Philosophers' Imprint

VOLUME 26, NO. 7
FEBRUARY 2026

A BEHAVIORAL THEORY OF SOCIAL INSTITUTIONS

Megan Henricks Stotts

McMaster University

doi.org/10.3998/phimp.7180

1. Introduction

When writing about the metaphysics of social institutions, philosophers rightly emphasize that institutions' existence and nature are "up to us" in a deep and important way. Objects such as insects and stars could exist with the same basic nature in the absence of human activity; corporations, governments, and religious institutions could not. This feature of institutions makes it natural to think that their existence and nature depend on our thoughts in some way, and accounts of institutions that appeal to mental phenomena dominate the literature.[1] But there are also important ways in which institutions' existence or nature can be less "up to us" than we might initially expect, which tend not to be emphasized. I'd like to try starting with these kinds of cases—that is, with cases that highlight ways in which institutional reality can get away from us. In my view, these cases push us toward a behavioral approach to social institutions. According to this approach, institutions are made up of behavior that clusters into copied roles, and the individuals who engage in that behavior, where all of the roles together promote some particular result(s).[2] I'll motivate this behavioral theory of institutions, discuss and develop its features, and finally, compare it to other accounts.

It is important to note that by 'social institution' I mean a complex, structured social entity such as the Colombian government, the Catholic Church, the U.S. National Science Foundation, Yamaha Corporation, or Major League Baseball. I will offer a behavioral account of just this specific kind of social entity, and not of social entities, *tout court*. For instance, social norms, laws, and social groups fall outside

1. See, *e.g.*, Miller (2001), Thomasson (2003), Searle (2009), Tuomela (2013), Guala (2016), and Ludwig (2017).
2. Some aspects of my behavioral theory of institutions were introduced in Stotts (2024), and various threads from that paper reappear in what follows, to be developed and justified in more detail. The main project of Stotts (2024) was to argue against accounts of social institutions that appeal to mental phenomena, ending with a preliminary sketch of the behavioral alternative I develop in detail below.

of the present paper's scope.[3]

2. A Behavioral Theory of Social Institutions**

We'll begin with three cases in which institutions seem to be less "up to us" than we might expect. Though they are inspired by real-life examples, they are described in hypothetical terms, to allow for the stipulation of certain details.

Racist Courts

Imagine a court system explicitly founded to promote justice. Most of its participants (such as attorneys and judges) proudly believe it succeeds. Imagine that one racialized group makes up a small percentage of the society's general population but a large percentage of the incarcerated population, due to factors such as intercultural misunderstanding in the courtroom, harsher sentencing, and a long history of unjust economic disadvantage. This court system oppresses members of that racialized group, and is thus a racist institution, despite the fact that few of its participants think of it as such. Here, one aspect of the institution's nature is not "up to" its participants: even though participants share beliefs about the institution's nature, its nature is not what they believe it to be.

Watchdog

Imagine an athletic institution organizing professional competition in a high-contact team sport. Serious injuries are common during games. The institution's leadership makes a big show of publicly purporting to create a new institution within their larger one: they announce that certain physicians are members of a new "watchdog" institution, responsible for assigning one physician to attend each game in order to stop the game whenever they deem serious injury to be likely. But then, as a matter of fact, no physicians are assigned to any games, and play is never stopped. Spectators and the athletic institution's leadership, however, assume that physicians do monitor the games. Here, institutional reality is again not "up to us" in an important way: people share beliefs about an institution's existence and nature, and yet it fails to exist at all (adapted from Stotts, 2024).[4]

Disputed Government

Imagine a longstanding, stable government in a country that has deprioritized civic education. The citizens have divergent beliefs about the structure of their government and how it operates, including its electoral process. After a contentious election, approximately half of the citizens genuinely believe that Candidate A is the individual selected by the country's electoral process, and the other half believe that Candidate B was selected. The country's longstanding electoral process implies that Candidate A is the leader, and Candidate A does, in fact, go on to enact the duties and privileges of the country's leader. But nonetheless, half of the population believes that Candidate B is the country's actual leader, unjustly prevented from fulfilling their duties. De-

3. Turner (1994, pp. 103–105) promotes a behavioral project of larger scope by arguing that the best way to understand society as a whole may be to see it as made up of just observable behavior, due to the problems we face when trying to spell out some notion of shared mental entities underlying that behavior. My project here is narrower than Turner's, focusing just on social institutions. I think that accounts of some social phenomena can successfully appeal to shared mental entities.

4. Watchdog is likely the most controversial of the three cases. For example, Hindriks (2013, p. 429) says that "an institution can exist without being instantiated," offering the continued existence of the papacy during a papal interregnum as his example of this phenomenon. I would describe Hindriks's case as an institution continuing to exist while one role is temporarily unoccupied, rather than an institution existing while completely uninstantiated. My hope is that Watchdog offers a strong reason to think that completely uninstantiated institutions are impossible. It seems natural to say that the agents in Watchdog *decided* (or pretended to decide) to create an institution, but failed to actually make one. But a reader who finds Watchdog unconvincing may focus on just the other two cases.

