realm of social interactions well before the emergence of the roles, statuses, and normative commitments characteristic of institutions.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors

Competing interests. None

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

References

Charafeddine, R., Mercier, H., Clément, F., Kaufmann, L., Berchtold, A., Reboul, A., & Van der Henst, J.-B. (2015). How preschoolers use cues of dominance to make sense of their social environment. *Journal of Cognition and Development*, *16*(4), 587–607.

Clément, F., Bernard, S., & Kaufmann, L. (2011). Social cognition is not reducible to theory of mind: When children use deontic rules to predict others' behaviors. *British Journal of Developmental Psychology*, *29*(4), 910–928.

Kaufmann, L., & Clément, F. (2007). How culture comes to mind: From social affordances to cultural analogies. *Intellectica*, *46/47*, 221–250.

Kaufmann, L., & Clément, F. (2014). Wired for society: Cognizing pathways to society and culture. *Topoi*, *33*, 459–475.

Kaufmann, L., & Clément, F. (2024). Wired for society? From ego-logy to eco-logy. *Behavioral and Brain Sciences*, *47*, e134.

Searle, J. R. (1995). *The construction of social reality*. Free Press.

Simmel, G. (1910). How is society possible? *American Journal of Sociology*, *16*(3), 372–391.

Slocombe, K. E., Kaller, T., Call, J., & Zuberbühler, K. (2010). Chimpanzees extract social information from agonistic screams. *PLoS One*, *5*, e11473.

Taylor, D., Gönül, G., Alexander, C., Züberbühler, K., Clément, F., & Glock, H.-J. (2023). Reading minds or reading scripts? De-intellectualising theory of mind. *Biological Reviews*, *98*(6), 2028–2048.

Thomsen, L., Frankenhuis, W. E., Ingold-Smith, M. C., & Carey, S. (2011). Big and mighty: Preverbal infants mentally represent social dominance. *Science*, *331*, 477–480.

How much is left for the theory-theory to explain?

Joseph Cruz*

Philosophy and Cognitive Science, Williams College, Williamstown, USA
joseph.l.cruz@williams.edu

*Corresponding author.

doi:10.1017/S0140525X26105627, e376

Abstract

Jara-Ettinger and Dunham treat the "mentalistic stance" as theory-theory, leaving the institutional stance vulnerable to a familiar reply: the tacit theory can be context-tuned to fit the surprise cases. I argue that the institutional stance is best motivated by pairing it with simulation-based mindreading and scripts, which together reduce theory-theory's distinctive explanatory role and preserve the authors' appeal to computational economy.

Jara-Ettinger and Dunham too quickly assimilate the mentalistic stance to theory-like mindreading. They therefore leave open their institutional stance to a familiar reply from theory-theorists, namely, that the target phenomena can be recovered by enriching the tacit theory. A more principled rebuttal of the theory-theory is required to motivate the institutional stance, and that, in turn, requires an alternative mentalist capacity. Simulation theory is one such alternative.

Simulation-based mindreading offers a mentalistic but non-theory-like alternative to theory-theory. To explain and predict others' behavior, one imaginatively places oneself in the other's epistemic and motivational situation. Much mental-state attribution thus proceeds by redeploying one's own practical and affective resources (Goldman, 1992; 2006; Gordon, 1986; Gordon & Cruz, 2003; Heal, 1986; Todd & Tamir, 2025).

This matters here because the "surprise" cases the authors emphasize invite a standard theory-theory reply. The maneuver is to accommodate the data by updating the theory to make it more granular and reflect changes in context. The Kate and Eric vignette (sect. 1) illustrates this. When Kate asks her workplace assistant Eric to get her snacks after running into him at a movie theater, Jara-Ettinger and Dunham argue that our surprise reflects reliance on the institutional stance. On their view, this is an interpretive framework appropriate to the workplace that does not fit the informal setting, so we are surprised when we observe the institutional behaviors manifested. Since the theory-theory attribution of Kate and Eric's beliefs and desires can be readily carried over to the informal context, the theory-theorist allegedly has no explanation for our surprise.

A theory-theorist can readily offer a reply. Nothing in the theory-theory requires attributing the same belief–desire profile across contexts. The theory can be tuned so that work and leisure contexts carry different default expectations. Contemporary defenses of theory-theory often stress this flexibility, especially in cases that are unexpected or involve idiosyncratic agents (Jara-Ettinger, 2019; Jern, Lucas, & Kemp, 2017). Kate can be treated as just such an agent, which would motivate a more complex attribution of her beliefs and desires. For example, she may have confused beliefs about the boundaries of an assistant's obligations. On this approach, surprise arises because Kate behaves as if the work context belief–desire profile applies in a leisure setting.

Both the institutional stance and the theory-theory have an explanation for the data. A way out of this standoff is to appreciate that the institutional stance is attractive because it explains the surprise without relying on unconstrained enrichment of a mentalistic theory. More generally, the motivation for the institutional stance consists in highlighting behavioral phenomena that resist, or are less parsimoniously explained, by the mentalistic stance. But the theory-theory *always* has resources to accommodate the instances that motivate the institutional stance. This flexibility is precisely why surprise cases do not, by themselves, force a choice between theory-like mentalizing and the institutional stance. Jara-Ettinger and Dunham effectively concede as much when they note that a mentalistic interpretation is available (e.g., sect. 6.3.2), but that it requires increasingly complex and intuitively implausible mental-state attributions. They contrast this with the institutional stance, which makes the same interactions appear seamless by appeal to role-based expectations. The institutional stance is therefore appealing not because it refutes the theory-theory on its own terms but because it explains the surprise without relying on open-ended enrichment of a tacit mentalistic theory. The worry about the theory-theory is not complexity per se, but case-by-case accommodation without any independent constraints on the theory. And there is a ready alternative to the theory-theory version of mentalism in the simulation theory.

Accordingly, I propose a division of labor. In structured contexts shaped by roles and norms, the institutional stance plus script knowledge does much of the explanatory work that is often assigned to theory-like mentalizing. In less structured contexts, simulation-based mindreading supports flexible and agent-specific interpretation. On this picture, the institutional stance does not primarily compete with mentalism as such. It mainly competes

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

with a theory-like construal of mentalism. The range of cases in which theory-theory is the most parsimonious option is therefore considerably narrowed. These are cases in which behavior is genuinely novel or nonstandard and is not readily explained by folk physics, as when someone trips and stumbles, not captured by a familiar script, and not made predictable by institutional role structure. They are also cases in which simulation is unreliable because the target is dissimilar, atypical, or otherwise not well modeled by first-person redeployment.

A second motivation Jara-Ettinger and Dunham offer is computational economy. They argue that role-based reasoning can reduce the cognitive costs and uncertainty of mentalizing and can reduce the online computational burden of tracking beliefs and desires in real time (sect. 6.3.1). This is a useful point, but it further highlights what is missing from their picture of the mentalistic stance. Simulationist accounts were motivated in part by the same computational contrast, since simulation predicts others by redeploying first-person resources rather than by applying an increasingly complex theory (Cruz, 1998). The institutional stance and simulation theory therefore look like parallel forms of cognitive offloading that apply in different domains.

Jara-Ettinger and Dunham suggest that mentalism combined with the institutional stance yields richer possibilities for representing and creating institutional structure (sect. 6.3.3). Simulation-based mindreading can play that same enriching role, since it supports agent-specific expectations about how particular individuals will interpret and negotiate institutional roles, without committing us to the theory-theory.

Once we acknowledge how readily the theory-theory can be modified, its treatment of the authors' surprise cases becomes difficult to state without an overly permissive enrichment strategy. If the theory can be freely revised to match each surprise, it no longer looks like the kind of stable, learnable resource that could guide online interpretation. The institutional stance treats the same phenomena by deploying independently learnable representations of social structure in terms of roles and norms, rather than by case-by-case elaboration of an underlying theory. It looks stronger still once simulation-based mindreading is counted as a mentalistic alternative.

Financial support. None.

Competing interests. The author declares no competing interests.

References

Cruz, J. (1998). Mindreading: Mental state ascription & cognitive architecture. *Mind & Language*, *13*(3), 323–340.

Cruz, J., & Gordon, R. (2003). Simulation theory. In L. Nadel (Ed.), *Encyclopedia of cognitive science* (vol. *3*, pp. 9–13). Nature Publishing Group.

Goldman, A. (1992). In defense of simulation theory. *Mind and Language*, *7*(1–2), 104–119.

Goldman, A. (2006). *Simulating minds: The philosophy, psychology, and neuroscience of mindreading*. Oxford UP.

Gordon, R. (1986). Folk psychology as simulation. *Mind and Language 1*(2), 158–171

Heal, J, (1986). Replication and functionalism. In J. Butterfield (Ed.), *Language, mind and logic* (pp. 135–150) Cambridge UP.

Jara-Ettinger, J. (2019). Theory of mind as inverse reinforcement learning. *Current Opinion in Behavioral Sciences*, *29*, 105–110.

Jern, A., Lucas, C., & Kemp, C. (2017). People learn other people's preferences through inverse decision-making. *Cognition*, *168*, 46–64.

Todd, A. R., & Tamir, D. I. (2025). Simulation-driven mentalizing facilitates projection and introjection. *Trends in Cognitive Sciences*. Advance online publication. https://doi.org/10.1016/j.tics.2025.10.013

Sacred values as role-constitutive norms in social groups

Katrina Fincher*

Psychology, New School for Social Research: The New School for Social Research, New York, USA
fincherk@newschool.edu

*Corresponding author.

doi:10.1017/S0140525X26105603, e377

Abstract

Jara-Ettinger and Dunham propose that institutional cognition tracks roles and their normative expectations. I argue that decentralized roles lacking formal authority require an additional mechanism to stabilize legitimacy. Sacred values – nonnegotiable commitments that impose costly obligations – function as role-constitutive norms that signal intrinsic devotion, thereby determining legitimate role occupancy and preventing free riding on group identities.

Jara-Ettinger and Dunham propose that an institutional stance, reasoning about behavior in terms of roles and their normative expectations, constitutes a core system of human social cognition distinct from mental-state reasoning. Many of the roles they discuss, however, appear to rely on a legal – rational scaffolding, leaving others, particularly those associated with group identities, unexplained. Identities such as "American" or "scientist" are not merely positions defined by codified rules but also roles embedded in norm-governed systems that define entitlement, moral standing, and membership boundaries. This raises the question of how decentralized groups determine who legitimately occupies a role and how free riding on group identity is prevented.

I propose that such roles are often stabilized through sacred values – nonnegotiable commitments that signal intrinsic devotion and define the normative foundations of rightful role occupancy. In such cases, commitments function not merely as behavioral guidelines but as role-constitutive norms that signal intrinsic commitment and thereby define legitimate role occupancy. These cases suggest that institutional cognition must represent two distinct classes of norms: compensatory norms governing role performance and role-constitutive norms defining legitimate role occupancy.

Jara-Ettinger and Dunham's conception of roles frequently presupposes a legal–rational scaffolding. As a result, role occupancy is stabilized through formal structures that define who is entitled to act and how violations are repaired. In such systems, legitimacy is grounded in formally codified rules and procedural regularity (Beetham, 2013; Weber, 1922/2019): a barista may negotiate payment terms and a janitor's entry hours without jeopardizing their role. Although Jara-Ettinger and Dunham's framework allows that institutional reasoning also applies to

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

informal roles such as group membership, it does not specify how the legitimacy of role occupancy is determined in the absence of formal authority.

In Weber's (1922/2019) terms, such roles more closely resemble traditional forms of authority, in which legitimacy rests not on procedural rules but on shared belief and customary recognition. Berger and Luckmann (1966) similarly argue that shared beliefs become institutionalized when they are widely recognized as legitimate foundations of social roles. In decentralized groups lacking bureaucratic enforcement, legitimacy must therefore be sustained through mechanisms that stabilize shared commitments and transform contingent norms into obligations.

Sacred values, commitments treated as absolute and resistant to instrumental tradeoffs, may provide one such mechanism. Unlike ordinary norms, sacred values resist tradeoffs and motivate costly, even self-sacrificial commitments. Across cultural and political contexts, individuals treat certain principles, such as land, sovereignty, or religious sites, as nonnegotiable, refusing material compromises even when they would produce substantial benefits (Atran & Ginges, 2012; Dehghani et al., 2010; Ginges & Atran, 2013; Tetlock et al., 2000). Empirical work shows that when such values become fused with group identity, individuals may express willingness to incur extreme personal costs, including fighting or dying to defend them (Hamid et al., 2019). Because these commitments demand sacrifice, they function as credible signals of legitimate role occupancy (Tetlock et al., 2000; Tetlock, Mellers & Scoblic, 2017). By making commitment visible and costly, sacred values stabilize expectations about legitimate role occupancy in decentralized communities.

From the perspective of the institutional stance, such commitments function as role-constitutive norms that help determine who legitimately occupies a role. Their function is not merely to regulate behavior but also to establish legitimate role occupancy itself. Because sacred values are non-compensatory – resistant to material substitution or procedural repair – their endorsement signals intrinsic commitment rather than strategic compliance. Violations, therefore, call an individual's standing within the role into question rather than simply triggering a sanction.

This dynamic can be seen in communities whose identities are organized around sacred values that demand absolute commitment. Identifying as a Quaker, for instance, does not depend solely on formal membership in the Religious Society of Friends or compliance with institutional rules. Rather, the identity is anchored in a central sacred value: the belief that every individual possesses an "inner light." Because this commitment imposes an obligation to extend forgiveness even in the face of serious wrongdoing, failing to uphold it calls into question one's claim to the identity – and thus one's standing within the role.

A similar dynamic operates in non-religious contexts. Although scientists may work in different institutions or disciplines and frequently disagree about particular theories, the identity of "scientist" is sustained by a shared commitment to the authority of the scientific method as a pathway to truth. When this epistemic commitment is violated – for example, through fabrication or deliberate disregard for evidence – the issue is not merely procedural misconduct. Such violations call into question whether the individual legitimately occupies the role of scientist at all.

Taken together, these perspectives suggest that decentralized institutions stabilize legitimacy through commitments treated as intrinsic rather than negotiable. Formal institutions achieve this through procedural legitimacy, whereas decentralized groups rely on costly commitments that signal devotion to shared principles. Social groups, therefore, reveal that the institutional stance must represent not only norms governing how roles are performed, but also norms determining who is entitled to occupy them.

Acknowledgements. None.

Financial support. No funding was used.

Competing interests. The author declares no conflicts of interest.

References

Atran, S., & Ginges, J. (2012). Religious and sacred imperatives in human conflict. *Science*, *336*(6083), 855–857.

Beetham, D. (2013). *The legitimation of power*. Bloomsbury Publishing.

Berger, P. L., & Luckmann, T. (1966). *The social construction of reality: A treatise treatise in the sociology of knowledge*. Penguin Books.

Dehghani, M., Atran, S., Iliev, R., Sachdeva, S., Medin, D., & Ginges, J. (2010). Sacred values and conflict over Iran's nuclear program. *Judgment and Decision Making*, *5*(7), 540–546.

Ginges, J., & Atran, S. (2013). Sacred values and cultural conflict. *Advances in Culture and Psychology*, *4*, 273–301.

Hamid, N., Pretus, C., Atran, S., Crockett, M. J., Ginges, J., Sheikh, H. & Vilarroya, O. (2019). Neuroimaging 'will to fight'for sacred values: An empirical case study with supporters of an Al Qaeda associate. *Royal Society open science*, *6*(6), 1–5.

Tetlock, P. E., Kristel, O. V., Elson, S. B., Green, M. C., & Lerner, J. S. (2000). The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*, *78*(5), 853.

Tetlock, P. E., Mellers, B. A., & Scoblic, J. P. (2017). Sacred versus pseudo-sacred values: How people cope with taboo trade-offs. *American Economic Review*, *107*(5), 96–99.

Weber, M. (1922/2019). *Economy and society: A new translation*. Harvard University Press.

There are no institutions without rules commentary on Jara-Ettinger & Dunham: The institutional stance

Ana Gantman[a,b,c*], Jordan Wylie[d] and Brendan Fee[b]

[a]Psychology, Brooklyn College, Brooklyn, USA; [b]Psychology, CUNY Graduate Center, New York, USA; [c]Philosophy, CUNY Graduate Center, New York, USA and [d]Psychology, Cornell University, Ithaca, USA

apgantman@gmail.com
bfee@gc.cuny.edu
jcw355@cornell.edu

*Corresponding author.

doi:10.1017/S0140525X26105639, e378

Abstract

The authors contend that the "institutional stance" paves the way for uniquely human institutions. However, it is formalized rules that are the integral ingredients for modern institutions and

sufficient for understanding people within them. We suggest that formalized rules are a technology for norm conformity that codify, legitimize, enforce, and explain permissible behavior across roles and contexts – and literally establish institutions.

The target article claims that people rely on representations of institutional structures and roles embedded within them to understand others' behavior. The authors suggest that these representations pave the way for the making of all kinds of social institutions, from book clubs to companies. However, this emphasis on the *understanding of roles* is missing a key facet of modern human institutions, without which we cannot understand their origins – the articulation of formal *rules*.

The creation of powerful, recursive, and ever-expanding (see Klotz, 2021) human institutions (e.g., governments, universities, recreational bowling leagues) cannot be explained by merely scaling up our understanding of social roles. Institutions (from the Latin *instituere* "to set up") are products of deliberate efforts to structure society with rules. Rules are articulated, legitimized, codified, and enforced directives that regulate permissible behavior. Rules are, in a sense, a technology, deliberately set up to facilitate norm conformity.

Norms are well-established as essential for large-scale human cooperation (Henrich & Muthukrishna, 2021), are appreciated by children from a young age (Weston & Turiel, 1980), and have a proto-form in animals that does not require the mentalistic stance (Andrews, 2009; 2012). Institutional rules are widely known, legitimized and enforceable by some authority (e.g., the police, the crown, "the people"), written down, and as a result, are perfectly transmittable across people and generations. Rules are a technology fashioned from the raw materials of norms.

Yet, rules and norms are ultimately different in kind. Norms are often underspecified, and their transmission is "lossy" (Heyes, 2018), changing as they are passed across time, contexts, and cultures. In contrast, rules are formalized and come with prespecified punishments (Wylie & Gantman, 2023), and usually require deliberate effort to change. People conform to the norms only of relevant reference groups (Paluck & Prentice, 2020; Sherif & Sherif, 1953), but expect that the rules apply equally to everyone in their jurisdiction (Hannikainen et al., 2021; Wylie & Gantman, 2023; Wylie et al., 2026). Rules can alter norms (e.g., Supreme Court ruling in favor of gay marriage; Tankard & Paluck, 2017), and rules and norms can suggest people behave in opposite directions (e.g., jaywalking; Wylie & Gantman, 2023).

Rules, then, are a very successful human innovation and are constitutive of modern institutions. This technology has been used at least as early as the earliest human writing, which details the coordination of sharing resources and repayment of debts (Graeber, 2011; 2015). Children create their own rules during play (Castle, 1998), and can distinguish between the spirit and the letter of the law by around age four (Bregant, Wellbery, & Shaw, 2019). By five, they can even find loopholes in the space between their meanings (Bridgers et al., 2025). Not until around seven years old, however, do children recognize that some actions are available to individuals only because of their status (i.e., role) within a particular institution (Noyes, Dunham, & Keil, 2023). The capacity and proclivity to engage with rules are consistent across time and emerge early in the lifespan.

The benefits of rules-tech are substantial in streamlining institutional administration. Rules facilitate predictable interactions among large groups of strangers (Hadfield & Weingast, 2012), even when they cannot communicate with each other (Hannikainen et al., 2022). In this way, rules reduce the cognitive burden of human interaction. Rules allow people to skip reasoning – one can rely on a predetermined course of action for what to do and what not to do; other considerations (e.g., moral, ethical, efficiency) can be brushed aside. Indeed, research suggests people feel a diminished sense of agency when following rules (Caspar, Cleeremans, & Haggard, 2018) and experience cognitive conflict when breaking them (Pfister et al., 2019).

Critically, rules are also extremely useful for understanding the behavior of others, as most people spend much of their time abiding by the law. Rule adherence can explain why people submit to taking their shoes off at the airport, returning their library books, driving on the left (or right) side of the road, paying their taxes, and avoiding plagiarism in their homework assignments. Rules can perhaps best describe people's behavior within institutions, when their assigned roles take precedence over their personhood. People who join institutions are explicitly trained (e.g., given a company "bible") on exactly what to do and how to do it. The obligation to comply with rules readily explains much coordinated behavior, and appreciating their importance complicates the apparent throughline the authors propose from worker bees to middle management. In sum, *rules* enable large-scale coordination, have roots in animal minds, are readily appreciated and adapted by children, and provide public, accessible attributions for behavior. Formal rules, then, do much of what the institutional stance is purported to do – no new system of social understanding required. The authors suggest we need "institutional norms," but we already have norms bounded by social roles (e.g., Eagly & Wood, 2012); what we do need is a cognitive science of formalized rules.

When we understand rules as technology, a kind of cultural artifact (Franssen, Lokhorst, & Van de Poel, 2009), we can evaluate what kinds of social configurations they make more or less likely (Winner, 1980). Rules create order, a turbo-charged way of controlling others, and give rise to inhumane forms of punishment (e.g., prison; see Wylie et al., 2025). Rules can consolidate the power of the rule-makers, serving the interests of some to the detriment of others (Lukina, 2023; Wylie & Gantman, 2023). Rules make institutions that make people ruleable.

Finally, the authors invoke Arendt's coverage of Eichmann's trial and legal defense of "just following orders" as a downside to taking the institutional stance. But this misunderstands Arendt and the power of rules. When she describes the Nazi bureaucrat as "thoughtless," she does not mean that *we* understand *him* through the institutional stance, but rather that people embedded in institutions follow rules given to them without seriously considering the actions they mandate. This is not a downside of the institutional stance but a downside of institutional *rules*. It's not the roles we play, it's the rules we follow.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. The authors declare none.

References

Andrews, K. (2009). Understanding norms without a theory of mind. *Inquiry*, *52*(5), 433–448.

Andrews, K. (2012). *Do apes read minds?: Toward a new folk psychology*. MIT Press.

