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

Downloaded from https://www.cambridge.org/core. IP address: 172.59.14.145, on 18 Sep 2026 at 16:35:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0140525X25103427

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

Downloaded from https://www.cambridge.org/core. IP address: 172.59.14.145, on 18 Sep 2026 at 16:35:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S0140525X25103427

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

Downloaded from https://www.cambridge.org/core. IP address: 172.59.14.145, on 18 Sep 2026 at 16:35:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S0140525X25103427

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

Downloaded from https://www.cambridge.org/core. IP address: 172.59.14.145, on 18 Sep 2026 at 16:35:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S0140525X25103427

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

Downloaded from https://www.cambridge.org/core. IP address: 172.59.14.145, on 18 Sep 2026 at 16:35:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S0140525X25103427

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-

Downloaded from https://www.cambridge.org/core. IP address: 172.59.14.145, on 18 Sep 2026 at 16:35:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S0140525X25103427

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

Downloaded from https://www.cambridge.org/core. IP address: 172.59.14.145, on 18 Sep 2026 at 16:35:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S0140525X25103427

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.

Downloaded from https://www.cambridge.org/core. IP address: 172.59.14.145, on 18 Sep 2026 at 16:35:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms.
https://doi.org/10.1017/S0140525X25103427

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.

Downloaded from https://www.cambridge.org/core. IP address: 172.59.14.145, on 18 Sep 2026 at 16:35:48, subject to the Cambridge Core terms of use, available at https://www.cambridge.org/core/terms. https://doi.org/10.1017/S0140525X25103427