> spite this disagreement, the government continues to function seamlessly, with Candidate A as leader, but with half of the population thinking of this as a sham. Here again, the institution's nature is in an important way not "up to us:" it exists with a determinate nature (*i.e.*, with a certain actual electoral process, and with Candidate A as its leader), despite widespread disagreement among participants about what that nature is.

Institutions, then, are in various ways less "up to us" than we often assume. We might all share beliefs about the existence and nature of an institution and it can either fail to have the nature we think it does (as in Racist Courts) or fail to exist at all (as in Watchdog). These two cases highlight ways in which we have less power than we might think to *determine* the existence and nature of institutions. On the other hand, an institution can exist with a determinate nature even though we disagree about what that nature is (as in Disputed Government). This last case highlights a way in which institutions and their natures are less *dependent* on us than we might suppose.[5]

To begin to construct a theory on the basis of these cases, we need to find features that unite Racist Courts, Disputed Government, and paradigmatic examples of institutions, and which are absent from Watchdog. My initial suggestion is the following: in all cases in which an institution exists, there is behavior that clusters into roles, working together to promote some particular result(s). In Disputed Government, although there is disagreement about the electoral process and thus about who is the leader, only one person is actually able to enact the *leader* role in ways that promote the government's results (such as peace and order within the society it governs). And in Racist Courts, there is behavior that clusters into roles such as *attorney* and *defendant*, of which everyone is aware, but what everyone is not aware of is one of the results that the behavior promotes: over-incarceration of a certain racialized group. Moreover, in cases that most theorists would treat as paradigmatic, such as a typical university, there is agreement about the institution's nature and purpose, but there is also actual behavior that clusters into roles, together promoting certain results (specifically, knowledge generation and transmission). On the other hand, in Watchdog, we can see that although there is agreement about the nature and purpose of the purported institution, no institution is actually created because the requisite behavior is absent.[6]

However, the picture of social institutions just suggested cannot differentiate institutional reality from mere social reality.[7] Imagine a neighborhood in which one person, Juana, is so kind and witty that her company is unusually enjoyable. Juana's neighbors visit her more frequently than they visit anyone else, to seek social interaction with her. Imagine too that Juana tends to refrain from seeking social interaction because her visitors meet her socialization needs. Here we can distinguish one behavioral role occupied by Juana, and another occupied by the other neighbors, with all of the roles together promoting the fulfillment of Juana's socialization needs. On the picture we've been considering, this would be a Juana-centered social institution. But it's

5. I am using 'determined by' and 'dependent on' in an ecumenical sense and not in, for instance, the technical sense Epstein (2015, pp. 106*ff.*) introduces.

6. One might wonder whether the three cases are better addressed by Thomasson's (2003) strategy of identifying them as downstream metaphysical consequences of primary institutional phenomena, where the latter are grounded in collective intentionality. Thomasson offers the example of a recession: "the existence of recessions certainly depends on collective intentionality (since there could not be economies at all without monetary systems and trade transactions, all of which require collective intentionality), but requires no beliefs of anyone's *about recessions*" (p. 288). Racist Courts is amenable to that kind of treatment (*cf.* Thomasson, 2003, p. 276), but the same cannot be said for Watchdog or Disputed Government. Thomasson's strategy can explain how additional institutional facts beyond those collectively accepted can come to obtain, but in Watchdog the problem is that collective intentionality purports to generate institutional facts that do not actually obtain. And Disputed Government involves lack of agreement about foundational facts within an institution (*e.g.*, its electoral process), rather than about downstream consequences of collectively accepted institutional facts.

7. On that distinction, *cf.* Wilson, 2007, pp. 148–149.

clear that all we have here is the merely social phenomenon of Juana being popular. Similarly, consider social phenomena among nonhumans. Ants in a colony evince behavior that clusters into roles (such as *queen*, *nurse*, and *forager*), together promoting the ants' survival and reproduction (Gullan and Cranston, 2014, p. 337). But we wouldn't want to call an ant colony a social institution.

We need to differentiate institutional reality from mere social reality while still properly categorizing Racist Courts, Watchdog, and Disputed Government. There are three features that I'll propose distinguish social institutions from mere social reality, falling into two main categories: *momentum*, and *optionality*. I'll motivate these features and then argue that all three are needed to differentiate institutions from mere social reality.

First, momentum: social institutions involve one particular way of doing something getting some momentum behind it and carrying itself forward. One aspect of this momentum, in my view, is that some of the behavior within an institution is *copied* from past behavior: things are done a certain way because that is how they were done in the past. I follow Millikan (2005, pp. 3, 5, 31) in seeing copying as a matter of causal connections among features of particular instances of behavior, which does not require any particular intentions or even awareness from the copying individual (*cf.* Stotts, 2023, p. 2123). If I flick my wrist in a particular way when cooking pancakes because my father did so, then I am copying him, though I may never realize I am copying him. As I see it, individuals enacting institutional roles are not making completely fresh decisions about how to act within their roles; rather, they are causally influenced by what has been done in the past.