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

Bregant, J., Wellbery, I., & Shaw, A. (2019). Crime but not punishment? Children are more lenient toward rule-breaking when the "spirit of the law" is unbroken. *Journal of Experimental Child Psychology*, *178*, 266–282.
Bridgers, S., Parece, K., Iwasaki, I., Broski, A., Schulz, L., & Ullman, T. (2025). Learning loopholes: The development of intentional misunderstandings in children. *Child Development*, *96*(3), 1066–1087.
Caspar, E. A., Cleeremans, A., & Haggard, P. (2018). Only giving orders? An experimental study of the sense of agency when giving or receiving commands. *PloS one*, *13*(9), e0204027.
Castle, K. (1998). Children's rule knowledge in invented games. *Journal of Research in Childhood Education*, *12*(2), 197–209. https://doi.org/10.1080/02568549809594884
Eagly, A. H., & Wood, W. (2012). Social role theory. *Handbook of Theories of Social Psychology*, *2*(9), 458–476.
Franssen, M., Lokhorst, G. J., & Van de Poel, I. (2009). Philosophy of technology. *Stanford Encyclopedia of Philosophy*.
Graeber, D. (2011). *Debt: The first 5,000 years*. Melville House.
Graeber, D. (2015). *The utopia of rules: On technology, stupidity, and the secret joys of bureaucracy*. Melville House.
Hadfield, G. K., & Weingast, B. R. (2012). What is law? A coordination model of the characteristics of legal order. *Journal of Legal Analysis*, *4*(2), 471–514. https://doi.org/10.1093/jla/las008
Hannikainen, I. R., Tobia, K. P., de Almeida, G. d. F. C. F., Donelson, R., Dranseika, V., Kneer, M., Strohmaier, N., Bystranowski, P., Dolinina, K., Janik, B., Keo, S., Lauraitytė, E., Liefgreen, A., Próchnicki, M., Rosas, A. & Struchiner, N. (2021). Are There Cross-Cultural Legal Principles? Modal Reasoning Uncovers Procedural Constraints on Law. *Cognitive Science*, *45*, e13024. https://doi.org/10.1111/cogs.13024
Hannikainen, I. R., Tobia, K. P., de Almeida, G. D. F. C. F., Struchiner, N., Kneer, M., Bystranowski, P., Dranseika, V., Strohmaier, N., Bensinger, S., Dolinina, K., Janik, B., Lauraitytė, E., Laakasuo, M., Liefgreen, A., Neiders, I., Próchnicki, M., Rosas, A., Sundvall, J., & Žuradzki, T. (2022). Coordination and expertise foster legal textualism. *Proceedings of the National Academy of Sciences of the United States of America*, *119*(44), e2206531119. https://doi.org/10.1073/pnas.2206531119
Henrich, J., & Muthukrishna, M. (2021). The origins and psychology of human cooperation. *Annual Review of Psychology*, *72*(1), 207–240.
Heyes, C. (2018). *Cognitive gadgets: The cultural evolution of thinking*. Harvard University Press.
Klotz, L. (2021). *Subtract: The untapped science of less*. Flatiron Books.
Lukina, A.. (2023). Chapter 38: The problem of evil law. In M. Tushnet & D. Kochenov (Eds.), *Research handbook on the politics of constitutional law*. Edward ElgarPublishing. https://doi.org/10.4337/9781839101649.00048
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
Pfister, R., Wirth, R., Weller, L., Foerster, A., & Schwarz, K. A. (2019). Taking shortcuts: Cognitive conflict during motivated rule-breaking. *Journal of Economic Psychology*, *71*, 138–147.
Prentice, D., & Paluck, E. L. (2020). Engineering social change using social norms: Lessons from the study of collective action. *Current Opinion in Psychology*, *35*, 138–142.
Sherif, M., & Sherif, C. W. (1953). *Groups in harmony and tension; An integration of studies of intergroup relations*. Harper & Brothers.
Tankard, M. E., & Paluck, E. L. (2017). The effect of a Supreme Court decision regarding gay marriage on social norms and personal attitudes. *Psychological Science*, *28*(9), 1334–1344.
Weston, D. R., & Turiel, E. (1980). Act–rule relations: Children's concepts of social rules. *Developmental Psychology*, *16*(5), 417.
Winner, L. (1980). Do artifacts have politics? *Daedalus*, *109*(1), 121–136. https://www.jstor.org/stable/20024652
Wylie, J., Chiu, C. P. Y., Dakin, N. M., Cunningham, W., & Gantman, A. (2025). The psychology of state punishment. *European Journal of Social Psychology*, *55*(2), 251–258. https://doi.org/10.1002/ejsp.3147
Wylie, J., & Gantman, A. (2023). Doesn't everybody jaywalk? On codified rules that are seldom followed and selectively punished. *Cognition*, *231*, 105323. https://doi.org/10.1016/j.cognition.2022.105323
Wylie, J., Milless, K. L., Sciarappo, J., & Gantman, A. (2026). The biased enforcement of rarely followed rules. *Personality and Social Psychology Bulletin*, *52*(1), 105–118.

Social prediction requires enriching theories of institutions with theories of mind

Ido Halevy[a] and Nir Halevy[b]*

[a]Stanford University, USA and [b]Stanford Graduate School of Business, USA
ihalevy@stanford.edu
nhalevy@stanford.edu

*Corresponding author.

doi:10.1017/S0140525X26105536, e379

Abstract

We address two important questions that the target article leaves largely open. First, we explain *why* institutional representations on their own struggle to reduce uncertainty in social interactions sufficiently and hence require mental state inferences to calibrate predictions. Second, we explain *how* supplementing theories of institutions with theories of mind provides the richness necessary to refine social predictions.

Navigating the social world requires predicting others' behavior. The target article posits that abstract institutional representations increase predictability by delineating the formal rules and social norms that govern the behavior of different role holders within organizational systems. It claims further that individuals form expectations concerning others' behavior by combining representations of others' roles in the social structure (e.g., boss, teacher, and parent) and representations of others' mental states (e.g., motives, beliefs, and feelings). However, the target article leaves open the questions of *why* and *how* individuals synthesize predictions derived from institutional and mental state representations.

Why individuals enrich institutional representations with mental state representations

The claim that institutional environments – with their roles, rules, and norms – enable calibrated expectations about human behavior implicitly assumes that institutional environments are strong situations that restrict variance in human behavior (Li et al., 2025). However, not all institutional environments are strong situations. Tight environments feature well-defined norms and enforce them strictly, whereas loose environments enable individuals to have many degrees of freedom when choosing their actions (Gelfand, Harrington, & Jackson, 2017). Supervisors and subordinates may rely on their institutional representations to predict each other's behavior in tight environments but will likely need additional information about others' mental states to form accurate predictions in weak situations and loose environments (e.g., are they motivated to do a good job?).

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

Further, individuals may mentally represent the same institution differently. For example, employees may represent their workplace as a hierarchy or as a network, with distinct consequences for perceived norms (Moyal, Tan, & Halevy, 2023). This variability in mental representations of institutions means that a theory of the institutional context (e.g., workplace) may not be enough to predict others' behavior. Doing so requires understanding how others mentally represent the institution.

Within organizations, many employees experience role ambiguity and role conflict, which capture lacking clear information about their role demands and facing contradictory role-based expectations, respectively (Schmidt et al., 2014). Role ambiguity and role conflict may increase variability in in-role behavior. Even when expectations are clear, employees may engage in deviant behavior, break rules, and innovate. By falling short of expectations, exceeding expectations, or violating expectations in other ways, coworkers may surprise each other (Halevy, Miclau, & Lee, 2026). The institutional stance is insufficient to reliably predict the behavior of role holders who do not behave as expected.

Consider the case of convenience store robberies. Research shows most people expect to find clerks behind the register when they enter a convenience store, even though clerks sometimes stock shelves, mop floors, and take bathroom breaks. Convenience store robberies that begin while clerks violate robbers' expectations by being away from the cash register end up being significantly more violent (DeCelles et al., 2022). Although the institution (convenience store) and roles (robber, clerk) remain the same irrespective of the clerk's position in the store, this deviation from a role-based expectation shifts the interaction toward greater violence. As this example illustrates, knowing the institutional context and people's roles in it may be insufficient for predicting how an interaction will unfold.

One institution mentioned frequently in the target article is hierarchy. There is considerable variation among hierarchies, not only across cultures and historical periods (Muthukrishna, Henrich, & Slingerland, 2021) but also within the same society and historical period. Hierarchies can be relatively steep or flat, well-functioning or dysfunctional, based on dominance or prestige, and so on (Cheng et al., 2013; Yu et al., 2019). Further, the same position in the hierarchy may entail different duties and responsibilities in different institutional contexts (e.g., a CEO in a private company, a public company, and a nonprofit organization). The fact that similar institutions often give rise to very different normative systems (i.e., organizational cultures) is evident also from the ubiquity of cultural clashes following organizational mergers and acquisitions (Weber & Camerer, 2003). Hence, knowledge of the institutional context and individuals' roles may be insufficient to predict others' actions.

How theories of mind help calibrate predictions derived from institutional theories

Our simple solution extends Figure 1c in the target article. We propose that people regularly integrate institutional and mental state representations in the same way they enrich nouns with adjectives. Because role representations are insufficient to form accurate predictions about others' behavior, people supplement roles with descriptors that capture others' mental states. Parents to multiple children, or instructors who teach multiple students, know that the roles "child" and "student" are insufficient to describe how different individuals will behave in the same role. This is why we have numerous adjectives that add meaningful nuance to social roles: a *helpful* child, a *curious* student, a *messy* roommate, a *nasty* neighbor, a *kind* boss. In all these cases, the mentalistic stance meaningfully clarifies what to expect from the role holder.

Importantly, the adjective does not have to represent an enduring personality trait. It can capture a transient state that helps calibrate expectations. Consider the role of a barista, mentioned repeatedly in the target article. Keeping the role holder (barista) and institution (café) constant, we may form different expectations depending on whether the barista seems overworked and tired or well-rested and energized on a given shift. Their expectations concerning our tipping behavior may likewise depend on their perceptions of our mood that day.

In sum, social prediction requires more than knowledge of institutional structure and people's roles in it. It requires mental state inferences. Trust between two romantic partners or business partners cannot be based only on roles and norms; it must incorporate beliefs about others' feelings, intentions, and capabilities. The same logic applies beyond dyads and small teams. The large-scale coordination and cooperation that are necessary to provide public goods and solve collective action problems need more than a theory of institutions; they require beliefs about others' mental states. Hence, important as the institutional stance is for forming expectations, understanding how people enrich institutional models with mental state inferences is essential for a complete and coherent theory of social prediction.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. None.

References

Cheng, J. T., Tracy, J. L., Foulsham, T., Kingstone, A., & Henrich, J. (2013). Two ways to the top: Evidence that dominance and prestige are distinct yet viable avenues to social rank and influence. *Journal of Personality and Social Psychology*, *104*(1), 103–125.

DeCelles, K. A., Kouchaki, M., & Halevy, N. (2022). Unexpected employee location is associated with injury during robberies. *Proceedings of the National Academy of Sciences*, *119*(39), e2200026119.

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.

Halevy, N., Miclau, E., & Lee, S. (2026). Abstractness, concreteness, and strategic surprises. *American Psychologist*. Available online first. https://doi.org/10.1037/amp0001679.

Li, R., Balliet, D., Thielmann, I., & de Vries, R. E. (2025). Revisiting situational strength: Do strong situations restrict variance in behaviors? *Journal of Personality and Social Psychology*, *129*(3), 551–575.

Moyal, A., Tan, J. C. Y., & Halevy, N. (2023). Brokering in hierarchies versus networks: How organizational structure shapes social relations. *Journal of Experimental Social Psychology*, *109*, 104515.

Muthukrishna, M., Henrich, J., & Slingerland, E. (2021). Psychology as a historical science. *Annual Review of Psychology*, *72*(1), 717–749.

Schmidt, S., Roesler, U., Kusserow, T., & Rau, R. (2014). Uncertainty in the workplace: Examining role ambiguity and role conflict, and their link to depression—a meta-analysis. *European Journal of Work and Organizational Psychology*, *23*(1), 91–106.

Weber, R. A., & Camerer, C. F. (2003). Cultural conflict and merger failure: An experimental approach. *Management Science*, *49*(4), 400–415.

Yu, S., Greer, L. L., Halevy, N., & Van Bunderen, L. (2019). On ladders and pyramids: Hierarchy's shape determines relationships and performance in groups. *Personality and Social Psychology Bulletin*, *45*(12), 1717–1733.

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

Relationship-bound roles and role-filler independence in social cognition

Hannah Hok[a,b] and Ashley J. Thomas[b]*

[a]Brain and Cognitive Sciences, Massachusetts Institute of Technology, USA and [b]Psychology, Harvard University, Cambridge, USA
hok1@mit.edu
athomas@g.harvard.edu

*Corresponding author.

doi:10.1017/S0140525X2610572X, e380

Abstract

The target article proposes an evolutionarily old "institutional stance," where people make sense of others' behavior through roles independent of individuals. We argue that representations of such role-filler-independent roles are likely not primitives. We offer an alternative: relationship-bound role representations that encode particular individuals. Role-filler independence may instead be a cultural tool that addresses coordination challenges in large-scale groups.

In many contemporary societies, some social roles and the behaviors they afford depend on the institutions that create them, regardless of the individuals who occupy them. A judge, a referee, or a cashier is treated primarily as a position within a larger institutional structure: whoever occupies the position is expected to behave in role-appropriate ways. The target article proposes an evolutionarily old "institutional stance," according to which people make sense of others' behavior based on institutional roles. We agree that social cognition is not solely, nor primarily, grounded in mental state reasoning and that an institutional stance captures an important feature of contemporary adult social cognition. However, we argue that role-filler independence – the intuition that anyone can occupy a role – is not an evolutionarily ancient default but a widely available innovation that helps solve coordination problems in large-scale societies. We connect this argument broadly to the implications of work on impartiality and suggest that another candidate for the foundation of social reasoning could instead be relationship cognition: abstract representations of relationships between particular individuals that include *relationship-bound* roles.

Roles can be thought of in *at least* two ways that are non-mutually exclusive: (1) role-filler-independent roles whose occupants are interchangeable and (2) roles embedded within interpersonal relationships (relationship-bound roles). Many of the authors' examples – baristas, janitors, and cashiers – are contemporary jobs. The behaviors afforded by these roles are shaped by institutions situated in large-scale economic systems, which we argue are particularly salient examples of role-filler independent roles. Take their movie-theater example, where a boss no longer expects their employee to fetch things for them when the employee is "off the clock." While intuitive to many readers, this example may not be intuitive to infants or young children or even people living in different economic systems. Indeed, even US adults likely do not "forget" their boss–employee relationship when off the clock. Alternatively, roles can be represented *within* interpersonal relationships. For example, people have expectations about how mothers (one role) behave toward their children (another role). However, while a child may have multiple caregivers whose roles have overlapping functions, children likely do not think of their caregivers as interchangeable. If a child's father leaves or dies, he cannot simply be replaced with another caregiver without changing the representation. In this way, relational representations often include particular people in the relation. These two ways in which roles can function are not mutually exclusive. Even highly institutional roles can become relationship-bound (e.g., building a relationship with a particular barista), and conversely, relational roles (e.g., mother) can be conceived of abstractly in the absence of particular people. We agree that infants have abstract representations of relationship categories, which allow for generalization. However, we argue that they work with representations of particular people. Research demonstrates that infants track structured social relations, including dominance hierarchies, asymmetric interactions, and stable patterns of affiliation. However, these findings do not uniquely support representations of abstract roles with role-filler independence, the way the institutional stance may suggest. For example, dominance hierarchies in other species may arise from patterns of interactions among particular individuals and their particular ties, rather than abstract dominant and subordinate role slots. Indeed, children and infants generalize dominance relations asymmetrically, generalizing dominance more readily than subordination (e.g., the role of the "dominant" is not interchangeable; Mascaro et al., 2023; Hok et al., 2024). Moreover, infants expect different social interactions when social partners change (Kudrnova, Spelke, & Thomas, 2024; Pepe et al., 2025; Thomas, Saxe, & Spelke, 2022). There is no evidence that infants change expectations of how particular dyads will interact when contexts change. While it is possible that infants can recognize institutions and would expect similar interactions even when a social partner is swapped out within the same institutional role, there is no such evidence. Thus, while existing infant research does not necessarily exclude the possibility that infants assume role-filler independence, the data can also be explained by representations that include particular individuals. While we think it is unlikely that the primitive foundation of social cognition includes role-filler independence, we agree that it plays an important role in shaping contemporary human adult cognition, which may allow for many new ways to relate, both positive and negative. For example, role-filler independence is closely related to other social tools such as impartiality. Thus, role-filler independence may be a representational format (focusing on the role rather than the individual). Impartiality may be a normative principle *that uses* this representational format. These types of social tools may emerge recurrently across historical and cultural contexts under conditions that require larger-scale coordination. In large-scale social environments – contemporary markets, bureaucracies, or contexts with many interactions with unfamiliar individuals – it is infeasible to maintain rich representations for every social partner. Institutional roles provide a solution by allowing expectations to generalize across individuals occupying the same functional position, rather than indexing across particular individuals and their relationships (for a similar argument on morality, see DeScioli & Kurzban, 2013). Consistent with this view, both adults and older children, at least in the USA, expect formal rules to emerge preferentially in large groups, where identity-based relationship tracking may be more difficult

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

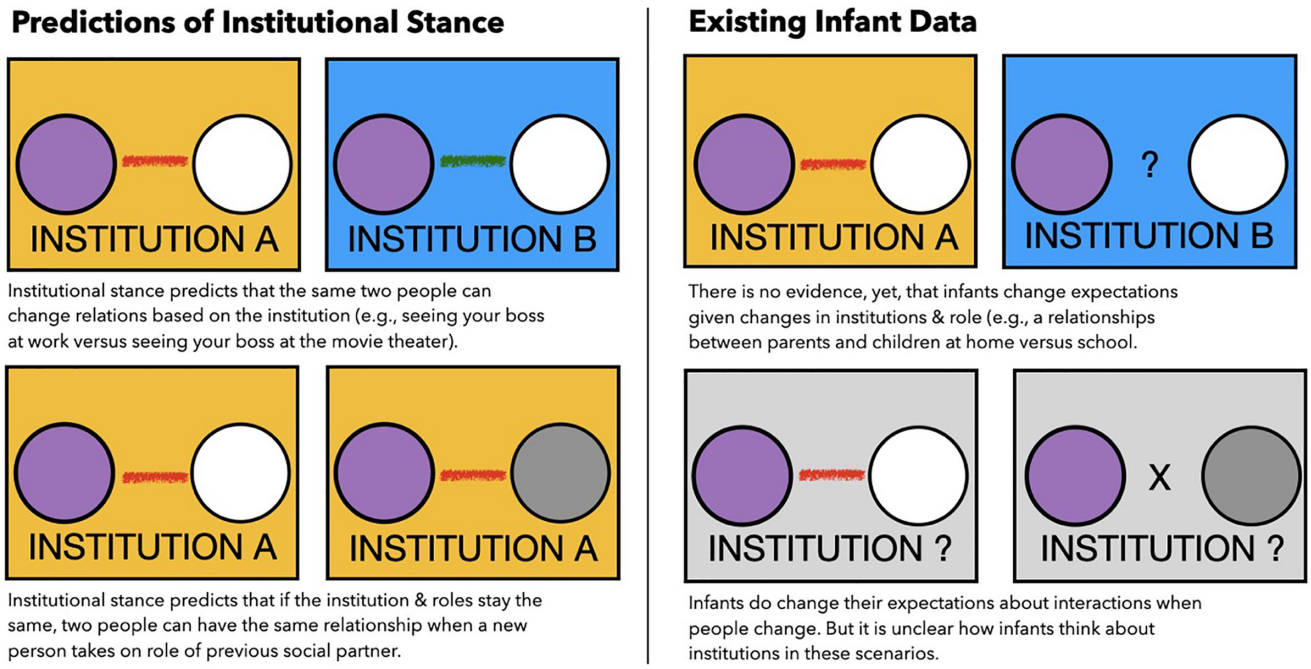


Figure 1 (Hok and Thomas). **Predictions of the institutional stance versus existing infant evidence**. The institutional stance predicts that expectations should track institutional roles rather than particular individuals. Existing infant work shows that expectations change when social partners change, but has not established role-filler independence. Circles represent individuals, background colors represent institutions, and links represent expected interactions.

(Hok, Thomas, & Saxe, 2025a).Classic sociology and historical analyses converge on a similar conclusion. Weber's account of bureaucratic institutions emphasizes the central role of impersonal, functionally defined positions in enabling scalable coordination in modern societies (Weber, 1921/1922/1978). Likewise, historians have argued that notions of impartiality "forcefully emerged" in the seventeenth and eighteenth centuries, as cultural tools for managing perceptions of parochial bias in increasingly larger and more fragmented social worlds (Murphy & Traninger, 2014). Furthermore, developmental work with children finds that notions of impartiality and procedural justice emerge fairly late in childhood: between 6 and 8 years (Schmidt et al., 2016; Shaw & Olson, 2014; Hok et al., 2025b). Together, this evidence suggests that an institutional stance does not form the foundation of social cognition. Rather, it may emerge as a social innovation to address issues that arise as humans form very large modes of organization. These inventions may be enabled by evolutionarily old capacities to represent relationships among particular individuals and by humans' prolific ability to learn from others and invent new ways of relating.

Acknowledgments. We thank Claire Augusta Bergey, Mack Briscoe, Aaron Chuey, Pablo Rodriguez Osztreicher, Rebecca Saxe, and Alex Shaw for comments and helpful feedback on earlier versions of this draft.

Financial support. Hannah Hok was supported by the Melon Foundation awarded to the Center for Research on Open and Equitable Science at MIT.

Competing interests. The authors have no competing interests.

References

DeScioli, P., & Kurzban, R. (2013). A solution to the mysteries of morality. *Psychological Bulletin*, *139*(2), 477–496. https://doi.org/10.1037/a0029065

Hok, H., Vasquez, K., Barakzai, A., & Shaw, A. (2024). The nonmeek inherit the earth: Children generalize dominance, but not submissiveness. *Developmental Psychology*, *60*(3), 469–484. https://doi.org/10.1037/dev0001696

Hok, H., Thomas, A. J., & Saxe, R., (2025a). Written in stone: Lay intuitions about the emergence of formal rules. *Proceedings of the Annual Meeting of the Cognitive Science Society* (47).

Hok, H., Gerdin, E., Zhao, X., & Shaw, A. (2025b). When should the majority rule?: Children's developing intuitions about majority rules voting. *Cognition*, *260*, 106128.

Kudrnova, V., Spelke, E. S., & Thomas, A. J. (2024). Infants infer social relationships between individuals who engage in imitative social interactions. *Open Mind*, *8*, 202–216.

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.

Murphy, K., & Traninger, A. (Eds.). (2014). *The emergence of impartiality*. Brill. https://doi.org/10.1163/9789004260849

Pepe, B., Woo, B. M., Thomas, A. J., & Powell, L. J. (2025). Helping and hindering guide infants' expectations about future behavior. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 47).

Schmidt, M. F., Svetlova, M., Johe, J., & Tomasello, M. (2016). Children's developing understanding of legitimate reasons for allocating resources unequally. *Cognitive Development*, *37*, 42–52.

Shaw, A., & Olson, K. (2014). Fairness as partiality aversion: The development of procedural justice. *Journal of Experimental Child Psychology*, *119*, 40–53. https://doi.org/10.1016/j.jecp.2013.10.007

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

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. https://doi.org/10.1073/pnas.2121390119

Weber, M. (1978). *Economy and society: An outline of interpretive sociology* (G. Roth, & C. Wittich, Eds.). University of California Press. (Original work published 1921/1922).

Roles or rules?

Samuel G. B. Johnson[1,2]*

[1]University of Waterloo, Canada and [2]University of Warwick, UK
samuel.johnson@uwaterloo.ca

*Corresponding author.

doi:10.1017/S0140525X26105731, e381

Abstract

Jara-Ettinger and Dunham's institutional stance conceives of social behavior in terms of roles; economists instead define institutions in terms of rules. I argue that rules, rather than roles, are logically (and likely cognitively) primary because (i) rules are underdetermined by roles, (ii) many rules are more abstract than roles, and (iii) rules only work if they rarely need to be enforced.

Understanding and improving institutions may be the most important task of social science. They determine much about every society's economic outcomes and the life prospects of every human being (Acemoglu, Johnson, & Robinson, 2005). Some economists recognize that neoclassical economics lacks the tools needed to fully understand institutions, pointing toward cognitive science to fill explanatory gaps (North, 2005). Developing a cognitive economics of institutions is thus a crucial, yet largely neglected, project (Johnson, in press), of which the evolutionary and developmental underpinnings would be a cornerstone.

It was with great interest, therefore, that I read about Jara-Ettinger and Dunham (J&D)'s institutional stance. J&D define "institutions" as a "network of interconnected roles" or "representation of a graph, where each node represents a role and each edge represents a relation." As examples, J&D mention universities, governments, committees, charities, homeowner associations, businesses (e.g., corporate offices, rental-car agencies, cafés), and clubs (e.g., sports teams, knitting groups, book clubs, sororities, study groups). I learned much from J&D's theorizing about role-based social inference.

I am less sure I have learned where "institutions" come from, in the sense that economists and political scientists use the term. They usually define "institutions" as the "rules of the game" within which economic and political activity occurs (Crawford & Ostrom, 1995; Hogdson, 2006; North, 1990). Prototypical examples include property law, insurance markets, democratic accountability, separation of powers, rule of law, accounting standards, schedules of taxation, and judicial rules of evidence. (J&D's sense of "institution" maps more closely to what institutional economists call "organizations" [North, 1990] – a subset of "institutions" in the broad sense. "Organization" is probably closer to the vernacular use of "institution.") The intricate combination of formal legislation and informal norms constituting these rules exerts pervasive but largely invisible effects on our lives.

Rules and roles are, of course, related. The rights and responsibilities entailed by roles follow from the rules associated with those roles (e.g., the role of property owner is defined by the rules governing what a person can do with their property), and formal rules gain their force in part from enforcement by individuals in roles imbued with authority (e.g., rules of evidence are enforced by judges). If rules and roles are two sides of the same coin, then perhaps the institutional stance gives us all we need to understand institutions in the economists' sense.

For several reasons, however, I believe that rules are more fundamental than roles, both logically and (I suspect) cognitively.

First, roles greatly underdetermine the rules. For example, an individual filling the role of "judge" has the power to issue rulings about intellectual property. But this tells us little about the content of those laws, and it is those details that matter for economic outcomes (Friedman, 2001). Copyrights provide weak but lengthy protections to authors against copying specific text (literally construed), whereas patents provide strong but short protections to inventors against imitating an idea (broadly construed). The stronger protections on patents are economically important because they raise the returns to innovation for the inventor; their shorter terms are equally important because they (eventually) reduce the cost of the innovation for the user. Although the judge has the power to enforce these rules, she has only modest discretion. Both the cognitive content and economic rationale of patent law focus on the rules, not the roles.

Second, some rules are abstracted so far from those who "enforce" them that it is unnatural to analyze them in terms of roles. A key "meta-institution" is the rule of law – the principle that abstract and consistent rules, rather than the particulars of who occupies positions of power, determine how institutions work (Fukuyama, 2011). For societies with rule of law, the enforcer of the rules is secondary to the rules themselves, enhancing the primacy of rules over roles. Just as importantly, the meta-institution of the rule of law is so abstract that it is hard to pinpoint any individual role as its "enforcer."

Third, because formal enforcement is costly, institutions can effectively guide society only when their enforcement is rarely required; this is why effective formal institutions are almost always built on top of informal norms that are largely self-enforcing. In this sense, the law has an existence independent of – and more fundamental than – the roles that enforce it (North, 1990).

None of this undermines the empirical claims made by J&D, which are both plausible and interesting. I also agree with the broader point that institutions rely on sociocognitive adaptations that are psychologically prior to institutions themselves. Institutions are cultural technologies built on top of evolved instincts, yet they come to transcend them by taking on a life of their own via cultural evolution. Where I differ from J&D is that I doubt that role-based construal is the primary sociocognitive adaptation needed for this job.

Consider an analogy. Spoken language is a set of instincts that biologically evolved at least 100,000 years ago, whereas written language is a technology that culturally evolved around 5,000 years ago. Reading and writing co-opt not only our language instinct but also other innate capacities such as object recognition and motor control. We likely wouldn't have written language if humans had not evolved bipedalism to free our hands – but we would not thereby identify bipedalism as the primary instinctual underpinning of writing. Interpreting behavior in terms of roles,

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

similarly, is at best one of many instinctual underpinnings of institutions.

What would a more comprehensive catalog of such instinctual precursors to institutional technologies include? As J&D would likely agree, norm psychology, theory-of-mind, and indeed, social-role construal surely all contribute. So too do many of our socio-cognitive adaptations for cooperation (e.g., reciprocity), intuitive theories (e.g., lay economics; Boyer & Petersen, 2018), the domain-general capacity for abstract reasoning, and the domain-specific capacity for reasoning about rational actors (as in Jara-Ettinger et al., 2016). To the extent that it makes sense to speak of an "institutional stance" that has *rules*, rather than roles, as its domain, it is more likely a patching together of distinct cognitive systems rather than a single adaptation.

Financial support. This work was supported by grants from the Social Sciences and Humanities Research Council of Canada (SSHRC Grants 435-2023-0232, 430-2023-00215, 2004-2024-00279), the Natural Sciences and Engineering Research Council of Canada (NSERC Grants DGECR-2023-00203 and RGPIN-2023-03700), and the Templeton World Charity Foundation (TWCF-2023-32568).

Competing interests. The author declares none.

References

Acemoglu, D., Johnson, S., & Robinson, J. A. (2005). Institutions as a fundamental cause of long-run growth. In P. Aghion & S. N. Durlauf (Eds.), *Handbook of economic growth* (pp. 385–472). Elsevier.

Boyer, P., & Petersen, M. B. (2018). Folk-economic beliefs: An evolutionary cognitive model. *Behavioral and Brain Sciences*, *41*, e158.

Crawford, S. E., & Ostrom, E. (1995). A grammar of institutions. *American Political Science Review*, *89*, 582–600.

Friedman, D. (2001). *Law's order: What economics has to do with law and why it matters*. Princeton University Press.

Fukuyama, D. (2011). *The origins of political order: From prehuman times to the French revolution*. Farrar, Straus, & Giroux.

Hodgson, G. M. (2006). What are institutions? *Journal of Economic Issues*, *40*, 1–25.

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

Johnson, S. G. B. (in press). A cognitive economics approach to institutions. In R. Viale (Ed.), *Elgar handbook of cognitive economics*. Edward Elgar.

North, D. C. (1990). *Institutions, institutional change and economic performance*. Cambridge University Press.

North, D. C. (2005). *Understanding the process of economic change*. Princeton University Press.

Taking institutions (and development) seriously

Alexander Noyes*

Boston College, USA
alexander.noyes@bc.edu

*Corresponding author.

doi:10.1017/S0140525X26105561, e382

Abstract

Jara-Ettinger and Dunham diminish institutions by conflating *institutional roles* with roles in a very broad sense. In doing so, they obscure the important conceptual changes that occur during middle childhood. I review the evidence, undiscussed by Jara-Ettinger and Dunham, that shows the absence of a Theory of Institutions in early childhood.

I previously argued that a "Theory of Institutions" (Noyes, Keil, & Dunham, 2020) develops during middle childhood and constitutes a conceptual change – i.e., a transformation and expansion of children's conceptual repertoire. Jara-Ettinger and Dunham take up the idea that there is a Theory of Institutions, but attribute that theory to infants and young children.

Jara-Ettinger and Dunham elide a critical distinction between roles in the very broad sense and *institutional roles*. An institution is (or defines) a self-reproducing structure that retains its identity across changes in role-occupiers. For example, the US national women's soccer team has existed for over 40 years, even though the members today are not the members of 1985. Contrast a soccer team with a friendship. We can refer to being a friend as a role, but a friendship is not an institution. The identity of a friendship is not independent of the people in it. When I became friends with Jonah, a new friendship was born. If Jonah stops being my friend, I cannot preserve *that* friendship by substituting him with someone else.

Infants and young children make sophisticated inferences about social interactions and relationships. They might infer, for example, that if Amy dominates Bill and Bill is friends with Carl, then Amy dominates Carl too (Mascaro et al., 2023). But no degree of sophistication transforms patterns of domination and friendship into an institution – in the sense of a self-reproducing structure that retains its identity across changes in role-occupiers.

In Noyes and Dunham (2020), I presented children with the following scenario: a target is friends with the members of group A (but not B), but has fulfilled the joining procedure for group B (but not A). Four-year-old children say the target is a member of group A, the group of their friends, whereas older children and adults say the target is a member of group B. Without the concept of an INSTITUTION, young children cannot represent a group as numerically distinct from its members. In turn, they can only render a group as an aggregate of agents. To be a group, for young children, is to cooperate, share, and play together. The aggregate can expand or shrink as agents start and stop cooperating, sharing, and playing together, but there is no self-reproducing structure that is numerically distinct from the aggregate (Noyes et al., 2023).

Let me return to the explanation, which is the focus of the target article. Jara-Ettinger and Dunham argue for two modes of explanation of human behavior: We can explain behavior via mental states or institutions. Everyone agrees that children can make sense of agent behavior by drawing on knowledge about things other than mental states. Some of that knowledge includes generalized expectations about people in role categories, such as "soccer player." Although these categories refer to institutional roles in the adult language, we cannot assume that these words are backed by institutional concepts in early childhood. Children might say, "she kicked the ball *because she is a soccer player*"; Jara-Ettinger and Dunham take this as evidence for an "institutional stance." But, if the explanation rests on generalized expectations about agent goals, rather than how the agent is situated in an institution, then children's reasoning may be accomplished by a Theory of Agents alone. To evidence a Theory of Institutions, we

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

need to examine cases where a Theory of Agents furnishes the wrong explanation of an institutional phenomenon, such that the right explanation is diagnostic of a Theory of Institutions.

One test case that serves this purpose is *institutional action*, which includes signing an executive order, officiating a wedding, and scoring a goal. An action is institutional when the agent performing the action must do so in their capacity as an institutional role-occupier in order to achieve the expected outcome, which is a change to some institutional fact. For example, pushing a ball across a threshold can obligate the other team to concede defeat and the governing body to award a trophy. Institutional actions are an ideal case for testing whether children have a Theory of Institutions: One cannot understand *why* an institutional action achieved its outcome based on an analysis of agent behavior alone. Without the concept of INSTITUTION, one cannot grasp that a soccer team is a separate entity from the agents; in turn, one cannot grasp that scoring a point depends on how an agent is situated within the structure of the soccer team. If an "institutional stance" is needed anywhere, it is needed here.

I examined children's reasoning about institutional actions in Noyes et al. (2020). I taught children that three people belonged to a role that kicks balls into trees to score points for their team. Then, I told the children that the team removed the original three people from the role and inserted three new people. I then asked children what would happen if one of the removed individuals kicked a ball into a tree. With a Theory of Agents alone, a child may see the behavior as weird or even wrong. But the child cannot see the behavior as ineffectual. The action *did* have the desired effect – it went across the threshold! The action is only ineffectual with respect to institutional facts. Therefore, to see the action as ineffectual requires a Theory of Institutions. I asked the children: Would kicking the ball into the tree score a point for one of the teams? Nine-year-old children said it would not. But four-year-old children said it would. They failed to see occupying the institutional role as explanatorily relevant. The same pattern of responding occurred for all the institutional actions I tested.

Jara-Ettinger and Dunham may interpret these data as the continued enrichment of children's Theory of Institutions. But, if children's ontology excludes institutional objects – i.e., self-reproducing structures that retain their identity across changes in role-occupiers – then there is no Theory of Institutions to be enriched. During middle childhood, a conceptual change unlocks a whole new ontological category and makes intelligible a whole new class of phenomena, including institutional groups and actions.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

References

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*, 1294–1306.

Noyes, A., & Dunham, Y. (2020). Groups as institutions: The use of constitutive rules to attribute group membership. *Cognition*, *196*, 104143.

Noyes, A., Keil, F. C., & Dunham, Y. (2020). Institutional actors: Children's emerging beliefs about the causal structure of social roles. *Developmental Psychology*, *56*, 70.

Noyes, A., Keil, F. C., Dunham, Y., & Ritchie, K. (2023). Same people, different group: Social structures are a central component of group concepts. *Cognition*, *240*, 105567.

The conceptual form of the institutional stance

Sandeep Prasada*

Hunter College, CUNY, USA
sprasada@hunter.cuny.edu

*Corresponding author.

doi:10.1017/S0140525X26105688, e383

Abstract

There cannot be a distinct grammar for the institutional stance because institutional kind concepts (i) are not self-contained, (ii) share formal characteristics with non-institutional kind concepts, and (iii) combine with non-institutional kind concepts to form coherent thoughts, implicating a common representational format. The stance is best thought of as one perspective provided by a generative grammar of kind concepts.

Jara-Ettinger and Dunham propose that the institutional stance is supported by a generative grammar on the basis of (i) the fact that we can represent and create an unbounded number of arbitrary institutional structures; (ii) the existence of a unique type of representation that is specialized for thinking and reasoning about institutions; and (iii) the fact that these resources can be used to think about social collectives from an institutional perspective whether or not they actually have institutional structure. Similar arguments can be made for other stances (e.g., essentialist and mentalist) potentially leading to the suggestion that our conceptual systems consist of multiple generative grammars. Such a view is problematic for multiple reasons.

I illustrate the problems for such a view by considering the institutional stance, though the problems apply equally to other stances. First, the institutional stance is not self-contained. Institutions cannot be fully characterized without making reference to representations of things that are not institutions (e.g., universities have buildings; baristas make coffee). Second, whereas Jara-Ettinger and Dunham do a great job of characterizing the ways in which institutional representations are distinctive, the institutional representations must also support modes of thought that are shared with concepts in other domains. For example, the concepts that represent institutions and their roles support instance-of-kind and kind perspectives, just as non-institutional concepts such as DOG and TABLE do. Thus, the concept BARISTA allows us to think and talk about particular persons as baristas as well as to think and talk about a particular kind of profession and thus baristas generally. Whereas instances of the kind barista may be characterized via predicates such as MALE or FEMALE, the kind itself cannot. On the other hand, baristas may be said to be widespread, even though instances of the kind baristas cannot be so characterized. Furthermore, whereas two instances of an institutional, or non-institutional, kind can be qualitatively identical,

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

differing only numerically, two institutional, or non-institutional, kinds cannot differ only numerically. Thus, two baristas, like two dogs, can be qualitatively identical but numerically distinct, but baristas and schmaristas, like dogs and schmogs, cannot be distinct kinds if they are merely numerically distinct, but qualitatively identical. In these, and other formal characteristics, kind and instance-of kind representations of institutional kinds do not differ from such representations for non-institutional kinds such as DOG and TABLE. If our conceptual systems contain multiple generative grammars, these shared conceptual capacities would have to be built into each grammar separately, a decidedly unparsimonious possibility. Finally, we can form an unlimited number of coherent thoughts involving both institutional and non-institutional kinds and their instances (e.g., THE SCHOOL DOES NOT ALLOW DOGS IN THE CLASSROOM) suggesting the need for a common format in which we can combine both institutional and non-institutional concepts.

A better solution than positing multiple generative grammars is to have a generative grammar that can capture both the unity and the diversity found in the ways we think and reason about things from different content domains. One attempt to formulate such a theory can be found in the generative theory of conceptual form (Prasada, 2026). According to that theory, all our concepts provide perspectives (stances) for thinking and talking about things and their properties. The perspectives are encoded in the formal structure of concepts – their *conceptual form* – which provides instructions for how to think about the contentful properties of the concept. Kind concepts are constituted of three types of formal components, a generative component, a descriptive component, and indexical elements. Kind concepts from formally distinct domains are generated via different combinations of the formal components of kind representations and how those components are realized. Prasada (2026) develops a fragment of the theory of conceptual form which generates kind representations for a wide variety of non-social kinds. It also suggests that the theory may be extended to generate social kind concepts by allowing that instances of social kinds are constituted of a multiplicity of persons, with some social kinds being constituted by persons in roles (e.g., FAMILY), while others do not require roles (e.g., COMMUNITY), and with some allowing independence between roles and the persons occupying those roles (e.g., COMMITTEE) and others not (e.g., FAMILY). Such extensions capture distinctive properties of formally distinct classes of social kind concepts. Those distinctive elements work in concert with formal components that characterize kinds across content domains. For example, both living kinds and artifact kinds are partially characterized via principled connections between the kind and certain functional characteristics (Prasada & Dillingham, 2006; Prasada, 2017, 2026). Given the normative dimension of principled connections and the relevance of functions for characterizing non-institutional kinds, it appears that the machinery needed for characterizing non-institutional kinds is well suited for characterizing institutional kinds as well as the roles of which those kinds are constituted. Incorporating Jara-Ettinger and Dunham's analysis of institutional kinds with the theory of conceptual form is likely to yield new insights into the mechanisms that guide our thoughts about institutions as well as other kinds of things. The resulting theory would have the characteristics that motivate Jara-Ettinger and Dunham's proposal that we have a generative grammar for institutional concepts without the drawbacks of thinking of it as being distinct from the formal dimension of commonsense kind concepts more generally. On that view, the institutional stance would be one of the many perspectives for thinking and talking made available by the generative grammar of conceptual form.

Acknowledgments. The author thanks Katherine Ritchie for comments on a previous version of this commentary.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

References

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition*, *99*, 73–112.

Prasada, S. (2017). The scope of formal explanation. *Psychonomic Bulletin & Research*, *24*, 1478–1487.

Prasada, S. (2026). *Conceptual form: The hidden structure of commonsense concepts*. MIT Press.

The fundamental attribution error as Institutional Stance neglect

Kyle G. Ratner*, Diego Padilla-Garcia, Qianqian L. Cui and Mariana Torres

University of California, Santa Barbara, Santa Barbara, USA
kyle.ratner@psych.ucsb.edu
diego05@ucsb.edu
qianqiancui@ucsb.edu
marianatorres@ucsb.edu

*Corresponding author.

doi:10.1017/S0140525X26105652, e384

Abstract

Social psychology tells us that perceivers tend to prioritize person-based explanations of behavior and fail to recognize the causal influence of social roles and other situational information. This cognitive bias seems to occur automatically but cultural factors matter. We explain the relevance of landmark social psychology contributions for understanding the Institutional Stance and its interplay with the Mentalistic Stance.

About 100 years ago, Lewin (1931) urged psychologists to move from Aristotelian to Galileian thinking. He argued that scholars should stop assuming that behavior emanates solely from a person's traits and mental states and recognize the power of situational forces. He firmly believed that social contexts guide desires and beliefs as much as the gravitational pull of the sun influences the Earth's orbit. If he were alive today, Lewin would love the Institutional Stance and how it offers a representational system for processing social relations to complement the dominant Mentalistic Stance. He might be dismayed, however, that Jara-Ettinger and Dunham largely overlooked the work of his intellectual progeny – the social psychologists – when formulating their ideas.

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

The argument for the Institutional Stance is mostly built from developmental psychology and ethology observations. The framework that emerges convincingly asserts that people are capable of deriving behavioral expectations from social roles. The points about role-filler independence and how perceivers might switch from an Institutional Stance to a Mentalistic Stance when individuals defy normative behaviors for their roles mesh well with social psychological perspectives on impression formation (e.g., Brewer, 1988; Fiske & Neuberg, 1990). There is a lot for social psychologists to laud. That said, we suspect that perceivers neglect to use the Institutional Stance more frequently than Jara-Ettinger and Dunham's reasoning would dictate. We derive this prediction from classic social psychology research on causal attribution.

A troubling reality emerged for the situationists of the 1960s and 1970s. Although studies were showing the power of the situation (e.g., Asch, 1951; Milgram, 1963), perceivers seemed to overestimate the impact of person factors and underestimate that of situational factors when explaining psychological and behavioral responses (Gilbert & Malone, 1995; Ross, 1977). These research participants were making the same mistake that irked Lewin about his scholarly contemporaries and that Jara-Ettinger and Dunham argue is still being made by researchers hyper focused on the Mentalistic Stance. Ross (1977) famously labeled this phenomenon the fundamental attribution error.

A lot of the evidence for the fundamental attribution error came from studies that put people in social roles designed to constrain their behavior and then documented how observers made dispositional inferences anyway. Here, we highlight a handful of prominent examples. Jones and Harris (1967) found that people assumed that an essay reflected a writer's true attitudes even when they were told that essay position was determined by a coin flip. Ross, Amabile, and Steinmeitz (1977) randomly assigned participants to either the role of questioner or contestant in a quiz game. Observers consistently rated the questioner as more intelligent than the contestant because they could answer more questions even though they were aware that the questioner chose the questions. Bierbrauer (1979) re-enacted Milgram's study about obedience to authority and found that observers failed to recognize the situational forces operating upon the teacher delivering the shocks unless they were given considerable time to reflect on what had transpired. Guided by studies like these, social role theory (Eagly, 1987) pioneered the idea that people's tendency for neglecting social role constraints has contributed to stereotypes, such as men are better leaders than women. In all of these examples, it appears that participants eagerly take a Mentalistic Stance and neglect to use the Institutional Stance to the degree warranted by their information processing task.

Bierbrauer's findings indicate that neglect of the situation might occur automatically, and deliberate reasoning can result in more appreciation of situational influences. Gilbert's three-stage model formalized this notion by arguing that observers first categorize the behavior of others, then characterize the actor in trait terms, and finally correct their initial inference with information about situational constraints if cognitive resources are available (Gilbert, Pelham, & Krull, 1988). According to Gilbert and colleagues, categorization and characterization occur automatically, consistent with Winter and Uleman's (1984) observations about spontaneous trait inferences, but the correction stage that considers situational explanations is a deliberate process. From this vantage, the Mentalistic Stance might be dominant because it can be accessed automatically when causal attributions need to be made, whereas it takes an extra step for the valuable social relation information tracked by the Institutional Stance to be incorporated into these judgments.

The last several decades of social psychology indicate that the interplay of the Mentalistic Stance and Institutional Stance might be culturally dependent. The moderating effect of culture is considered by Jara-Ettinger and Dunham, and they reference social psychological work by Gelfand, Nishii, and Raver (2017) to make this point. Pertinent to attribution, the fundamental attribution error seems less prevalent in people from East Asian cultures compared to those from the Western ones that filled the participant pools of the foundational attribution studies (e.g., Choi, Nisbett, & Norenzayan, 1999; Morris & Peng, 1994). Other studies suggest that East Asian individuals process situational information more automatically than Westerners (Knowles et al., 2001; Lee, Hallahan, & Herzog, 1996). The reasoning is that East Asian cultures prioritize understanding context and interdependence, whereas Western cultures emphasize the actor independent of the surrounding context (Nisbett et al., 2001). Thus, it is sensible to predict that the Institutional Stance is used more automatically by people reared in cultures that encourage interdependent construals.

We hope that our commentary highlights the value of the social psychology literature for unlocking theoretical insight about the Institutional Stance that might not otherwise be apparent. Jara-Ettinger and Dunham, for instance, state as self-evident that observers will infer a person's liking for olive oil differently when cues in the environment suggest that the olive oil holder is a customer versus a grocer stocking the shelf (Figure 3 in the target article). Social psychology suggests that this might not occur. It could depend on whether automatic or deliberate processes are engaged. It could depend on the cultural orientation of the observer. More broadly, we encourage scholars of various disciplines studying social behavior to mine the social psychology literature for inspiration. We totally understand apprehension due to replication concerns (Nelson, Simmons, & Simonsohn, 2018; Spellman, 2015). However, within the muck, there are a lot of insightful gems if you dig beneath the surface.

Financial support. The writing of this commentary was not supported by a specific grant from any funding entity.

Competing interests. None.

References

Asch, S. E. (1951). Effects of group pressure upon the modification and distortion of judgments. In H. Guetzkow (Ed.), *Groups, leadership, and men* (pp. 177–190). Carnegie Press.

Bierbrauer, G. (1979). Why did he do it? Attribution of obedience and the phenomenon of dispositional bias. *European Journal of Social Psychology*, *9*(1), 67–84. https://doi.org/10.1002/ejsp.2420090106

Brewer, M. B. (1988). A dual process model of impression formation. In T. K. Srull, & R. S. Wyer (Eds.), *Advances in social cognition* (Vol. *1*, pp. 1–36). Lawrence Erlbaum.

Choi, I., Nisbett, R. E., & Norenzayan, A. (1999). Causal attribution across cultures: Variation and universality. *Psychological Bulletin*, *125*(1), 47–63. https://doi.org/10.1037/0033-2909.125.1.47

Eagly, A. H. (1987). *Sex differences in social behavior: A social-role interpretation*. Lawrence Erlbaum.

Fiske, S. T., & Neuberg, S. L. (1990). A continuum of impression formation, from category-based to individuating processes: Influences of information and motivation on attention and interpretation. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (Vol. *23*, pp. 1–74). Academic Press.

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

Gelfand, M. J., Nishii, L. H., & Raver, J. L. (2017). On the nature and importance of cultural tightness–looseness. *Annual Review of Psychology*, *68*, 309–335. https://doi.org/10.1146/annurev-psych-010416-044033

Gilbert, D. T., & Malone, P. S. (1995). The correspondence bias. *Psychological Bulletin*, *117*(1), 21–38. https://doi.org/10.1037/0033-2909.117.1.21

Gilbert, D. T., Pelham, B. W., & Krull, D. S. (1988). On cognitive busyness: When person perceivers meet persons perceived. *Journal of Personality and Social Psychology*, *54*(5), 733–740. https://doi.org/10.1037/0022-3514.54.5.733

Jones, E. E., & Harris, V. A. (1967). The attribution of attitudes. *Journal of Experimental Social Psychology*, *3*(1), 1–24. https://doi.org/10.1016/0022-1031(67)90034-0

Knowles, E. D., Morris, M. W., Chiu, C. Y., & Hong, Y. Y. (2001). Culture and the process of person perception: Evidence for automaticity among East Asians in correcting for situational influences on behavior. *Personality and Social Psychology Bulletin*, *27*(10), 1344–1356. https://doi.org/10.1177/01461672012710013

Lee, F., Hallahan, M., & Herzog, T. (1996). Explaining real-life events: How culture and domain shape attributions. *Personality and Social Psychology Bulletin*, *22*(7), 732–741. https://doi.org/10.1177/0146167296227006

Lewin, K. (1931). The conflict between Aristotelian and Galileian modes of thought in contemporary psychology. *Journal of General Psychology*, *5*(2), 141–177. https://doi.org/10.1080/00221309.1931.9918387

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, *67*(4), 371–378. https://doi.org/10.1037/h0040525

Morris, M. W., & Peng, K. (1994). Culture and cause: American and Chinese attributions for social and physical events. *Journal of Personality and Social Psychology*, *67*(6), 949–971. https://doi.org/10.1037/0022-3514.67.6.949

Nelson, L. D., Simmons, J. P., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology*, *69*, 511–534. https://doi.org/10.1146/annurev-psych-122216-011836

Nisbett, R. E., Peng, K., Choi, I., & Norenzayan, A. (2001). Culture and systems of thought: Holistic versus analytic cognition. *Psychological Review*, *108*(2), 291–310. https://doi.org/10.1037/0033-295X.108.2.291

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. *10*, pp. 173–220). Academic Press. https://doi.org/10.1016/S0065-2601(08)60357-3

Ross, L., Amabile, T. M., & Steinmetz, J. L. (1977). Social roles, social control, and biases in social-perception processes. *Journal of Personality and Social Psychology*, *35*(7), 485–494. https://doi.org/10.1037/0022-3514.35.7.485

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, *10*(6), 886–899. https://doi.org/10.1177/1745691615609918

Winter, L., & Uleman, J. S. (1984). When are social judgments made? Evidence for the spontaneousness of trait inferences. *Journal of Personality and Social Psychology*, *47*(2), 237–252. https://doi.org/10.1037/0022-3514.47.2.237

How does institutional reasoning develop?

Marjorie Rhodes* and Duygu Yilmaz

Psychology, New York University, New York, USA,
marjorie.rhodes@nyu.edu
duygu.yilmaz@nyu.edu

*Corresponding author.

doi:10.1017/S0140525X26105640, e385

Abstract

The authors propose that humans are endowed with the "institutional stance" – a capacity to interpret social behavior through roles within institutional structures. We consider that this capacity might reflect a developmental achievement rather than a starting point. Early social reasoning may reflect domain-general tendencies to represent kinds with normative properties, with generative institutional reasoning emerging later with differentiated representations of individuals.

The target article proposes an elegant account of social cognition as reasoning about the roles people occupy, rather than only their mental states. The authors view this "institutional stance" as an evolutionarily endowed system that operates from infancy. Two features of the developmental trajectory of norm-based reasoning, however, raise the possibility of alternate origins and greater developmental change than the current proposal predicts. Rather than reflecting an early-emerging, domain-specific system, early norm-based reasoning may originate in domain-general representations of kinds with normatively connected features, with institutional reasoning emerging later as children develop richer representations of individuals and their mental states.

First, the weighting of normative and mental-state information shifts across development – young children show a surprising bias toward norms. For example, young children (e.g., ages 4–5) predict that agents will follow norms instead of their personal preferences (Berndt & Heller, 1986; Chalik, Rivera, & Rhodes, 2014; Kalish & Shiverick, 2004), and expect norms more than preferences to generalize across groups and be stable over time (Kalish, 2002, 2012; Kalish & Lawson, 2008; Kalish & Shiverick, 2004). They also misremember behavior driven by mental states as instead reflecting norms (Kalish, 2012). By around age 7, children instead prioritize mental states in each of these cases. These systematic developmental changes are not clearly specified within the proposed framework, which instead emphasizes early emergence and stability of both institutional and mental-state reasoning.

Second, children's early norm-based reasoning is not restricted to the social domain. Young children extend normative reasoning to animal categories, responding that animals *should* display their characteristic features (Foster-Hanson et al., 2021; Haward et al., 2018). They also view animals that most clearly exemplify these normative features as the best examples of their kinds (Foster-Hanson & Rhodes, 2019a). As in the social domain, these tendencies weaken with age, as children increasingly incorporate individuating information.

Domain-generality, along with developmental changes in the weighting of normative information, are both consistent with a general, early-emerging bias to represent kinds with characteristic, normatively connected features (Prasada & Dillingham, 2006; 2009), with a developmental shift toward more robust representations of individuals. Such a general shift is well-documented. Young children preferentially remember, choose to learn about, and generalize based on kinds more than individuals, with a shift across development toward more consideration of within-category variation and of individuals as distinct from the categories to which they belong (Brandone, 2017; Cimpian & Erickson, 2012; Cimpian & Park, 2014; Gelman, Sánchez Tapia, & Leslie, 2016; Gülgöz & Gelman, 2015; Leslie & Gelman, 2012; Rhodes, Brickman, &

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

Gelman, 2008; Rhodes, Gelman, & Brickman, 2008; Riggs, Kalish, & Alibali, 2014).

From this perspective, early norm-based reasoning may not yet reflect institutional reasoning in the sense described by the authors, but instead broader features of how children reason about kinds. A key distinction, as the authors note, is that in kind-based reasoning, normatively connected features are properties of the category members themselves, whereas in institutional reasoning, such features arise from occupying a role. A critical test, then, is whether children treat norms as intrinsic properties of people or as contingent on roles[1].

Although direct evidence on this question is limited, existing findings are suggestive. Young children often treat role-based obligations as identity-defining, viewing individuals who occupy certain roles as being category members "deep down" (Foster-Hanson & Rhodes, 2019b). In addition, children tend to conflate norms and desires, assuming that people want to fulfill their obligations (Costanzo, Grumet, & Brehm, 1974; Kalish, 1998; Kalish & Shiverick, 2004; Riggs & Fast, 2025). Only later do children understand that norm compliance can diverge from individual preference, and respond that only those who genuinely care about fulfilling a role are members of the category "deep down." Together, these findings suggest that young children may initially represent norms as intrinsic to category membership rather than as contingent on roles.

Thus, perhaps young children begin with domain-general kind representations that support normative expectations. Over time, they develop richer representations of individuals and come to understand norms as arising from roles rather than as an intrinsic part of category membership. One way to reconcile this developmental account with the target article is to interpret early norm-based reasoning as proto-institutional cognition. On this view, young children represent categories with normative expectations but lack the role-filler independence and structured coordination that characterize mature institutional reasoning. Ontogeny may therefore parallel the phylogenetic progression proposed, with fully flexible and generative institutional reasoning emerging later in development.

A central question is how this transition occurs. One possibility, which could present a challenge to the proposed framework, is that it depends in part on the development of mental-state reasoning. Understanding that people sometimes act counter to their own preferences, and that different people, with different mental states, can occupy the same role and behave similarly, may help children represent roles as abstract positions that constrain behavior independently of the people who occupy them. Without thinking about mental states, role and person may remain fused. Consistent with this possibility, the ability to predict behavior that conflicts with category norms is related to theory-of-mind development (Chalik et al., 2014). When children explicitly represent mental states, deviations from norms become informative: the misalignment between a person's preferences and the category demands makes the category visible as something operating outside of individual membership. In this way, representing mental states might enable a shift from a kind of representation where category membership reflects something deep and internal to a role representation where the role is an external position that people can occupy, resist, or leave. We view this type of interplay between mental state and norm-based reasoning in the development of institutional thought as consistent with current developmental evidence, but a departure from the authors' proposal, as we understand it.

Taken together, these considerations suggest that institutional reasoning may emerge through the interaction of domain-general kind representations and developing capacities for representing individuals and their mental states. This view retains the authors' emphasis on norm-based reasoning as central to social cognition, but suggests that full-fledged institutional reasoning may characterize a mature endpoint of social cognition rather than its starting point.

Acknowledgements. We thank Eric Mandelbaum for the discussion of these ideas and comments on our previous drafts.

Financial support. No external funding supported this work.

Competing interests. We have no competing interests.

Note

[1] We are not challenging the general claim that infants and young children are capable of reasoning about role-filler independence (Hafri, 2024; Quilty-Dunn et al., 2023); the issue is whether that capacity underlies their representations of social roles.

References

Berndt, T. J., & Heller, K. A. (1986). Gender stereotypes and social inferences: A developmental study. *Journal of Personality and Social Psychology*, *50*(5), 889–898. https://doi.org/10.1037/0022-3514.50.5.889

Brandone, A. C. (2017). Changes in beliefs about category homogeneity and variability across childhood. *Child Development*, *88*(3), 846–866. https://doi.org/10.1111/cdev.12616

Chalik, L., Rivera, C., & Rhodes, M. (2014). Children's use of categories and mental states to predict social behavior. *Developmental Psychology*, *50*(10), 2360–2367. https://doi.org/10.1037/a0037729

Cimpian, A., & Erickson, L. C. (2012). Remembering kinds: New evidence that categories are privileged in children's thinking. *Cognitive Psychology*, *64*(3), 161–185. https://doi.org/10.1016/j.cogpsych.2011.11.002

Cimpian, A., & Park, J. J. (2014). Tell me about Pangolins! Evidence that children are motivated to learn about kinds. *Journal of Experimental Psychology: General*, *143*(1), 46–55. https://doi.org/10.1037/a0031687

Costanzo, P. R., Grumet, J. F., & Brehm, S. S. (1974). The effects of choice and source of constraint on children's attributions of preference. *Journal of Experimental Social Psychology*, *10*(4), 352–364. https://doi.org/10.1016/0022-1031(74)90031-6

Foster-Hanson, E., & Rhodes, M. (2019a). Is the most representative skunk the average or the stinkiest? Developmental changes in representations of biological categories. *Cognitive Psychology*, *110*, 1–15. https://doi.org/10.1016/j.cogpsych.2018.12.004

Foster-Hanson, E., & Rhodes, M. (2019b). Normative social role concepts in early childhood. *Cognitive Science*, *43*(8), e12782. https://doi.org/10.1111/cogs.12782

Foster-Hanson, E., Roberts, S. O., Gelman, S. A., & Rhodes, M. (2021). Categories convey prescriptive information across domains and development. *Journal of Experimental Child Psychology*, *212*, 105231. https://doi.org/10.1016/j.jecp.2021.105231

Gelman, S. A., Sánchez Tapia, I., & Leslie, S. J. (2016). Memory for generic and quantified sentences in Spanish-speaking children and adults. *Journal of Child Language*, *43*(6), 1231–1244. https://doi.org/10.1017/S0305000915000483

Gülgöz, S., & Gelman, S. A. (2015). Children's recall of generic and specific labels regarding animals and people. *Cognitive Development*, *33*, 84–98. https://doi.org/10.1016/j.cogdev.2014.05.002

Hafri, A. (2024). Cognitive development: The origins of structured thought in the mind. *Current Biology*, *34*(18), R856–R859. https://doi.org/10.1016/j.cub.2024.07.096

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

Haward, P., Wagner, L., Carey, S., & Prasada, S. (2018). The development of principled connections and kind representations. *Cognition*, *176*, 255–268. https://doi.org/10.1016/j.cognition.2018.02.001

Kalish, C. (1998). Reasons and causes: Children's understanding of conformity to social rules and physical laws. *Child Development*, *69*(3), 706–720. https://doi.org/10.2307/1132199

Kalish, C. W. (2002). Children's predictions of consistency in people's actions. *Cognition*, *84*(3), 237–265. https://doi.org/10.1016/S0010-0277(02)00052-5

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. https://doi.org/10.1111/j.1467-8624.2008.01144.x

Kalish, C. W., & Shiverick, S. M. (2004). Children's reasoning about norms and traits as motives for behavior. *Cognitive Development*, *19*(3), 401–416. https://doi.org/10.1016/j.cogdev.2004.05.004

Kalish, C. W. (2012). Generalizing norms and preferences within social categories and individuals. *Developmental Psychology*, *48*(4), 1133–1143. https://doi.org/10.1037/a0026344

Leslie, S. J., & Gelman, S. A. (2012). Quantified statements are recalled as generics: Evidence from preschool children and adults. *Cognitive Psychology*, *64*(3), 186–214. https://doi.org/10.1016/j.cogpsych.2011.12.001

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition*, *99*(1), 73–112. https://doi.org/10.1016/j.cognition.2005.01.003

Prasada, S., & Dillingham, E. M. (2009). Representation of principled connections: A window onto the formal aspect of common sense conception. *Cognitive Science*, *33*(3), 401–448. https://doi.org/10.1111/j.1551-6709.2009.01018.x

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2023). The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, *46*, e261. https://doi.org/10.1017/S0140525X22002849

Rhodes, M., Gelman, S. A., & Brickman, D. (2008). Developmental changes in the consideration of sample diversity in inductive reasoning. *Journal of Cognition and Development*, *9*(1), 112–143. https://doi.org/10.1080/15248370701836626

Rhodes, M., Brickman, D., & Gelman, S. A. (2008). Sample diversity and premise typicality in inductive reasoning: Evidence for developmental change. *Cognition*, *108*(2), 543–556. https://doi.org/10.1016/j.cognition.2008.03.002

Riggs, A. E., Kalish, C. W., & Alibali, M. W. (2014). When you've seen one, have you seen them all? Children's memory for general and specific learning episodes. *Developmental Psychology*, *50*(6), 1653–1659. https://doi.org/10.1037/a0036130

Riggs, A. E., & Fast, A. A. (2025). Children's beliefs about the emotional consequences of norm adherence and violation. *Journal of Experimental Child Psychology*, *249*, 106071. https://doi.org/10.1016/j.jecp.2024.106071

Interpreting and reimagining social kinds using the institutional stance?

Katherine Ritchie*

Philosophy, University of California Irvine, USA
kcritchie@gmail.com

*Corresponding author.

doi:10.1017/S0140525X26105573, e386

Abstract

Jara-Ettinger and Dunham argue that the institutional stance is used to interpret the behavior of individuals and represent alternative structures. While they focus on roles like barista, they argue that the stance also applies to kinds like race and gender. I argue that there are several differences suggesting this is incorrect, or at least not yet supported by their arguments.

Jara-Ettinger and Dunham argue that the institutional stance can be used to interpret and analyze behavior and that it is generative, allowing us to represent and create arbitrary structures. While most of their examples focus on roles for individuals like barista, assistant, or the most dominant member of a baboon troop, they also connect the institutional stance to social kinds like race, gender, and caste kinds, arguing that the institutional stance "offers the possibility of envisioning and enacting new and more just futures" (39). Here I draw out differences between these cognitive projects applied to, on the one hand, individuals playing roles and those roles themselves (e.g., a barista, the barista role), and, on the other, to individual kind members and social kinds themselves (e.g., a woman, the kind *women*). These suggest the institutional stance is not utilized in both cases, or, at least, that more work is needed to show that it is.

Imagine a basketball team. It has a structure with positions like point guard and center. Team members play these roles. Players' behavior can be interpreted through the roles they play. Other basketball teams share the same structure, but on those teams, different people play each role. And each team might have different players at different times. We can also imagine roles themselves changing – for instance, the role of point guard coming to involve more shooting as part of its function.

Now think of social kinds like *women* and *men* or *White*, *Black*, and *Asian people*. Like all other kinds, these can vary in members across time. And like many other kinds, any member is usually thought to be a member for life. As Jara-Ettinger and Dunham note, many theorists have argued that these kinds should be theorized as structural. For example, gender might be analyzed in structural terms so that what it is to be a woman, for instance, is to be positioned as subordinate to men (Haslanger, 2000). Here, theorists are not considering some particular group with a hierarchical structure (like a particular basketball team with its structure). Rather, it is the kinds themselves that are analyzed as structural (Ritchie, 2020).

Does the institutional stance enable the interpretation of individual behavior through these social kinds? For instance, might people interpret the behavior of an individual woman or Black person not through their mental states, but through their being members of the kind of *women* or *Black people*, respectively? It seems that people do interpret behavior and aptitudes in terms of kind membership, but this is compatible with many ways of reasoning that do not involve the institutional stance. For instance, people might reason that these kinds have essences, and that this explains the behavior of kind members.

Further, there are several pressures that support the use of the institutional stance for roles like barista that do not apply well to gender, race, and caste. First, individuals can shift between roles across time. A creature might move up in a dominance hierarchy. A bee might shift from forager to receiver. Eric might be an assistant at one point, an acquaintance later. Second, there is

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

fungibility in role-fillers. Steve might play the role of point guard, then Chris subs in. One employee at a car rental agency might step in to continue a conversation with a customer. By representing roles and structures themselves, we have the tools to apply different roles to the same person and a single role to different people. Being able to represent roles and structures themselves also provides resources to analyze and imagine new roles and structures.

However, neither of these features applies well to kinds like *women*, *Brahmins*, or *Black people*. Return to our example of gender social kinds. As noted above, people tend to think of gender as stable across a person's lifetime. There are not many changes in roles played (in a game, after 5 pm, etc.) to suggest that these are roles that one might play only temporarily. And, it would be very strange (and sexist!) to imagine that Sarah could just "sub in" for Laurie as a woman. Roles like barista and kinds like *women* pattern differently on both features. So, it is not clear that interpreting via roles or imagining modification to roles (baristas, point guards) gives us resources to interpret via or imagine alternative ways a social kind might be.

Perhaps an appeal to structural thinking, which Jara-Ettinger and Dunham say "aligns" with their arguments about the institutional stance, will provide the needed support to bridge this gap (19). Research on structural thinking has found that children's and adults' explanations and judgments of mutability are sensitive to structural information (Vasilyeva & Lombrozo, 2020; Vasilyeva, Gopnik, & Lombrozo, 2018; Yang, Naas, & Dunham, 2021). The structural conditions tested rely on differences in access (e.g., to more education or to different toys) or constraints (e.g., fewer restrictions on hiring). Given different environmental structures, children and adults reason that a group (girls, boys, fictional Borunians) might hold higher-paying jobs or play with different toys. Does this work suggest that we represent social kinds themselves in structural ways (i.e., as positions or roles in structures)? Does it show that people take social kinds themselves to be changed in such scenarios? Not directly. Rather, what I take it to show is that judgments about the stability of a property, inductive potential, or adequacy of an explanation can be connected to external structural sources. But this is compatible with the kind itself, not figuring in the interpretation of behavior. It is compatible with non-structural or even essentialist views of kinds and kinds themselves remaining unchanged, even if in some environments, individuals in these kinds display different behaviors.

Future work on the institutional stance ought to further explore its applicability to social kinds themselves, especially if this is useful for analyzing, imagining, and manifesting a more just social world.

Acknowledgements. Thanks to Carolina Flores, Rebecca Mason, and Sandeep Prasada for comments and discussions on an earlier version of this commentary.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors

Competing interests. None.

References

Haslanger, S. (2000). Gender and race: (What) are they? (What) do we want them to be?. *Noûs*, *34*(1), 31–55.

Ritchie, K. (2020). Social structures and the ontology of social groups. *Philosophy and Phenomenological Research*, *100*(2), 402–424. https://doi.org/10.1111/phpr.12555

Vasilyeva, N., Gopnik, A., & Lombrozo, T. (2018). The development of structural thinking about social categories. *Developmental Psychology*, *10*, 1–14.

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

Yang, X., Naas, R., & Dunham, Y. (2021). Testing the limits of structural thinking about gender. *Developmental Science*. https://doi.org/10.1111/desc.13169

The barista "must be friendly": Dispositional entailments as role content

Hagop Sarkissian*

CUNY Graduate Center, USA
hagop.sarkissian@baruch.cuny.edu

*Corresponding author.

doi:10.1017/S0140525X26105615, e387

Abstract

Jara-Ettinger and Dunham formalize role content as tasks, powers, and limitations, but their barista who "must be friendly" points to another role category: dispositional entailments, or expectations about orientation and manner. These neither fit their taxonomy nor reduce to mental state attributions. Recognizing them, I argue, expands the institutional stance's evaluative reach beyond the assessment of discrete role requirements.

Jara-Ettinger and Dunham argue that the institutional stance constitutes a distinct cognitive system for interpreting social behavior through role-based structures. Their three-part formalization of role content – tasks, socially granted powers, and limitations – is a genuine advance, specifying behavioral boundaries around role performance. All three occupy the same conceptual space: they delineate the behavioral requirements and permissions that attach to a role. Tasks specify what must be done, whereas powers and limitations define what one can and cannot do. And while a power described with its boundary conditions is itself a limitation (and vice versa), they may be pragmatically useful as separate categories.

This is what makes one of their own central examples so conspicuous. Jara-Ettinger and Dunham classify as a *limitation* the requirement that the barista "must be friendly" (Section 4.1). The requirement is real, yet it is not a behavioral boundary described from any direction. It cannot be cashed out as a discrete action or sequence of actions. It specifies, instead, an *orientation*: a quality of engagement with which the role's tasks are to be carried out. One might respond, however, that "must be friendly" *just is* a constraint on how tasks are performed and therefore a proper limitation whose violation constitutes role failure: a barista can observe every behavioral boundary while being surly or be warm and accommodating while making procedural errors. These are indeed familiar patterns of role failure, evaluable from within the

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

institutional stance without mentalizing (that is, without positing beliefs or desires to explain the failure), yet the institutional stance's formalization offers no way to distinguish them. I call the norms of orientation *dispositional entailments*, extending the sense in which normative entailments of roles are already discussed in this literature – namely, as expectations. Recognizing them as a distinct type of role content enriches the institutional stance without collapsing it into the mentalistic stance.

This matters because many roles central to social life – parent, teacher, leader, judge, host – have dispositional entailments as core content. A parent who feeds, clothes, and shelters a child but does so in a manner marked by indifference or contempt is failing in the *role*, not failing on some separate dimension that only mentalizing could detect. The failure is, to borrow their language, *institutional*: the role of the parent specifies that care should be enacted with a certain orientation. The current formalization of role content does not capture this difference – but an expanded formalization could. Indeed, the ease of identifying roles without dispositional entailments (compared to the difficulty of finding roles without limitations) suggests it may be a more natural third category.

One might pause: aren't dispositional entailments just requirements to hold certain mental states, as Jara-Ettinger and Dunham discuss in Section 6.3.3? The examples they provide, though, involve propositional attitudes: a stock person must know where items are located, and a role-occupant may be expected to want to fulfill their role's goals. Dispositional entailments are different. "Be caring" does not specify that the role-occupant must believe P or desire Q. Any attempt to decompose it into a propositional attitude either misses the target ("must believe caring is important" is compatible with cold, perfunctory conduct) or covertly reintroduces the dispositional content ("must want to engage warmly" simply restates an orientation requirement in mentalistic vocabulary). Dispositional entailments specify a quality of relational engagement, not an inference about an individual's psychology.

A community that expects its parents to exhibit care is not thereby predicting what individual parents think or feel; it is specifying a condition of adequate role performance. And like other role content, dispositional entailments are socially transmitted and publicly assessable. They are learned as part of what the role demands, and failures to meet them are recognized and evaluated by others within the institutional framework itself.

The dispositional entailments of roles become especially visible in normative traditions – such as classical Confucianism – that have theorized them explicitly. Confucian ethics are organizes around roles specifying not only what the role-occupant must do but also the orientation with which they must do it (King 2025; Sarkissian 2023, 2025). A passage from the canonical *Mengzi* (4B28) illustrates the point. When a role-occupant is treated with unexpected hostility by an institutional counterpart, the prescribed response is not to infer the offender's mental states, much less to examine their own. Instead, they should ask whether their scrutable role conduct may have prompted it somehow (perhaps inadvertently). Only when this kind of self-examination confirms full compliance with dispositional norms (tenderheartedness, decorum, and commitment) is it appropriate to reach a judgment about the other. What's more, the judgment is institutional, not mentalistic: this person is no longer within the space of role-governed interaction – they are simply "lost." The entire sequence operates within the institutional stance: the role-occupant evaluates their own performance against institutional standards and from the perspective of the counterpart. The line of interrogation is "What did I do?" (role assessment), not "What was in my mind?" (introspection).

Compare this to Jara-Ettinger and Dunham's account of competition between the stances (Section 6.3.2). They propose that explanatory or predictive failure in one stance triggers recruitment of the other. Their examples include both discrete task failures (the coffee doesn't arrive) and failures of manner (a barista-friend acting unusually formal). Yet, their formalization accommodates only the first kind. Once dispositional entailments are recognized as distinct role content, the institutional stance gains evaluative resources that extend beyond compliance with discrete role requirements.

A frivolous judge, an inattentive host, a callous parent – these are recognizable instances of role failures assessed against dispositional standards without shifting to inferences about individual beliefs and desires. The failures can be read from the overt, scrupulous behavior of the role occupant (their tone, manner, and demeanor). The standard itself, however, cannot be reduced to any enumerable set of behavioral requirements. The institutional stance thus does more evaluative work than Jara-Ettinger and Dunham's formalization, with its emphasis on prediction, currently allows. But that only becomes visible once we recognize dispositional entailments as part of what roles specify – something their barista who "must be friendly" already suggests.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

References

King, R. A. H. (2025). *"The lord a lord, the minister a minister, the father a father, the son a son." Roles and virtues in Plato, Aristotle, the Mencius and the Xunzi*. Felix Meiner Verlag.

Sarkissian, H. (2023). Me-knowledge and effective agency. In T. Szabó Gendler, J. Hawthorne, & J. Chung (Eds.), *Oxford studies in epistemology* (Vol. *7*, pp. 261–277). Oxford University Press.

Sarkissian, H. (2025). From character trouble to role solutions: A way forward in contemporary moral psychology. *Analysis*. https://doi.org/10.1093/analys/anaf035

When roles replace minds: The gendered costs of the institutional stance

Konrad Szocik[a]* and Andrzej Rozmus[b]

[a]Department of Social Sciences, University of Information Technology and Management in Rzeszow, Poland and [b]Institute of Education Analysis, University of Information Technology and Management in Rzeszow, Poland
kszocik@wsiz.edu.pl
arozmus@wsiz.edu.pl

*Corresponding author.

doi:10.1017/S0140525X26105524, e388

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

Abstract

Jara-Ettinger and Dunham argue that people often adopt an institutional stance, predicting behavior through roles and rules rather than individual minds. In our paper, we show that in feminized domains such as care, service, and administration, role-based processing more readily turns persons into functions, reproducing gender inequality. We propose institutional design interventions that rehumanize through individuation cues and interactional norms.

Jara-Ettinger and Dunham offer a powerful reframing of social cognition: in addition to a mentalistic stance, humans deploy an institutional stance that predicts and explains interaction by representing role structures and their rule-like entailments. This proposal captures how people navigate complex coordination without continuously constructing fine-grained models of individual minds. It also clarifies why institutional markers, such as uniforms, formal disclosures, and status-coded contexts, so reliably shape interpretation. The authors explicitly recognize a moral downside: institutional cognition can reduce attention to others' mental and emotional lives, thereby lowering empathy and enabling dehumanization (Harris & Delgado, 2025). In our paper, we want to argue that this downside is not merely incidental but politically patterned in a way that matters directly for gender equality. In this context, it is significant that the first two examples illustrating the institutional position given by Jara-Ettinger and Dunham concern gender relations: the Kate–Eric nexus and the Laurie–Paul nexus.

We suggest that the institutional stance often operates as a gendered filter on mind perception and empathy. When interaction partners are processed primarily as interchangeable role occupants, we rely on role-based entailments (what the role is for and what it requires) rather than on the person's lived experience, intentions, and vulnerability. In many institutional contexts, this substitution is efficient and benign. But in gender-coded domains including care, service, education, and especially administrative support, roles have historically been structured to minimize friction and maximize replaceability. Feminized work is often organized around norms of smoothness, emotional availability, and support, which can make the worker's personhood less salient in interaction and the demands of others feel procedurally legitimate. The moral risk is not just instrumental treatment; it is the quiet conversion of a person into a function.

This is a distinctively cognitive pathway for reproducing inequality even when explicit sexist beliefs are absent. Feminist social ontology emphasizes that gender is not just an internal identity but a socially meaningful position that structures expectations, competencies, and entitlements across contexts. On conferralist views, being positioned as a woman is partly constituted by socially authorized patterns of treatment: others' expectations, their perceived entitlements, and the harms they tend to treat as role appropriate. If so, the institutional stance is not only a tool for coordination. It is one mechanism by which gendered social meaning becomes cognitively automated: the stance stabilizes what fits the role and screens off the moral salience of the person behind it (Ásta, 2018).

Objectification provides a clear marker of what attenuated mentalizing removes from social perception. It is often operationalized as reduced mind attribution and moral status – less agency, less experience, and correspondingly thinner moral regard. In organizational settings, broader workplace objectification is treated as a systematic temptation: people are reduced to instruments for extrinsic goals, undermining their needs for control, belonging, and self-esteem. When institutions are designed to treat certain roles as infinitely substitutable while extracting affective labor, the institutional stance can make that design feel like a neutral social fact (Kellie, Blake, & Brooks, 2019; Zhang, Wisse, & Lord, 2025).

Jara-Ettinger and Dunham already suggest that institutional processing can lower empathy and facilitate dehumanization. Here, we add that we should expect this effect to be asymmetrically distributed along gendered lines. The more a setting is institutionally marked by procedures, hierarchy, and uniforms, the less observers will mentalize workers and the more they will treat dignity violations as deviations from role performance rather than harms done to a person. This is particularly evident in totalitarian organizations, where meetings, or more broadly, social interaction, are aimed at quickly creating an organizational identity that maintains and reproduces organizational systems of rules. The agenda of a meeting in a given environment will therefore not be an attempt to negotiate between a person's current beliefs and their organizational role, but rather to exert a more or less nuanced, effective influence of the organization on that person's life and identity (Howe & Hinderaker, 2018).

Moreover, this shift should be stronger for roles culturally coded as feminine, including especially reception, teaching, care coordination, service, than for masculinized roles because feminized roles are more likely to be socially framed as support and care functions whose success is measured by invisibility and compliance. Studies of feminized ancillary occupations, including receptionist work, highlight a recurring pattern: complex interpersonal skill is undervalued and rendered socially invisible, making it easier for institutions to erase the person by treating the work as peripheral (Cortis & Naidoo, 2025).

If this diagnosis is right, the normative upshot is not to reject institutional cognition, but to treat rehumanization as a problem of institutional design. Coordination need not be purchased at the price of mentalizing; institutional scripts can be arranged so that role-based processing does not routinely eclipse attention to persons. In practice, this points to small but durable shifts toward individuation and recognition: using names rather than role labels, adding brief perspective cues, and shaping interactional norms so that expectations of compliance and availability give way to more basic forms of acknowledgment. Experimental evidence also suggests that even relatively simple affective shifts, such as inducing gratitude, can reduce objectification in interpersonal and workplace contexts (Shi et al., 2023). Framed this way, the central prediction is both descriptive and action guiding: where institutions introduce stable cues that make personhood salient within otherwise role-dominant settings, the institutional stance should dampen empathy less, and its gendered asymmetries should correspondingly recede.

Financial support. None.

Competing interests. None.

References

Ásta. (2018). *Categories we live by: The construction of sex, gender, race, and other social categories*. Oxford University Press.

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

Cortis, N., & Naidoo, Y. (2025). Gender segregation and women's "ancillary" occupations: The case of health care receptionists. *The Economic and Labour Relations Review*, *36*(2), 351–369. https://doi.org/10.1017/elr.2025.15

Harris, L. T., & Delgado, N. (2025). The functional role of interpersonal dehumanization and associated brain networks. *Nature Reviews Psychology*, *4*, 336–346. https://doi.org/10.1038/s44159-025-00439-9

Howe, W., & Hinderaker, A. (2018). The rule was the rule": New member socialization in rigidly structured totalistic organizations. *Atlantic Journal of Communication*, *26*(3), 180–195. https://doi.org/10.1080/15456870.2018.1472093

Kellie, D. J., Blake, K. R., & Brooks, R. C. (2019). What drives female objectification? An investigation of appearance-based interpersonal perceptions and the objectification of women. *PLOS ONE*, *14*(8), e0221388. https://doi.org/10.1371/journal.pone.0221388

Shi, J., Wang, X., Teng, F., & Chen, Z. (2023). A little appreciation goes a long way: Gratitude reduces objectification. *The Journal of Positive Psychology*, *18*(4), 627–635. https://doi.org/10.1080/17439760.2022.2053877

Zhang, B., Wisse, B., & Lord, R. G. (2025). Workplace objectification: A review, synthesis, and research agenda. *Human Resource Management Review*, *35*(4), 101104. https://doi.org/10.1016/j.hrmr.2025.101104

Infants' early understanding of interactions enables the acquisition of institutional roles

Tibor Tauzin*

Department of Linguistics, University of Vienna, Vienna, Austria
tibor.miklos.tauzin@univie.ac.at

*Corresponding author.

doi:10.1017/S0140525X26105500, e389

Abstract

Infants demonstrate a number of early-emerging sophisticated communicative and linguistic skills that allow them to infer situational constraints, interpret instrumental and communicative interactions, attribute goals to others, recognize group membership, and construct higher-level, abstract distinctions through language. Here, I argue that the understanding of institutional roles develops gradually from these uniquely human cognitive capacities, rather than as an independent skill.

Infant research has accumulated evidence showing that young humans can understand abstract *interactional* roles, guided by the individual goals of other agents. For example, infants can identify the role of a chaser as opposed to a chasee (Yin & Csibra, 2015) based on their intention to follow and catch another entity. Infants can also identify a giver as opposed to a taker (Tatone & Csibra, 2024) or a communicator as opposed to another agent who does not or cannot engage in communicative information transfer (Tauzin & Gergely, 2018, 2019). The foundations of recognizing interactional roles might also be present in nonhuman animals, allowing them to identify predetermined interactional roles, such as "prey" or "predator," in order to survive. In contrast to recognizing interactional roles, representing *institutional* roles should allow infants to understand that the behavior of an individual in a certain context is guided by the internalized *aims and norms of the group* to which this individual belongs. Crucially, however, there is no direct evidence that the comprehension of institutional roles – or, an institutional stance – is present early in development, nor that it has a specialized brain region or neural mechanisms dedicated to process such information. Although the absence of evidence is not evidence of absence, it is nevertheless necessary to examine whether an institutional stance constitutes a genuinely distinct ability, or whether it can be parsimoniously explained as an additional functional adaptation of an early-emerging cognitive skill with well-characterized neural underpinnings.

Accumulating empirical evidence suggests that human infants have precocious inferential skills to interpret the instrumental (Gergely & Csibra, 2003) and communicative behavior of others (e.g., Tomasello et al., 2005; Vouloumanos, Onishi, & Pogue, 2012; Vouloumanos, Martin, & Onishi, 2014) and perform adequate responses based on these inferences. When relying on these inferential skills, infants can take into account the situational context provided by the environment while assigning and maintaining representations of other agents' goals and states of informedness. In communicative interactions, for instance, they can draw adequate inferences to interpret a request about retrieving a target object based on its location, the informedness of this person about it, and the intention of the communicator (Király, Oláh, & Kovács, 2023; Tauzin, Call, & Gergely, 2024). Another line of research shows that, beyond their early inferential capacities, young infants can identify group membership on the basis of perceptual cues (Powell & Spelke, 2013, 2018). Infants expect members of a given group to behave similarly, interpreting such behavioral similarity as a signal of social affiliation (similar to other species, see e.g., Paukner et al., 2009).

Crucially, these abilities may form a sufficient basis to *gradually understand and learn about* institutional roles without possessing an institutional stance. By relying on their understanding of goal-directed behavior, infants may infer the instrumental goals of other agents in social interactions. For instance, they may understand that an exchange of objects (i.e., money and goods) occurs during shopping without yet understanding the institutional role of a shop clerk. In communicative interactions infants might also be able to draw inferences about the epistemic goals of others. For example, they may recognize that different educators possess the same relevant information about generic facts (Egyed, Király, & Gergely, 2013) even if they have no understanding of the institutional role of such individuals. Thus, by relying on the situational context as a cue and their ability to recognize group members, they may learn over the course of repeated interactions that, irrespective of the identity of the interacting individual, different people behave in a similar manner in given situations. Presumably, this is sufficient for infants to begin forming abstract representations of simple institutional roles (such as a shop clerk or a teacher) after repeated exposure to interactions with different individuals occupying these roles.

Importantly, this capacity is likely to be complemented by another early-emerging ability in human infants: their communicative and linguistic competence. Infants can infer that labels

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

refer to the same group of entities (Pomiechowska et al., 2024). Acquiring labels can, therefore, highlight commonalities across different individuals fulfilling the same role by referring to them with the same word, such as "clerk" or "educator," through a categorization effect. Consequently, even with limited direct experience of institutional roles, children can learn about them through language.

I thus propose that identifying institutional roles is supported by infants' sophisticated inferences about situational constraints, instrumental and communicative interactions, group behavior, and language at the beginning of development. These abilities have been shown to be present in young human infants and are supported by dedicated neural circuits that process domain-relevant information (such as the posterior superior temporal sulcus with respect to goals, see Begus et al. (2020); the right temporo-parietal junction with respect to informational asymmetries, see Hyde et al. (2010); Liu et al. (2025); the language circuit with respect to linguistic information). Possessing such skills may enable infants to recognize that certain interactional roles are stable across contexts and independent of the individual who occupies a given role in a specific interaction. This recognition requires active and presumably effortful inferential processes early in development in order to learn and generalize from experience about interactional roles, but it does not necessitate an institutional stance. Thus, understanding institutional roles may be a natural outcome of a prolonged learning process grounded in the aforementioned capacities, rather than an independent skill. This, in turn, may enable humans to develop fast access to, implicit recognition of, and sophisticated inferences about institutional roles as they grow older, through the automatization of these processes, similar to a music enthusiast who does not adopt a *musical stance*, but instead draws on a set of basic cognitive mechanisms to differentiate between genres. In addition to being more parsimonious, this account is consistent with current evidence, as the understanding of institutional roles appears to emerge slowly and gradually in children, perhaps without involving specialized brain circuits to understand institutions.

Financial support. This research received no specific grant from any funding agency, commercial, or not-for-profit sectors.

Competing interests. The author declares none.

References

Begus, K., Curioni, A., Knoblich, G., & Gergely, G. (2020). Infants understand collaboration: Neural evidence for 9-month-olds' attribution of shared goals to coordinated joint actions. *Social Neuroscience*, *15*(6), 655–667.

Egyed, K., Király, I., & Gergely, G. (2013). Communicating shared knowledge in infancy. *Psychological Science*, *24*(7), 1348–1353.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: The naive theory of rational action. *Trends in Cognitive Sciences*, *7*(7), 287–292.

Hyde, D. C., Boas, D. A., Blair, C., & Carey, S. (2010). Near-infrared spectroscopy shows right parietal specialization for number in pre-verbal infants. *Neuroimage*, *53*(2), 647–652.

Király, I., Oláh, K., & Kovács, Á. M. (2023). Can 18-month-olds revise attributed beliefs?. *Open Mind*, *7*, 435–444.

Liu, Y., Moss, E., Ting, F., & Hyde, D. C. (2025). Neural sensitivity to others' belief states in infancy predicts later theory of mind reasoning in childhood. *Cortex*, *184*, 96–105.

Paukner, A., Suomi, S. J., Visalberghi, E., & Ferrari, P. F. (2009). Capuchin monkeys display affiliation toward humans who imitate them. *Science*, *325*(5942), 880–883.

Pomiechowska, B., Takács, S., Volein, Á., & Parise, E. (2024). The nature of label-induced categories: Preverbal infants represent surface features and category symbols. *Proceedings B*, *291*(2035), 20241433.

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.

Powell, L. J., & Spelke, E. S. (2018). Human infants' understanding of social imitation: Inferences of affiliation from third party observations. *Cognition*, *170*, 31–48.

Tatone, D., & Csibra, G. (2024). The representation of giving actions: Event construction in the service of monitoring social relationships. *Current Directions in Psychological Science*, *33*(3), 159–165.

Tauzin, T., Call, J., & Gergely, G. (2024). Infants produce optimally informative points to satisfy the epistemic needs of their communicative partner. *Open Mind*, *8*, 1228–1246.

Tauzin, T., & Gergely, G. (2018). Communicative mind-reading in preverbal infants. *Scientific reports*, *8*(1), 9534.

Tauzin, T., & Gergely, G. (2019). Variability of signal sequences in turn-taking exchanges induces agency attribution in 10.5-mo-olds. *Proceedings of the National Academy of Sciences*, *116*(31), 15441–15446.

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

Vouloumanos, A., Onishi, K. H., & Pogue, A. (2012). Twelve-month-old infants recognize that speech can communicate unobservable intentions. *Proceedings of the National Academy of Sciences*, *109*(32), 12933–12937.

Vouloumanos, A., Martin, A., & Onishi, K. H. (2014). Do 6-month-olds understand that speech can communicate? *Developmental Science*, *17*(6), 872–879.

Yin, J., & Csibra, G. (2015). Concept-based word learning in human infants. *Psychological Science*, *26*(8), 1316–1324.

First principles for any proposal of a generative grammar for complex sociality

Lotte Thomsen[a,b]* and Jennifer Sheehy-Skeffington[c]

[a]Department of Psychology, University of Oslo, Oslo, Norway; [b]Department of Political Science, Aarhus University, Aarhus, Norway and [c]New York University Abu Dhabi, United Arab Emirates
lotte.thomsen@psykologi.uio.no
lotte.thomsen@gmail.com
jss610@nyu.edu

*Corresponding author.

doi:10.1017/S0140525X26105706, e390

Abstract

Social coordination stems not simply from individual institutional "stances," but from an intersubjective, deeply phylo- and ontogenetically-rooted shared generative grammar of core relational forms, likely evolved for navigating critical socio-ecological (historically and contextually varying) structural dilemmas in adaptive ways. This relational grammar coordinates and regulates concrete social interactions and the way we see ourselves and others, relating to each other through it, as well as social institutions, giving meaning to human social life and society.

"The question of how it is that humans can understand and coordinate infinitely varying and novel instances of meaningful social interactions is roughly parallel to the classical learnability question in theoretical linguistics – how it is that all children within a language community converge in learning the same grammar of their mother tongue, although the verbal stimuli they meet are impoverished, noisy, and highly variable. Ultimately, it amounts to asking how human social life is understandable, and hence possible.

The problem does not reduce to understanding that individuals are agents who act intentionally to fulfil their goals. Although by definition a social relationship takes place between intentional agents, knowing that several intentional agents are present does not specify what kind of social relationship they will have with one another. Pursuing the analogy with language acquisition, it is possible there is something analogous to universal grammar in the domain of social relations – a universal set of abstract kinds of social relations, and innate skeletal knowledge of these that guides learning, inference, and indeed experience in the social domain."

The basic point of the paragraphs above (quoted from, p. 18) is remarkably similar to Jara-Ettinger & Dunham's (hereafter, J-E&D) proposal that "a generative grammar that builds structured models of social collectives," not just reasoning about the mental states of other individuals, is necessary for understanding the social world (J-E&D, p. 1). We whole-heartedly agree: Just as you know which sentences are correctly formed in your mother-tongue even without any explicit knowledge of formal syntax, so this shared social grammar enables you to intuit that whenever you are hungry you can eat food from the fridge in your family home (regardless who bought it) but not at the office, or that a boss can order her PA to fetch her coffee at work, but not Saturday night at the movies (J-E&D, p. 1). It also licenses inferences of the moral consequences and reasons for violating these social rules. To our minds, the basic, critical questions of any social grammar proposal are 1) *what* are its conceptual primitives and evolved fundamental rules, and 2) *how* are these recognized, instantiated, and functioning in coordinating social interaction – making social life understandable, guiding socialization and adaptive social strategies, and sustaining and regulating different kinds of social institutions. To these questions, the answers of J-E&D are unfortunately vague.

Seeking to understand the same interface between social mind and society that concerns the target article, relational models (Fiske, 1991, 2004, 2025) and social dominance (Ho et al., 2025; Sidanius & Pratto, 1999) theories each build on classic social theory and decades of empirical social-scientific evidence. Fiske (1991, 2004, 2025) draws on vast records of ethnographic data to outline the evolved cues that universally signal and instantiate the fundamental relationships of communal unity, hierarchy, equality, and proportionality, complementing it with experimental evidence that people are cognitively encoded in terms of the basic kind of relationships one has with them (e.g., Fiske, Haslam, & Fiske, 1991). Sidanius and Pratto (1999) argued that other groups are predominantly seen and treated not in terms of their specific identities but as a function of their position in the societal hierarchy and the basic, abstract hierarchical motives of the perceiver, a claim supported by thousands of research papers (Ho et al., 2025).

The Thomsen-Carey proposal synthesized these insights with the theory of core cognition (Carey, 2009), arguing that structural relational models (e.g., communal/kin sharing and caring; hierarchy; equality) form part of an evolved and finite set of conceptual primitives for a coordinating, generative grammar that manifest even in the preverbal mind to make socialization possible, taking the perceptual input specified by Fiske's ethnographic review (e.g., sharing bodily fluids/synchrony/close touch for communal clique structures; vertical/size/numerical transitive differentiation for hierarchy) (see also Ho et al., 2025; Sheehy-Skeffington & Thomsen, 2020, 2023; Thomsen et al, 2011).

The very point of these prior proposals is that *the way in which* infants and all other people understand any specific relationships and social interaction is precisely as concrete instantiations of basic and general, abstract, latent, and compositional relational forms and the coordination rules they entail: *The structures of social life* (as Fiske titled his 1991 book) scale and compose to societal institutions, not only in the way they are represented, but in the strategic and adaptive ways they are endorsed and navigated, thereby forming political psychology (Ho et al., 2025; Sidanius & Pratto, 1999; Thomsen, 2010).

Surprisingly, J-E&D refer only briefly to this work with the claim that "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." (sect 5.3, para.3). Yet, the very point of a compositional, generative grammar is that it operates over some finite set of (likely innate, evolved) primitives. What J-E&D think these fundamental relations or roles are and how they function remains unclear. That they repeatedly turn to examples of communal care and hierarchy suggests that the notion of innate relational models (Fiske, 1991), as part of a compositional generative social grammar that scales up to society (Sheehy-Skeffington & Thomsen, 2020, 2023; Thomsen, 2010, 2020, 2025), is more central to institutional understanding than the authors will allow.

Arguing that reasoning about mental beliefs/intentions and social structures/constraints is separate stance neglects the centrality of intersubjective rules for understanding self, others, *and* social institutions, in a manner reminiscent of methodological individualism, ignoring seminal work in the social sciences (e.g., Berger & Luckman, 1966; Cooley, 1902; Garfinkel, Rawls, & Lemert, 2015; Goffman, 1959; Mead, 1934; Vygotsky, 1978). Institutions and, we'd argue, any concrete social relation*ships* are "*structurated*" (Giddens, 1984) through the repeated enactment of different relational roles they encompass. Arguing that a generative grammar builds structured models that constrain behavior begs the question of *what* its conceptually primitive relational/structural forms are that allow us to generate infinite rule- and role-based social coordination in the first place.

Financial support. This work was supported by the European Research Council (LT, grant number 101040978-COODINATE) and the Norwegian Research Council (LT, grant number 335600). LT was also supported by the Danish Research Foundation (grant number DNRF-144).

Competing interests. None.

References

Berger, P. L., & Luckmann, T. (1966). *The social construction of reality: A treatise in the sociology of knowledge*. Irvington Publishers.

Carey, S. (2009). *The origin of concepts*. Oxford Academic Press.

Cooley, C. H. (1902). *Human nature and the social order*. Charles Scribner's Sons.

Fiske, A. P. (2004). Four modes of constituting relationships: Consubstantial assimilation; space, magnitude, time, and force; concrete procedures; abstract symbolism. In Haslam, N. (Ed.), *Relational models theory: A contemporary overview* (pp. 61–146). Psychology Press.

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

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.
Fiske, A. P. (2025). *Representing relationships: Modes of cognition and connection*. Cambridge University Press.
Fiske, A. P., Haslam, N., & Fiske, S. T. (1991). Confusing one person with another: What errors reveal about the elementary forms of social relations. *Journal of Personality and Social Psychology*, *60*(5), 656.
Garfinkel, H., Rawls, A., & Lemert, C. C. (2015). *Seeing sociologically: The routine grounds of social action*. Routledge.
Giddens, A. (1984). *The constitution of society: Outline of the theory of structuration*. Polity Press.
Goffman, E. (1959). *The presentation of self in everyday life*. Doubleday Anchor Books.
Ho, A. K., Kteily, N. S., Sheehy-Skeffington, J., & Thomsen, L. (2025). Social dominance orientation: The motivational basis of intergroup inequality. In B. Gawronski (Ed.), *Advances in experimental social psychology* (Vol. *72*, pp. 125–208). Academic Press.
Mead, G. H. (1934). *Mind, self & society*. University of Chicago Press.
Sheehy-Skeffington, J., & Thomsen, L. (2020). Egalitarianism: Psychological and socio-ecological foundations. *Current Opinion in Psychology*, *32*, 146–152.
Sheehy-Skeffington, J., & Thomsen, L. (2023). Ideology as a moral-relational language. *Psychological Inquiry*, *34*(1), 35–42.
Sidanius, J., & Pratto, F. (1999). *Social dominance: An intergroup theory of social hierarchy and oppression*. Cambridge University Press.
Thomsen, L. (2010). *Seeing social relations*. Doctoral dissertation, Harvard University.
Thomsen, L. (2020). The developmental origins of social hierarchy: How infants and young children mentally represent and respond to power and status. *Current Opinion in Psychology*, *33*, 201–208.
Thomsen, L. (2025). Relational core cognition is for understanding the meaning of social life, not just enduring interpersonal relationships. *Behavioral and Brain Sciences*. https://www.cambridge.org/core/journals/behavioral-and-brain-sciences/article/abs/cognitive-representations-of-social-relationships-and-their-developmental-origins/E8FEB3ADE6576EF41162B5FA8BB79B18
Thomsen, L., Frankenhuis, W. E., Ingold-Smith, M., & Carey, S. (2011). Big and mighty: Preverbal infants mentally represent social dominance. *Science*, *331*(6016), 477–480.
Vygotsky, L. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

How uniquely flexible is the human institutional stance, and what is it for?

Felix Warneken[a]* and Alexandra G. Rosati[b]

[a]Department of Psychology, University of Michigan, Ann Arbor, USA and [b]Departments of Psychology and Anthropology, University of Michigan, Ann Arbor, USA
warneken@umich.edu
rosati@umich.edu

*Corresponding author.

doi:10.1017/S0140525X26105585, e391

Abstract

What aspects of the "institutional stance" are really unique to humans? We argue that some animals are *more* flexible than described, and that humans are *less* flexible than assumed. We finally suggest that attention to the possible functions of different institutions suggests many possible commonalities between humans and animals.

Jara-Ettinger and Dunham propose that humans have a uniquely flexible and generative capacity to represent social groups, such that the same individual can show different behaviors depending on their specific role. Animals, in this view, possess a proto-institutional stance, but are fundamentally limited in how they use it. We argue that the divide between humans and (some) animals may be smaller than described, and that a focus on the functions of institutions can help bridge this gap.

Are other great apes really inflexible?

Claims of a uniquely human institutional stance centers on our species' flexibility, as animals' capacities are proposed to be limited to a few categories like dominance and kinship, and are assumed to involve attributions of social roles as essentially immutable individual traits. We argue that at least our closest relatives, chimpanzees (*Pan troglodytes*), are more flexible than this characterization. Consider dominance hierarchies, a key example from the article. While animal dominance is often thought to reflect physical intimidation, chimpanzee hierarchies are substantially more fluid (de Waal, 1982; Muller, Wrangham, & Pilbeam, 2017). Dominant males can acquire rank in part through social support and may differ substantially in their use of aggressive versus affiliative strategies (Foster et al., 2009; Muller & Mitani, 2005). As such, any fixed concept of "dominant" will not fully capture this role in chimpanzee groups.

Similarly, while we agree that kinship roles are often fairly ingrained, chimpanzee kin interactions are also context-dependent (Pusey, 1983; Walker & Pusey, 2020). For example, mothers tend to be more sociable when they have infant sons than daughters (Murray et al., 2014). Moreover, while chimpanzee mothers overwhelmingly provide offspring care, younger siblings are sometimes "adopted" by older siblings following maternal death – even by males who would not normally exhibit such caretaking (Hobaiter et al., 2014; Reddy & Mitani, 2019). This suggests that chimpanzees can sometimes fill "empty" institutional roles, even ones as seemingly predetermined as "mother."

Finally, we consider a potentially overlooked proto-institution: social play. Play provides a meaningful point of continuity as it encompasses many of the sorts of leisure-based social institutions characterized as uniquely human (e.g., book clubs, sports leagues). Play is a crucial social behavior shown by juvenile and adult chimpanzees (Sabbi et al., 2024; Samuni et al., 2024), and the same individual may show either aggressive intent or signal playfulness through postures, facial expressions, and vocalizations. Moreover, play can change the "meanings" of objects: a stick may be a utensil when foraging, a weapon in an aggressive context, or a toy during play (Kahlenberg & Wrangham, 2010; Watts, 2007). In that sense, while animals' overt institutional *behaviors* are obviously limited

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

by their lack of language and material possessions, it is unclear whether the *cognitive representations* that allow for institutional flexibility may nonetheless be shared.

Are humans really that flexible?

We agree that humans can participate in a wide variety of institutions, ranging from coffee shops to universities to national governments. The target article proposes that humans can do so because we possess generative institutional representations. We argue that humans may be *less* cognitively flexible than assumed, and that this apparent variety may be mostly due to the variety of institutional *content*. This content may be generative, even while the underlying structure reflects strong commonalities: clubs sharing a passion for books, soccer, or Pokémon cards – or governing organizations of homeowners, cities, or nations – all share overlapping purposes (e.g., for socializing and leisure or for collective decision-making).

Work by Fiske (1991) described in the target article synthesizes cross-cultural insights to propose that there are only three or four *core* types of social relations defining an individual's roles within groups. For example, both a soccer league and a book club could follow a "communal sharing" or an "equality matching" principle. While there might be a myriad of ways in which these social relations manifest, the core structures are limited by strong boundary conditions. Similarly, while children seem to intuitively grasp these core social relations, they are also limited in the kinds of institutional structures they represent and create (Heck, Shutts, & Kinzler, 2022; Thomas, 2024), often emulating familiar social interactions (Harris, 2021) rather than generating a wide range of arbitrary institutional structures.

Finally, while the human institutional stance might open doors to understanding and creating new institutional structures, we think much of the specific institutional practices described in the article could be due to *historical* events and *cultural* changes in technologies and social exchange. For example, many small-scale human societies lack formalized, temporally-enduring political institutions and rather organize informal leadership as needed (Boehm, 1999; Garfield, Syme, & Hagen, 2020). Conversely, many institutions in postindustrial societies are shaped by the modern professionalization of these institutions and such historical changes may impact our psychological processes directly (Elias, 1939; Foucault, 1977). Lastly, as readers may experience when filing tax returns or reading Kafka novels, intuitive theories of institutions can break down when trying to navigate institutional bureaucracies. For this reason, the wide range of current institutional structures should not be taken as *direct* evidence for a generative mental ability to comprehend these institutions.

What are they for? The functions of institutions

We emphasize that understanding the *function* of different institutional structures is critical, both for formal scientific theories and for people navigating their social world. The target article uses "function" to describe the task or goal of a particular role within an institutional structure, such as a barista brewing coffee or blending sugar-free iced-vanilla-protein matcha. We think the holistic purpose of the institution must also be considered: coffee shops brew coffee, but more broadly serve foraging and socializing functions. Many of the diverse institutions mentioned in the article may meet a quite limited number of needs, such as acquiring food, maintaining social bonds, or collective decision-making. While human social interactions may manifest in diverse ways in terms of the *content* of institutional structures, some animals also engage in group-level behaviors with parallel functions. We predict that the number of core institutional *functions* and their associated structures will be more limited, and more widely shared, in phylogeny and human history.

Acknowledgements. We thank Josef K. for helpful insights.

Financial support. Nothing to declare.

Competing interests. The authors declare no competing interests.

References

Boehm, C. (1999). *Hierarchy in the forest: The evolution of egalitarian behavior*. Harvard University Press.

de Waal, F. B. M. (1982). *Chimpanzee politics*. Harper & Row.

Elias, N. (1939/1978). *The civilizing process: The history of manners* (E. Jephcott, Trans.). Basil Blackwell.

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.

Foster, M. W., Gilby, I. C., Murray, C. M., Johnson, A., Wroblewski, E. E. and Pusey, A. E. (2009), Alpha male chimpanzee grooming patterns: Implications for dominance "style". *American Journal of Primatology*, *71*, 136–144. https://doi.org/10.1002/ajp.20632

Foucault, M. (1977). *Discipline and punish: The birth of the prison* (A. Sheridan, Trans.). Pantheon Books.

Garfield, Z. H., Syme, K. L., & Hagen, E. H. (2020). Universal and variable leadership dimensions across human societies. *Evolution and Human Behavior*, *41*(5), 397–414. https://doi.org/10.1016/j.evolhumbehav.2020.07.012

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

Heck, I. A., Shutts, K., & Kinzler, K. D. (2022). Children's thinking about group-based social hierarchies. *Trends in Cognitive Sciences*, *26*(7), 593–606. https://doi.org/10.1016/j.tics.2022.04.004

Hobaiter, C., Schel, A. M., Langergraber, K., & Zuberbühler, K. (2014). Adoption by maternal siblings in wild chimpanzees. *PLOS ONE*, *9*(8), e103777. https://doi.org/10.1371/journal.pone.0103777

Kahlenberg, S. M., & Wrangham, R. W. (2010). Sex differences in chimpanzees' use of sticks as play objects resemble those of children. *Current Biology*, *20*(24), R1067–R1068. https://doi.org/10.1016/j.cub.2010.11.024

Muller, M. N., & Mitani, J. C. (2005). Conflict and cooperation in wild chimpanzees. *Advances in the Study of Behavior*, *35*, 275–331. https://doi.org/10.1016/S0065-3454(05)35007-8

Muller, M. N., Wrangham, R. W., & Pilbeam, D. P. (Eds.). (2017). *Chimpanzees and human evolution*. Harvard University Press.

Murray, C. M., Lonsdorf, E. V., Stanton, M. A., Wellens, K. R., Miller, J. A., Goodall, J., & Pusey, A. E. (2014). Early social exposure in wild chimpanzees: Mothers with sons are more gregarious than mothers with daughters. *Proceedings of the National Academy of Sciences*, *111*(51), 18189–18194. https://doi.org/10.1073/pnas.1409507111

Pusey, A. E. (1983). Mother-offspring relationships in chimpanzees after weaning. *Animal Behaviour*, *31*(2), 363–377. https://doi.org/10.1016/S0003-3472(83)80055-4

Reddy, R. B., & Mitani, J. C. (2019). Social relationships and caregiving behavior between recently orphaned chimpanzee siblings. *Primates*, *60*(5), 389–400. https://doi.org/10.1007/s10329-019-00732-1

Sabbi, K. H., Kurilla, S. E., Monroe, I. G., Zhang, Y., Menante, A., Cole, M. F., Otali, E., Kobusingye, M., Emery Thompson, M., Muller, M. N.,

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

Wrangham, R. W., Machanda, Z. P. (2024). Ecological variation in adult social play reveals a hidden cost of motherhood for wild chimpanzees. *Current Biology*, *34*, 1364–1369. https://doi.org/10.1016/j.cub.2024.02.025
Samuni, L., Mielke, A., Crockford, C., & Wittig, R. M. (2024). Social play fosters cooperation in wild adult chimpanzees. *Current Biology*, *34*(24), 5839–5845. https://doi.org/10.1016/j.cub.2024.10.058
Thomas A. J. (2024). Cognitive representations of social relationships and their developmental origins. *The Behavioral and Brain Sciences*, 1–53. Advance online publication. https://doi.org/10.1017/S0140525X24001328
Walker, K. K., & Pusey, A. E. (2020). Inbreeding risk and maternal support have opposite effects on female chimpanzee dispersal. *Current Biology*, *30*(2), R62–R63. https://doi.org/10.1016/j.cub.2019.11.081
Watts, D. P. (2007). Tool use by chimpanzees at Ngogo, Kibale National Park, Uganda. *International Journal of Primatology*, *29*(1), 83–94. https://doi.org/10.1007/s10764-007-9227-4

When do we take the institutional stance?

Evan Westra* and Daniel Kelly

Department of Philosophy, Purdue University, West Lafayette, USA
ewestra@purdue.edu
drkelly@purdue.edu

*Corresponding author.

doi:10.1017/S0140525X26105664, e392

Abstract

While role-based representations likely play some role in social cognition, we are skeptical that this amounts to a distinctive "stance." We encourage the authors to make more specific predictions about when we rely more upon institutional reasoning rather than mentalizing. We suggest that this might depend on the amount of perceived individual agency afforded to occupants of different social roles.

Walking down the street, someone you don't recognize makes eye contact, smiles, and waves while excitedly approaching you. You panic: *Should I know this person? What do they want? What should I do?* Then you see their clipboard and the non-profit logo on their t-shirt. You avert your gaze and scurry by.

Cases like this show how everyday social cognition efficiently integrates both institutional and mentalistic information as it navigates us through complex social interactions. The stranger's clipboard and t-shirt provide crucial cues about their social role – a canvasser for a non-profit organization. Grasping these cues enables you to infer the mental causes of their behavior. You conclude that they are a stranger who does not really recognize you, and that their friendliness is a charade. Rather, they are enacting a familiar social script because they want to persuade you to sign on to their cause. (But you aren't fooled.)

Jara-Ettinger and Dunham's account of the institutional stance leaves plenty of room for such collaborations between mentalistic and institutional reasoning. However, their account goes further, positing that in some social interactions, we may rely *entirely* upon the institutional stance. This is nicely illustrated in the examples of Kate and Eric, of Laurie, Paul, and Alex, and of our vending machine-like interactions with baristas, which depict contexts where we can reliably predict and interpret behavior using only structured role representations, and where relying upon mentalizing would be unnecessary, cognitively inefficient, and even misleading.

We are mildly skeptical about this aspect of their account. We suspect that social cognition is *isotropic* (Fodor, 1983): it holistically integrates all kinds of information when predicting and interpreting behavior – including mentalistic and institutional cues, but also cues about social group membership, artifacts and their functions, social norms, the broader physical environment, etc. Different kinds of information might be more relevant in certain contexts, but we remain unconvinced that it's useful to theorize about this in terms of distinct socio-cognitive "stances."

But we are curious to hear more. More specifically, we wonder what kinds of contextual factors might lead us to rely solely (or just more heavily) upon the institutional stance rather than mentalizing. The authors give some hints, highlighting how certain proximal cues – uniforms, verbal disclosures of institutional status, linguistic markers like the *t/v* distinction, and more subtle signs of automatic, norm-governed, or scripted behavior – might activate institutional reasoning. They also suggest that reliance upon role-based reasoning will be moderated by familiarity with the institution in question. We think there could be more to the story.

Here is one possibility: perhaps, all else equal, we should expect preferential reliance on the institutional stance in familiar institutional contexts. This idea still does not generate very useful predictions, though, as all else that could be relevant is rarely equal. There are many familiar institutional contexts in which grasping someone's role leaves open *crucial* questions about how to interact with them that can only be answered through sophisticated mentalistic reasoning (e.g., a student wondering how to respond to a stern teacher, or faculty strategizing about how to handle an erratic provost). Institutions are sites of highly complex social interactions, and we typically need all of the socio-cognitive tools at our disposal to navigate them. Maybe *some* institutional contexts require little or no mentalizing, but it's not at all obvious which these will be.

So, it is unlikely that institutional reasoning dominates mentalizing in every familiar institutional context. What Jara-Ettinger and Dunham need, then, is a more detailed theory (and set of predictions) about when and why *certain* institutional contexts invite more or less mentalizing than others.

Our suggestion: the predictions and interpretations produced by mentalizing and institutional reasoning differ greatly in terms of their *specificity*. Mentalistic reasoning is individualistic: it most naturally explains why a particular individual is engaged in a specific action and is well suited to predict what they are likely to do (*Alice likes to surf in the mornings, so she'll schedule her work meetings in the afternoons*). In contrast, institutional reasoning is inherently generic: it applies generally to whomever happens to be occupying a kind of role *(cashiers scan food items while customers place them in bags; servers ask for drink orders before food orders*). This suggests that mentalistic reasoning will be more useful and reliable in contexts where the target is known to have greater scope for individual agency, while role-based reasoning will be more useful and reliable in contexts where that scope is constrained or absent.

Notably, this prediction also holds for target individuals who are deeply ensconced in an institutional context. Some roles afford greater scope for individual agency than others. Often, these sit higher on the institutional hierarchy: bosses have more agency

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

than workers; customers have more agency than service staff; generals have agency more than infantrymen. The more institutional power an individual has, the more they are able to bring the idiosyncrasies of their own personalities to bear on their decisions and actions, and the more predicting their behavior, even within the context of their institutional role, will require reasoning about their particular attitudes and traits. Conversely, the less institutional power an individual has, the more their behavior will be constricted by their institutional role, and hence the more apt and reliable role-based representations will be for predicting their behavior.

Our prediction aligns with the spirit of many examples the authors use to illustrate the institutional stance in action, from Eric the hapless flunky, to the interchangeable customer service workers Alex and Paul, to the barista-as-vending machine. These are all low-status roles that afford little to no individual agency. Little personalized knowledge of their occupants' mental states is needed to predict what they will (or, perhaps, *must*) do.

Our suggestion represents only one route the authors might pursue; surely there are others. But we think it exemplifies the sort of amendment that will generate more specific predictions about when we rely more heavily upon the institutional stance: a more detailed theory of what, respectively, role-based predictions and mentalistic predictions are *good for*.

Financial support. This research received no specific grant from any funding agency, commercial or not-for-profit sectors.

Competing interests. None.

Reference

Fodor, J. (1983). *The modularity of mind*. MIT Press.

Authors' Response

What are institutions made of?

Julian Jara-Ettinger[a]* and Yarrow Dunham[b]

[a]Department of Psychology, Yale University, USA and [b]Wu Tsai Institute, Yale University, USA
julian.jara-ettinger@yale.edu
yarrow.dunham@yale.edu

*Corresponding author.

doi: 10.1017/S0140525X2610692X, e393

Abstract

How do people understand a social world built from hundreds of interconnected social structures? The commentaries on our proposed answer, the institutional stance, clustered around three questions: (1) What is the conceptual structure within the stance and where does it come from? (2) How does the stance integrate with mentalizing, and to what extent are they separate systems? (3) What changes when we think of people not as observers of social life but as participants in it? The commentaries also highlight new aspects of social cognition we had not considered but which may fall within the stance's scope. Here we address these questions and challenges, clarify key aspects of our account, and offer friendly amendments and future empirical directions aimed at addressing important challenges raised in the comments.

Across the animal kingdom, two things are clear. First, many species have complex social lives that are remarkably stable across ecologies. A chimpanzee troop in Uganda organizes much like a troop in Cote d'Ivoire. Second, even the most socially complex nonhuman species organize themselves in only a handful of ways. Humans, by contrast, have hundreds, perhaps thousands, of social structures. They vary radically across cultures, and they can endure for centuries or be invented and dissolved overnight. How do people track and infer these structures? What explains this productivity? How does it emerge and develop? We are glad to see broad agreement that these questions call out for further explanation.

The institutional stance offers an account of the cognitive machinery that allows us to effortlessly understand and navigate large-scale societies built from hundreds of interacting social structures. Our fundamental claim is that people can make sense of others' behavior by conceptualizing it as the product of interconnected networks of norm-governed roles. We can take this stance toward people who explicitly take on roles, but also toward people who may not think of themselves as occupying roles at all. And we can approach these tasks as impersonal observers or as embedded, personally accountable "members." Once we represent behavior this way, we can predict it, explain it, enter into it, and even think about how to contest or change it.

The commentaries clarify where the institutional stance has explanatory power, highlight aspects of our proposal that were underdeveloped, and most importantly point out things we got wrong. A running theme is that our article focused too heavily on examples of formally codified roles such as baristas, cops, and cashiers, but our proposal was meant to cover much more. Caregivers and kin, a group of friends, and even the strangers on a bus can all be understood as enacting interconnected roles. Using the word *institution* for both formal structures and informal ones may have misled in some cases, and much of our response here expands and refines our proposal with particular attention to how people apply an institutional stance beyond formally codified roles. Our response revises and develops our proposal, organized around seven questions.

R1. What the institutional stance represents

To take an institutional stance is to (1) interpret an agent's behavior as driven by a *role* they occupy and (2) treat the role as embedded in a network of other roles. These roles allow us to predict and explain behavior based on the behavioral expectations and norms embedded in them. In its full form, the stance

represents roles via role-filler independence; expectations about a role are stored separately from the individual who fills it.

Several commentators argued that our emphasis on role-filler independence implies that people are more substitutable than they actually are. After all your parent is *your* parent; in close relationships, the person and the role are bound together. Hok and Thomas and Tauzin take this to suggest that children must learn to separate role from role-occupier through experience, and that developmentally early representations focus on specific relationships rather than abstract roles. Noyes argues that relationships like friendship do not let us swap one person for another, as a role-based account implies. But role-filler independence does not imply that anyone can step into any role at any time. It only means that we partition our knowledge of the role from our knowledge of the person who occupies it, and that this separation lets us generalize each set of expectations appropriately. This is precisely why infants can quickly recognize a relationship as *caregiving* and generalize their expectations to new instances without having to learn what behaviors are characteristic of each specific relationship from scratch (e.g., Johnson, Dweck, & Chen, 2007; see Thomas, 2026 and our target article for review). More formally, this is a type/token distinction. Roles are types, and tokens are particular instances of those types. A child can simultaneously understand that their dad is *their* dad (a token), and that many other children also have their *own* dads (other tokens of the same type). There is no confusion about whose dad is whose, but because they are all dads, they all act in similar ways toward their kin. Of course, role-filler independence also lets us create highly transient roles that can be quickly exchanged across agents, such as who drives and who DJs during a long roadtrip. But this is a possibility the representations permit, and not a feature they force on every role. Even in the kinds of cases these commentators focus on, such as close friendships, a role-like understanding can do a lot of work. If I ask a friend to join me on a trip and they cannot, I might be disappointed and may think that no one can wholly replace them. But a role-like understanding of my friendship network will still be useful if you want to predict who I am likely to invite next. Indeed, even institutional settings that seem to allow clean substitution usually show these same dynamics. Consider a game of basketball. When Lebron James subs out and another forward subs in, the player's position lets us predict their behavior, but that in no way implies that we think nothing has changed! A familiar individual with person-specific attributes can still figure in our thinking. We do not see any of this as a challenge to role-filler independence but rather the very point of it: some things go with the role, others with the occupier, and the critical cognitive work is keeping the two distinct.

These considerations highlight an important point raised by Adams; the normative expectations embedded in roles cannot be so loose that they predict nothing, nor so tight that too many actions inevitably violate them. And this balance cannot be static; it must adapt dynamically, on the fly, to maintain the regulatory limits of the individuals acting within the roles. In this we agree with Adams. An interesting idea following from these considerations is that we can predict how much someone will need to mentalize by looking at the relative tightness of role expectations; as roles become looser or more negotiable, we will be more likely to need to supplement our predictions via mental state inference. These dynamics will likely differ on a role-to-role basis but also across different institutional domains.

R2. The institutional stance and mentalizing

Our article did not give enough attention to a critical point. To reason about a token (the roles as enacted by these particular people in this particular context), it is helpful to integrate role expectations with what we know about the specific people enacting them. We suspect people do this often with people they know well. For example, we may be surprised when a typically opinionated person is successfully carrying out a role that requires withholding judgment.

Because humans are intelligent and flexible agents, roles can allow different degrees of latitude. Some specify behavior tightly (a cashier's script), and others leave wider discretion (a CEO's leadership style). We agree with Westra and Kelly and Halevy and Halevy that the more latitude a role permits, the less the institutional stance alone can predict behavior, and the more observers will be obliged to take a mentalistic stance. Westra and Kelly point out that higher-status roles usually carry more discretion than lower ones, and this dovetails with work on social status pointing out that lower-status agents have to devote more cognitive resources to modeling higher-status agents than the reverse (e.g., those higher up in power structures show reduced perspective taking relative to those lower in the structure: Blader, Shirako, & Chen, 2016; Galinsky et al., 2006).

Halevy and Halevy here propose another way to integrate the two that we agree with. Rather than modeling how a specific mind implements a role, we can represent dispositional descriptors that capture how a mind carries out role-linked actions. For example, we can represent a barista as *eager*, *moody*, or *inattentive*. Each of these captures a particular way a mind can implement the barista role.

We also agree with Sarkissian who goes further to suggest that dispositional entailments can be built into the role itself. For example, we can embed *friendliness* into the barista role, or expect anyone in the role of a parent to be patient. These are, at heart, instructions about how a mind should implement a role. Roles, in other words, are not only enacted by minds; sometimes a role includes instructions about what kind of mind one should try to have. Here an interesting question is the extent to which this requires mental state reasoning to understand. Some cases, like embedding politeness norms in a role ("friendly barista"), can also likely be understood by thinking of the disposition through normative script-like expectations. That seems less plausible for other examples ("moody barista").

Linking roles to dispositions raises concerns that several other commentators pushed us on. If roles can specify how minds should work, in what sense is the institutional stance really distinct from mentalizing? Surely when people represent roles, they do not represent mindless automata. Most sharply, Westra and Kelly are right that people probably bring in all the tools at their disposal to understand others, such that "pure" deployments of a single stance may be rarer than our initial framing implied. These points raise important empirical questions, but our view does not hinge on any specific outcome. Our main commitment is not that people regularly separate these stances in practice. Rather, it is that the two stances are computationally separable. Each has its own minimal sufficient inputs, and thus the computational resources to generate predictions about behavior without appeal to the other system. If we tell you that Sally wants cookies and believes there are some in

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

the cabinet, you can predict what she will do without any information about roles. And if we tell you that Sally is playing tag and she is "it," you can predict what she will do without asking what she believes or desires. These minimal bundles mark the computational boundaries of each system. This can also be seen in how the bundles cannot be cut down further. Knowing that Sally wants cookies is not enough on its own (you need her beliefs), and that Sally is playing tag is not enough either (you need to know which role she is occupying).

Thus, although we agree that the two stances will frequently be coactivated, we do suspect that the balance between them can vary dramatically across contexts. Ratner et al. and Cikara raise interestingly different cases in which people favor one stance at the expense of the other. Ratner et al. point out that the fundamental attribution error might be an underuse of the institutional stance, over representing the role of the mind even when a role is at play. Cikara suggests that when the institutional stance is working well, it dampens our motivation to mentalize, and we risk forgetting there is a mind there at all, even in our closest relationships. Szocik and Rozmus raise similar concerns about larger institutional structures and propose ways to re-humanize the people occupying roles within them. Putting these together with the cases that Westra and Kelly and Halevy and Halevy describe, we see at least three different interactions across the two stances: cases where we see the mind in the role, cases where we mistake a role for a mind, and cases where we forget the agent behind a role has a mind. A critical open question is when and why we shift across stances, but we doubt there will be a single universal answer. As we noted in our target article, the relative use of an institutional stance has varied across historical periods, and Ratner et al. point out that it likely varies across cultures as well.

One important consideration here is that throughout, we conceptualize mentalizing in terms of the theory–theory approach (Wellman, 2018; see also our main article for our review on this approach). Cruz suggests that this creates a problem that could be avoided. If mentalizing is done through a theory of mind, then any behavior can be explained in mental-state terms if we sufficiently elaborate the mental states at play, e.g., by attributing complex sets of desires to act in particular ways in particular contexts. If so, why ever take an institutional stance? This leads Cruz to suggest that simulation-based mentalizing would be a better companion to the institutional stance. On the one hand, we are happy to acknowledge that simulation is a good way to model some aspects of other minds. But the empirical record shows that people often do hold a more theory-like understanding of other minds (e.g., Saxe, 2005). As to the worry that a theory–theory approach allows someone to recast nearly any institutional inference as a complex mental state inference, we just think that this is generally true: One *can* do so. Our claim is more narrowly empirical: That is not what people *in fact* do in a wide range of everyday cases. For example, when watching a uniformed person stock shelves in a supermarket, we generally interpret their behavior in simple, role-governed terms. But if pressed to explain why they take this seemingly menial job day after day, we can and indeed will generate a complex, theory-driven mentalistic explanation. We see the institutional stance as offering a distinct, low-cost interpretive pathway that we often resort to, but we are not troubled that the behavior can, in principle, be explained in other ways.

R3. Conceptual primitives and generativity

Moving from what lies within a role to what lies between roles, a key part of our explanation for the sheer number and range of human institutions is *generativity*; we combine and compose roles productively, so that we can represent social structures of arbitrary size and recombine existing representations to build new ones. For example, we can represent caregiver/kin relations with variable numbers of caregivers and kin, and with varying criteria for who can enter those roles (e.g., in alloparenting, same-sex relationships, etc); we can identify and reason about the local structure of roles we encounter in workplaces such as universities (e.g., faculty, graduate student, chair, registrar) and gradually extend the scope of what we know as our experience grows (dean, provost, ombudsperson). We proposed that a grammar might support this productivity because that is what grammars do, and we suggested that the elements of this grammar can be understood as a generative graph.

Prasada raises a challenge here. Institutional concepts do not exist in isolation. Professors are tied to universities in buildings and cities; baristas are tied to coffee and working hours. Our thoughts freely combine institutional and noninstitutional content ("the school does not allow for dogs in the classroom"). If institutional concepts have their own grammar, how do they combine with the rest of thought? Or does positing a separate grammar for institutional concepts duplicate conceptual machinery that a unified grammar of concepts already supplies? On the one hand, we are not sure the problem disappears under a unified grammar. Institutional kinds have unique structural properties – most notably a graph-like structure that determines who and how agents interact. Ordinary kinds such as *dog* and *table* can also be thought of as standing in relation to one another, but the kind of relation is of a fundamentally different kind. A unified grammar would have to encode these properties as special cases available to a subset of kinds, and from our perspective that might be sufficient to label it an institutional grammar. If Prasada's generative theory of conceptual form (Prasada, 2026) can handle the structural properties institutional representations need, all the better: the question is precisely how the system draws boundaries within a single representational architecture.

On the other hand, we certainly agree that institutional representations must combine with noninstitutional representations, but we do not think we are committed to a grammar so walled off that it cannot. The problem of conceptual combination across grammars has a long-standing solution in formal language theory: A context-free grammar can embed another within it. For instance, the grammar of arithmetic expressions has its own rules, while also living inside the grammar of nearly any programming language, existing as a self-contained system the host reaches into. The same might be true in human cognition. For instance, both intuitive physics and intuitive psychology have their own internal compositional structures (Xu et al., 2021; Velez-Ginorio et al., 2017), and they can still integrate seamlessly (Liu et al., 2026), which is why we can think about the physical and the institutional layers of a building at the same time (a recent proposal for exactly how to formalize these combinations is provided in the work of Chandra et al., 2026).

Thomsen and Sheehy-Skeffington, on the other hand, agree that a generative grammar of social structures exists and, as they

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

note, suggested this more than a decade before we did (Thomsen & Carey, 2013), an intellectual history we were remiss in not giving more attention to. Their proposal, building on Fiske's relational models, is that the primitives are four basic relational forms: communal sharing, authority/hierarchy, equality matching, and proportionality (Thomas, 2026 has advanced a closely related proposal). We are not convinced that the Fiske primitives are at the right level of analysis to ground the institutional stance. This is because the level at which they describe relationships is so abstract that it strips out much of what we actually think about when we reason through a specific social collective (see also Powell & Tompkins, 2026). A university and a for-profit company are both authority-ranked structures. But the inferences we draw, the expectations we form, and the actions we plan inside each of them are different (as Halevy and Halevy note). Reducing them to "hierarchy" loses what makes them what they are, loses the specific behavioral entailments we need to know to navigate them. Plopped down in a different cultural setting, knowing that the structure I am in is "authority ranked" helps only a little. Put differently, taxonomies may reveal the skeleton of social relations, but acting within a complex social world requires putting meat on those bones. This is the reason for the main point of disagreement as to what lies in a generative grammar of institutions.

In contrast to a grammar like Thomsen and Sheehy-Skeffington's, which specifies primitive role relations, we posit a leaner grammar that simply requires the capacity to productively place roles in graph structures of arbitrary size, leaving the exact content within roles and between them more flexible. Grammars like those being developed in computational models of mentalizing might offer a promising parallel (e.g., Velez-Ginorio et al., 2017). Within mentalizing, mental states such as beliefs and desires are propositional attitudes that can take on any content. The grammar-like structure does not capture the *content* inside the mental states, but rather how mental states transition and how they can be composed productively. This includes, for example, representing beliefs about desires, or representing desires with logical connectives (e.g., John wants coffee or tea, but not both). In the same way, the institutional stance could be supported by a lean grammar that simply generates graph structures of roles, marking their kinds and connections. The content inside each role could then fall outside of the formal grammar, either via more general conceptual machinery or perhaps a more specialized internal grammar of norms and scripts. Indeed, in some recent work we have shown how this combination of a general graph-grammar, along with culturally specific behavioral scripts, can capture how people infer the internal structure of social collectives, predict future behavior of individual role occupiers, and even plan for how to influence them (Davis et al., 2026).

Adjudicating between these competing views will be hard, but not impossible. The key question is whether every role relationship must always fall cleanly into one of a few set of fixed categories, or whether some role relations can fall between the categories, or outside of them entirely. We think they can. For example, an older and a younger sibling can be communal in some ways, equal in others, and hierarchical in others, all at the same time, all because they are siblings (a kind of combination Fiske himself also thought possible; Fiske, 1991). Other role relations seem to fall outside of the scope of Fiske primitives. For example, a "chaser and chased" relation makes little sense to think about as a kind of communal, authority, market or equality relation.

Warneken and Rosati raise a related observation. They argue that we overestimate the diversity of human social structures by failing to acknowledge that most cluster into a small set of recurring forms like "interest groups" or "governance" organizations. The apparent diversity, on this view, is merely relatively superficial content variation over a limited set of cognitive forms. We agree that human social structures *can* be grouped into a relatively small number of high-level categories. We expect that this occurs because all human societies of a given size and complexity will tend to face recurrent problems of social coordination, such as advancing the shared aims of a subgroup. These dynamics will, at least from a certain level of abstraction, produce institutions that can be taxonomically organized in consistent ways. What we doubt, though, is that this grouping captures what an agent needs to understand to navigate them. Suppose you have been going to a book club and you are about to attend an organized soccer fan club for the first time. Knowing that they are both "shared interest groups" tells you very little about how the new one will organize; the roles that exist, who is allowed to take them on, what is expected of whom, what counts as a local norm violation, and so on. From the perspective of perceiving, inferring, and acting inside a structure, two structures that a coarse taxonomy groups together can be radically different. How often have we really encountered two different social structures that felt indistinguishable beyond the people that happened to be there? We suggest that the answer is: almost never.

Clément and Kaufmann raise a larger worry about our evolutionary speculation. They argue that calling an ant colony or a dominance hierarchy "proto-institutional" is a form of retrospective determinism, where we interpret evolutionary diversity as if it followed a linear trajectory with humans at the endpoint. This is an important worry, but our claim is different. We do not believe in a linear trajectory from insects to mammals to humans. Instead, we see recurrent solutions to the fundamental problem of how to coordinate when conspecifics cannot read each other's minds. Structures for social coordination might evolve independently many times, in many species that are in no way on a trajectory toward human-like cognition (Sumpter, 2010). Our claim is only that these evolved systems of social organizations, once in place, become regularities that minds can learn to model, in the same way that minds model other regularities in their environment. This makes social organization a necessary precursor for the institutional stance, but this does not imply that these systems of organization were teleologically headed toward the generative institutional stance seen only in humans. As an analogy, the evolution of wings in insects probably began with the emergence of small meshes that could be flapped to help dissipate heat. It is useful to talk about these heat dissipators as "proto-wings" to highlight that they are not wings, but necessary precursors for their evolution. We do not think doing so implies that this was always their ultimate destination.

R4. Development of the institutional stance

One thing we and our developmental commentators agree on is that much of the institutional stance is learned. No baby is born with the concept of *cashier*, and a critical part of development is learning what parts of behavior to encode as roles and what parts as particularities of individual minds. We believe that the representational capacity to separate role from role-occupier is evolutionarily ancient. But Tauzin argues that this could be learned, by tracking what behaviors follow particular individuals across contexts and which behaviors are stable within a context across individuals. We agree it is possible to sort behaviors once we can

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

store them in two representational categories. But this mechanism does not explain how the two categories arise in the first place, or how they are distinguished from other classic forms of inference, like attributing behavior to a situation, or to an essentialized social category. Each inference posits a different link between behavior and the underlying attribute. Once all these potential representational categories are in place, we agree with Tauzin that covariation tracking, enhanced by adult labeling of roles, helps children learn the critical content attributed to roles. We also agree with Rhodes and Yilmaz that mental-state reasoning becomes an important tool for inferring role-governed action because roles become most visible precisely when the occupier acts against their own preferences.

Tauzin and Hok and Thomas both suggest that infants begin with representations that are simpler than institutional roles, and that the institutional stance is built on top of them. For Tauzin, the simpler primitives are interactional roles such as *chaser* and *chased*. For Hok and Thomas they are specific individuals embedded in relational roles (perhaps derived from Fiske's taxonomy, such as agents engaged in *communal sharing* versus *equality matching*, as per Thomas, 2026). From our perspective, we're not clear how different these proposals actually are. What unifies interactional roles, relational roles and the roles our article emphasized is that they all represent behavior in terms of the positions agents occupy relative to one another, rather than in terms of who those agents are. This is the bedrock upon which these seemingly alternative proposals must rest. We would need to hear more to understand how, e.g., interactional roles diverge from what we have in mind.

Perhaps the most direct challenge comes from Rhodes and Yilmaz and Noyes, who both argue that the literature on child development is inconsistent with our account. Rhodes and Yilmaz highlight that children are initially overly focused on norms and then begin to prioritize mental states around age 7, pointing out that an early emergence of both stances does not explain why this shift would occur. They suggest that, rather than being constrained to role representations, young children apply a normative stance across many conceptual domains, even natural kinds. We take the point. But this concession remains compatible with the possibility that children represent both kinds and roles in appropriate ways, and that their role-based inferences diverge from their kind-based inferences. Based on other work we cite in our target article, we think this is indeed the case, but we acknowledge that establishing it more definitively is a particularly critical empirical challenge. We are actively tackling this in ongoing work attempting to document specific "signatures" of role-based reasoning that diverge from more well-studied individual- or kind-based reasoning (Baker, Dunham, & Jara-Ettinger, 2024). In particular, a key test is whether children realize that an agent lacking a particular property can gain that property by entering a role and then lose that property when they exit the role, and then that a new agent entering that same role will gain that same property in the same fashion. Although more needs to be done, we again think that children can indeed make these inferences in broadly appropriate ways.

Noyes sees the relatively late emergence of one particular form of institutional understanding as evidence against our account. He highlights his work showing that, before age eight, children do not understand that the power of an institutional object like money, or an institutional action like scoring a point in a game, can be revoked by group consensus. He takes this as showing that the more basic forms of role understanding we emphasize cannot be full-fledged institutional cognition. But this is ultimately a matter of definition. We prefer a more minimalist definition of institutional cognition as an understanding of interconnected roles. The more complex cognition Noyes studies is then built on top of the simpler components we highlight, necessary for it but not sufficient on its own.

R5. The view from inside: participation and normativity

Institutions are not just things we think about; they are also things we are seen as part of and shaped by. Several commentators correctly pointed out that our article did not give enough attention to this first-person, participant side of the institutional stance. After all, the institutional stance does not just let me predict you. It also helps me know what I should do, and in doing so lets me make my own behavior predictable to others. By thinking of myself as occupying a role, I can use what I know about the role to fit into the structure around me and even to come to understand my own place in the world and thus who I am as a person.

Butler and Schmidt and An argue that our proposal treats roles as overly descriptive, and so it fails to capture the normative force of role-linked norms. Norms do not just describe what people *will* do; they say what people – including ourselves – *should* do. We agree this is sometimes true. But we are also willing to bite the bullet and say that it is not always the case. That is, in some instances, the "should" is not doing any work. As outside observers, we can make sense of group behavior by representing the roles and rules that structure it, without attributing a binding *ought* to participants. This is the third-party analysis the framework was built to support – much like ethologists can describe another species' role structure without supposing that the animals experience their behavior as in any sense obligatory.

That said, we agree that becoming part of a social system is more than simply applying the third-party stance to yourself, and it often does induce a sense of obligation. Again, though, we think we differ in our take on how much work a richer sense of obligation performs. In many cases, prudential motivations may well suffice. Violating norms disrupts smooth social functioning. It imposes costs on others by defying their expectations, and on the violator too, because interactions have more friction and deviations invite sanctions. One evolutionary route to stable group behavior is to build creatures that *feel* a joint commitment in such cases, and our species seems to have taken it. But the same surface pattern can also emerge from a behavioral equilibrium in a coordination game, without those deeper feelings: We know what to do, we have no reason to think we could do better by doing otherwise, and so we keep doing what is expected (Guala, 2016). Indeed, in line with recent evidence in humans and great apes (Sánchez-Amaro et al., 2026), we suspect that this more prudential form of rational maximizing may underlie more collaborative behavior than is sometimes assumed. In other words, although the "ought" may often be present in structured social life, the interpretative stance need not depend on it to do its explanatory work.

An interesting connection comes from Fincher, who argues that decentralized groups sometimes need to socially organize with ad hoc roles with no explicitly agreed-upon process for who gets to occupy a role. For instance, a group of people doing a joint task might all sense they need someone to act as their leader, and they might represent this role, but do not have any agreed upon process for who gets to take on that role. This creates the challenge of determining who truly occupies the role. In this same example, someone might start acting like the leader even when everyone is not inclined to accept it. They suggest this can be solved through roles whose enactment requires showing commitment at personal

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

cost (what she calls "sacred values"). Based on this, they argue that institutional reasoning requires both representing norms that govern behavior (compensatory) and roles that define who is a legitimate role occupant (constitutive). Our proposal talked about entry conditions, but in a way that only seemed a clear match for more formalized systems. In less formal ones, we implicitly suggested that, from the perspective of an observer, to occupy a role is to act as if you are in it. We agree this may not always be enough, neither for the formal creation of social organization (Section R7) nor for understanding behavior. There are many cases where we might see an agent enacting a role and believe it is more appropriate to represent them as *wishing* they held that role rather than actually holding it. The constitutive norms needed to understand this may be a cultural innovation that takes longer to develop and understand, as Noyes' data suggest.

Baumeister et al. suggest that our account can illuminate first-person experiences of a different sort, related to belonging and the basic need to find one's place in the social order. Modern societies, they note, offer so many possible roles that finding one's place has become harder. Earlier in human history, societies were simpler, with fewer, more rigid, and less voluntary roles. This allowed roles to give you a fixed anchor for understanding who you are and where you stand in relation to others: your self was, at least in large part, your role. As societies grew in complexity and individual autonomy became more valued, roles became more fluid and more a matter of choice. Rather than being defined by the roles we are assigned into, roles become things we must seek out as part of our personal quest for meaning. We were not thinking about these more existential aspects of life when we developed our framework, but we find the suggestion intriguing. It points to a need to connect our view more explicitly to affective processes, and to other sources of individual and cultural variation, such as cultural tightness versus looseness (Gelfand, Harrington, & Jackson, 2017), which helps to capture the degree to which roles are assigned to you versus freely chosen, as well as the extent to which you can freely modify aspects of the normative constraints entailed by a role you have chosen.

R6. The dark side of the institutional stance

Our target article ended on a somewhat cheerful note, suggesting that the institutional stance can help us notice entrenched systems of subordination and exploitation that damage society but are easy to miss. If you see a system of relations as natural, efforts to change it can seem futile. But if you can reconceptualize a social structure as a human-made system of roles and relations, and analyze what those roles are and how they came to be, change becomes accessible.

Asma, Szocik and Rozmus, and Ritchie all suggest that this picture is too rosy, for several reasons. When we interpret others through an institutional stance, we may begin to enforce role expectations on them, and they might begin to accommodate their behavior accordingly. And when the roles we occupy feel natural or invisible (consider gender) the constraints they impose become easy to miss, so that role-driven action can get (mis)read as the expression of free preference. We take this to be Asma's central point, and we are compelled by it. Social change requires historical and moral perspectives that go beyond what the institutional stance provides. At the same time, the institutional stance does enable us to take new perspectives on old structures, which we think is necessary for structural change. Social hierarchies are pernicious and prone to reification. But we can and do work against them. In this we may stand alone in the natural world: how often has a chimpanzee troop decided to reconstrue or restructure their male dominance hierarchy?

Szocik and Rozmus note that the dehumanizing side effect of the institutional stance is not evenly distributed. Gender-coded roles such as those in care, service, and administrative support are structured to maximize replaceability, which reduces our tendency to think of the people in them and to represent their minds. We agree that the costs of the institutional stance fall on individuals in asymmetrical and even unjust ways, and thus are happy to endorse their design interventions, such as using names rather than role labels, as actionable directions to come out of this literature.

In a related vein, Ritchie questions whether the institutional stance can really be used to analyze large-scale social-structural systems like race or gender. She rightly points out that our article's main examples involve individuals in roles, and transferring that model to groups in a hierarchy is not trivial. The idea that agents can, in principle, move into and out of roles, and one occupant can substitute for another, fits cases like baristas well, but does not work for members of essentialized social kinds. We did not mean to equate these two levels of analysis, but we do think they share something cognitively and conceptually. The philosophical work that reinterprets naturalized social kinds as positions in a hierarchy (e.g., Ásta, 2018; Haslanger, 2000) and the use of the institutional stance to analyze a specific social structure share an interpretive question: is someone being treated this way because of the position they occupy, or because of who they are as an individual? For individual roles and social kinds alike, we can recognize that someone is being treated based on the social position they occupy, and not based on their individual identity. We can further use those observations to imagine alternative arrangements or even the dissolution of the structure as it currently exists. But the conditions for entering and leaving differ sharply across these cases, and we were too hasty in drawing these parallels.

More generally, the optimistic note at the end of the target article was not meant to imply that humans usually or even frequently apply the institutional stance to social kinds and subtle structures. Nor was it to say that, on balance, the institutional stance is a force for social change rather than for entrenching exploitation. It may well be the latter. But humans do sometimes engage in the work of seeing through and seeking to change naturalized structures, and when they do, we think the cognitive machinery we describe is a necessary ingredient in that process.

R7. Intuitive versus formal institutions

A major source of tension across the commentaries lies in assumptions about what an institution actually *is*. The question has no single answer. A legal institution analyzed in jurisprudence is a different kind of thing from what a layperson has in mind when they think of a legal institution, and both are different from the cognitive representations that let a child or an adult navigate the social structures around them by representing them as sets of interlocking roles. The institutional stance is a theory about that third thing. By *institution*, we mean a cognitive structure, a representation, built from an interconnected set of roles with persistence and continuity. Each role bundles norms and interaction scripts. And when an agent takes on a role, they enact the model of agency it defines.

Several commentators were dissatisfied with this and pushed back on the institutional stance by importing criteria from formal institutions. Noyes argues that a genuine institution must be a self-

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

reproducing structure that retains its identity across changes in its role-occupants. He is joined by Clément and Kaufmann, who argue that institutions require Searlean collective acceptance of status functions ("X counts as Y in context C"). Johnson argues that institutions are the rules of the social game, and that rules are prior to roles. Gantman et al. agree, specifically pointing to formalized rules as the raw material out of which institutions are made.

We agree that these comments identify features that may be necessary parts of a proper theory of formal institutions. In doing so, they highlight the gap between what is sufficient to understand social collectives in terms of roles, and what may be needed to formalize those collectives or create new kinds of social reality. But we are not persuaded that all these features belong inside the intuitive theory we are trying to characterize. When you watch two people interact and represent them as guided by roles of caregiver and cared-for, or barista and customer, you make sense of the interaction perfectly well without representing a self-reproducing structure, explicitly representing collective buy-in, or wondering what is in the Starbucks Employee Handbook. You may do more of this upon reflection or in more specific contexts. But we doubt that any of this is what the intuitive theory does when it interprets social collectives in real time. Asking the intuitive theory of social collectives to do all this work is akin to asking intuitive psychology to deliver a formal theory of neuroscience.

Nonetheless, our article was too quick to suggest that the institutional stance gives rise to real-world institutional structures. To clarify here, we see the institutional stance as *necessary* for the more demanding cognitive achievements the commentators raise, but not *sufficient*. Reasoning about a self-reproducing institution that persists across changes in role-occupiers requires the ability to represent roles in the first place, and that is what the institutional stance provides. For example, we see the cognitive transition between ages four and nine that Noyes discovered as a conceptual elaboration of this core intuitive theory. This same logic applies to Searlean collective acceptance that Clément and Kaufmann emphasize (and that we discuss as a likely outgrowth of our institutional stance in the target article). For these commentators, naive sociology rests on the recognition that people and things can gain new properties via collective attribution. To an extent, we agree. Many role-linked properties can emerge out of cultural practices through collective action. But we also think that social structures can emerge more organically. The push and pull of social life often leads to stable equilibrium patterns of interaction that were never collectively agreed upon (Hindriks & Guala, 2015). In some cases, these patterns may not be collectively recognized, but people can still represent them in role-like ways. This makes us suspect that rooting institutional reality in uniquely human and relatively modern cultural developments such as formalized rules or the emergence of writing (see also Gantman et al.) misses longer-running commonalities in our species' deeper cultural history.

Johnson suggests that rules precede roles, both in how formal institutions are built and as the cognitive foundation of the institutional stance. Formal institutions are certainly governed by rules that sit outside any single role, and our theory said little about these. But a rule alone cannot do anything until its scope is defined. That is, it has to apply to someone in some context. We think that doing this work will require sorting people based on where they stand within a structure, which is, in the end, to reinvent the roles that we began with. In other words, to make rules do the work these commentators want, you would need to tie them to social positions, and to the relations between those positions. In doing so, you will have arrived back at a position much closer to ours. That said, we do agree that some rules apply broadly to any member of a social collective and people intuitively understand that. It is an open question whether these rules are represented outside the institutional stance, or whether we instead attach them to a broad "member" role.

Gantman et al. raised a related distinction between norms and rules. They point out that norms are messy, often underspecified, and have lossy transmission. Rules, by contrast, are a cultural innovation that explicitly specify how agents should engage with one another. Our article did not draw this distinction at all, but we think it is a useful one. The jump from norms to rules might be a critical step that serves two functions. First, it allows humans to build formal and increasingly complex systems of social organization. Second, once learned, it allows people to embed them into the institutional stance, to reason about roles whose behaviors are explicitly codified. Gantman et al. point out that the rigidity of rules, and the consequences of breaking them, make role-driven behavior more predictable. If so, this cultural innovation regarding rules might lead to an even further decline in the necessity of mentalizing.

Acknowledging the gap between the everyday psychology of the institutional stance and the study of formal institutions is not a cop-out. It can help us predict when and why people find some institutional or economic contexts difficult to think about. Białek et al. review a set of puzzles in behavioral economics that can all be precisely explained by a gap between the ground-truth of complex settings and an intuitive theory that fails to accommodate them. Forcing this fit leads us to place complex behaviors into different structures, such as into communal or exchange logics that are an imperfect fit. In this way, the gap between formal institutions and our naive theory helps explain why the real structure of complex institutions is hard to understand, and why we need formal fields like economics and political science to effortfully determine the best formal way to think about them.

That said, we may be underestimating what the intuitive theory contains. One way to extend our original proposal is by letting people represent roles nested inside one another within a graph. That is, each node is still a single irreducible role, but clusters of nodes can be grouped together under a broader role. If we continue scaling up in this manner, we can eventually represent all nodes in the graph as falling under the widest role representing being a member of the collective. We think this may help with Johnson's concern that some rules seem to exist outside of any role. They could be captured, at least in part, by attaching them to the broadest "member" role that applies to everyone in the graph. At the same time, the kinds of rules Johnson alludes to, such as who gets to make judgments about intellectual property, are likely outside of our naive theory of social collectives.

Relatedly, inspired by Warneken and Rosati, we suspect that people represent the networks of roles in an institution as having a function, maybe even when they think no role member is necessarily aware of that function. This ability to give functional explanations to social collectives lets us reason about the structure as a whole, its coherence, and ways in which it could be different. In an analogous way to an expectation that agents act rationally in the

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

mentalistic stance, there might be an expectation that agents also organize rationally in the institutional stance, such that we assume that role structures are a rational attempt to solve a social coordination problem. These considerations point the way toward interesting empirical work exploring how observers attribute agency across levels of institutional structure.

R8. Looking ahead

The proposal we introduced in our target article, along with the revisions here, is a starting point, not a complete story. Together with the commentaries, it highlights understudied research directions in social cognition. We end by laying out questions we find both pressing and possible to answer.

First, what are the conceptual primitives and generative structure of the institutional stance? We think the system has two parts. The first is a foundational capacity to represent abstract roles and connect them generatively with no fixed limit – the grammar. Here the critical questions are the constituent parts, the "primitives," of the generative apparatus. Are they a limited set of Fiske-style relations, or, as we have argued, can even the infant's mind represent a much broader repertoire of potential relations, such as chaser and chased, giver and receiver, driver and passenger? The other part is the capacity to insert information inside those roles. These include regulatory and constitutive rules, dispositional mental states, and interaction scripts. This raises a question we cannot yet answer: can anything, in principle, be placed within a role, much like almost anything can be the object of a belief or a desire in the mentalistic stance?

Second, what in this system is uniquely human? Warneken and Rosati suggest we underestimate the richness of nonhuman primate social structures, and, given their expertise, we are inclined to trust them. Even so, we still think that there is a qualitative gap across species. We suspect that a chimpanzee looking at a human society could not understand how it works, and chimpanzees do not seem to critically reflect about how they organize, nor try to fundamentally change the existing organizations. There are, of course, empirical questions here.

A closely related issue is what actually develops. We all agree that development happens, and the commentaries offer many useful, more specific suggestions (along with our own proposal that pretend play is suspiciously focused on roles and might be a big part of how children learn and practice them). Open questions include which parts of the representation are learned and, if any are, what kind of system can generate genuinely new social concepts. And then, how do children learn the vast library of roles that exist in a society, their contents, and how to apply them? And is normativity a broad tendency that gradually gets centered on roles, or something more intimately related to institutional reasoning from the start?

A further question is how the institutional stance relates to mentalizing, and how this relationship works across different timescales. Do people move fluidly between the two stances within a single interaction, or switch only across contexts? Can we predict the balance between them by treating the two as a joint effort to capture the predictability of structured behavior, while accommodating the flexibility humans inevitably show? Do we ever turn off mentalizing completely to rely wholly on the institutional stance?

This also leads to questions about how the institutional stance is implemented in the brain. We speculated that specialized circuits might support it, but, as Tauzin noted, this was only speculation. In many ways, there is little at stake here. Nothing in our proposal requires dedicated brain areas, and a capacity can be important without having its own patch of cortex. That said, the answer would be informative, because it would tell us whether this reasoning shares circuitry with mentalizing, draws on more domain-general areas, or might rely on specialized algorithms of its own.

One of the largest open questions lies in characterizing the gap between using the institutional stance to understand social collectives and actually building complex structures with explicit roles, and in particular formalized institutions. We think developing the computational structure behind the institutional stance will become a cornerstone of any account of how real institutional structures are built. This is because creating social structures requires understanding them, but also because, with a formal computational model in hand, we will be able to see much more clearly what is missing, and so what is additionally required for the creation of formalized institutions. In some ways, we think the version we presented here was already helpful, as the comments show, in highlighting the gap between interpretation and creation. At the same time, we think it is critical to distinguish between the social-scientific project of characterizing commonality and variation across real-world formal institutional structures and the cognitive-scientific project of understanding how the mind uses roles and relations to simplify the problem of social inference at scale. This institutional stance speaks directly only to this second project.

As a final closing question, what is the richer phenomenology running alongside institutional perception? Choosing to interpret a social collective in terms of roles is not to just apply some abstract algorithm over data. It is how we make sense of one of the most important parts of our lives – our social world. These interpretive frameworks shape how we see ourselves and others, our humanity and theirs. They lead us to misrepresent and restrict others, but sometimes, with effort, to recognize and even combat injustice. These are not abstract inferences. They determine what it feels like to be human in our deeply social world.

Financial support. This work was supported by NSF award BCS-2438827.

References

Ásta. (2018). *Categories we live by: The construction of sex, gender, race, and other social categories*. Oxford University Press.

Baker, A., Dunham, Y., & Jara-Ettinger, J. (2024). Roles guide rapid inferences about agent knowledge and behavior. *Proceedings of the Annual Meeting of the Cognitive Science Society* 46.

Blader, S. L., Shirako, A., & Chen, Y. R. (2016). Looking out from the top: Differential effects of status and power on perspective taking. *Personality and Social Psychology Bulletin*, *42*(6), 723–737.

Chandra, K., Ragan-Kelley, J., Tenenbaum, J., & Saxe, R. (2026). Assembly Instructions for the Modular Mind. https://osf.io/xqtbp_v2

Davis, I., Jara-Ettinger, J., & Dunham, Y. (2026). Inferring the internal structure of groups through the integration of statistical learning and causal reasoning. Nature Communications.

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.

Galinsky, A. D., Magee, J. C., Inesi, M. E., & Gruenfeld, D. H. (2006). Power and perspectives not taken. *Psychological Science*, *17*(12), 1068–1074.

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press

Gelfand, M. J., Harrington, J. R., & Jackson, J. C. (2017). The strength of social norms across human groups. *Perspectives on Psychological Science : A Journal of the Association for Psychological Science*, *12*(5), 800–809. https://doi.org/10.1177/1745691617708631

Guala, F. (2016). *Understanding institutions*. Princeton University Press.

Haslanger, S. (2000). Gender and race: (What) are they? (What) do we want them to be? *Noûs*, *34*(1), 31–55. https://doi.org/10.1111/0029-4624.00201

Hindriks, F., & Guala, F. (2015). Institutions, rules, and equilibria: A unified theory. *Journal of Institutional Economics*, *11*(3), 459–480.

Johnson, S. C., Dweck, C. S., & Chen, F. S. (2007). Evidence for infants' internal working models of attachment. *Psychological Science*, *18*(6), 501.

Liu, S., Karakose-Akbiyik, S., Outa, J., & Kim, M. J. (2026). How physical information is used to make sense of the psychological world. *Nature Reviews Psychology*, *5*(1), 59–73.

Powell, L. J., & Tompkins, R. (2026). Relational models are too elaborate for basic social relationship reasoning. *Behavioral and Brain Sciences*, *49*, e44.

Prasada, S. (2026). *Conceptual form: The hidden structure of commonsense concepts*. MIT Press.

Sánchez-Amaro, A., Stengelin, R., Maurits, L., Jabeen, Z., Ramoser, C., Grueneisen, S., & Haun, D. (2026). Human children, chimpanzees, and orangutans act rationally in an adaptive collaborative choice task. *Journal of Experimental Psychology: General*, *155*(6), 1536–1549. https://doi.org/10.1037/xge0001923

Saxe, R. (2005). Against simulation: The argument from error. *Trends in Cognitive Sciences*, *9*(4), 174–179.

Sumpter, D. J. (2010). *Collective animal behavior*. Princeton University Press.

Thomas, A. J. (2026). Cognitive primitives, learning, and the structure of early social relationship cognition. *Behavioral and Brain Sciences*, *49*, e56.

Thomsen, L., & Carey, S. (2013). Core cognition of social relations. In M. R. Banaji & S. A. Gelman (Eds.), *Navigating the social world: What infants, children, and other species can teach us* (pp. 17–22). Oxford University Press.

Velez-Ginorio, J., Siegel, M., Tenenbaum, J. B., & Jara-Ettinger, J. (2017). Interpreting actions by attributing compositional desires. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 39)

Wellman, H. M. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology*, *15*(6), 728–755.

Xu, K., Srivastava, A., Gutfreund, D., Sosa, F., Ullman, T., Tenenbaum, J., & Sutton, C. (2021). A bayesian-symbolic approach to reasoning and learning in intuitive physics. *Advances in Neural Information Processing Systems*, *34*, 2478–2490.

https://doi.org/10.1017/S0140525X25103427 Published online by Cambridge University Press