But there is more to institutional momentum than copying, leading to a second differentiating feature for social institutions: the vacating and passing on of institutional roles. In my view, institutional momentum is strong enough to go beyond any particular set of individuals. Once something has become an institution, as opposed to just some particular individuals' way of doing things, it goes on to influence the behavior of other individuals as well, leading roles to be vacated and passed on. In this way, the institution takes on a life of its own (*cf.* Millikan, 2017, p. 20).

In addition to institutional momentum, institutions have a distinctive kind of *optionality*: when people form a new institution, they are instituting one particular way, *among others*, of doing something. Their way of promoting a certain result was not the only option, and thus it needed to be instituted in order to become their way of doing things. This motivates the third key feature of social institutions: equally accessible alternative ways to promote the institution's result(s).[8] Crucially, the alternatives need not be equally *good*. If there is one objectively best form of government, governments taking that form would still be social institutions. There would still be the right kind of optionality to the process of instituting that particular form of government, as opposed to others that might have equally easily been instituted instead.

So, there are three features that I take to be distinctive of social institutions: copying, the vacating and passing on of roles, and alternatives. Having provided some initial motivation for these features, I'll show how they differentiate social institutions from mere social reality.

First, the ant colony. Ants do vacate and pass on roles as they age and as the colony's circumstances change. But ants do not copy each other. They do not learn how to be foragers or nurses from other ants; rather, they are just responding to changes in their environment, in ways presumably shaped by their genes (Gordon, 2010, pp. 30–44). They also do not have equally accessible alternatives to living in colonies (p. 15).

Second, the popularity case. That case does involve equally accessible alternatives: there are other ways of promoting fulfillment of Juana's socialization needs, such as offering her free transportation to visit friends in other neighborhoods. But, the popularity case does not

8. The notion of accessibility used here comes from Stotts (2017, p. 875), where "[a]n option's level of accessibility to an individual or group is the degree of ease with which that individual or group can encounter and use it."

involve copying. People enact the roles of *popular person*, *friend of popular person*, or *seeker of popular person's friendship*, but without copying each other in doing so. They are just (we're supposing) independently motivated to pursue Juana's company. The popularity case also does not involve the vacating and passing on of roles. It has all of the optionality of an institution, but none of the momentum.

At this stage, there may be some doubt about whether all three features are necessary. Neither the ant colony nor the popularity case involves copying, so couldn't we require just copying and omit the other two features? Or, given that the ant colony does not involve equally accessible alternatives and the popularity case does not involve roles that are vacated, couldn't we require just those two features and eschew copying? To assuage this worry, I'll argue that no two of the three features can rule out all cases of mere social reality, without the addition of the third.

First, if we required both equally accessible alternatives and the vacating and passing on of roles, but did not require copying, a series of transient social arrangements could count as an institution. Imagine that at some particular beach on Saturday, every two hours a different group of four people happen to play Bocce on the same stretch of sand. So, the players' roles get vacated and passed on, and there is optionality: there are other options for leisurely outdoor competition. But, there are no causal connections between the successive groups of people. This mere coincidence would amount to a social institution, just as much as if the players had organized a weekend Bocce club. Requiring copying rules out this kind of case.

Second, consider what would happen if we required both the passing on of roles and copying, but not equally accessible alternatives. If we were then to discover that ants do, in fact, learn from each other how to forage or build nests, we'd have to count ant colonies as social institutions. I'm not averse to countenancing social institutions among non-humans, but the mere addition of copying to ant colonies seems insufficient. If their options are to learn from older ants how to live together in an ant colony, or to die, that seems fundamentally different from the way humans establish social institutions in the face of multiple options.

And finally, consider what would happen if we required equally accessible alternatives and copying, but not the passing on of roles. This last requirement is perhaps the most controversial of the three. But consider, for instance, a group of people stranded on an island who organize themselves into *leader*, *judge*, and *citizen* roles. Over time, all participants form habits in enacting their roles, thus being causally influenced by their own past behavior as role-holders (which does count as copying). They, of course, had other options for organizing themselves, and all of the roles together promote the result of peace among them. In my view, this arrangement falls short of institutionality. They have certainly created a social structure, but whether it will have the staying power to become an institution has yet to be seen. If they all die before anyone has passed on a role, it seems to me that we can say that they organized themselves effectively while they were alive, but they did not establish a new institution. On the other hand, if some of them have children to whom they pass on their roles, or even if just one role shifts from one person to another, then we have a social structure with the right kind of momentum behind it to count as an institution. It has some amount of staying power, beyond the habits of any particular set of individuals. Institutions may vary in how much of this kind of staying power they have, but they all have it to some degree.[9]

Here, then, is the account of social institutions I propose, which I will call the Behavioral Theory of Institutions (BTI):

9. I do not take the distinction between social institutions, and social structures that fall short of institutionality due to roles not having been passed on, to imply that such social structures are less deserving of respect. A new leadership structure may initially not qualify as an institution, but that does not mean its authority is necessarily illegitimate or that it is unworthy of respect on the international stage.

Behavior that clusters into roles, and the individuals who engage in that behavior, together compose a social institution, provided that:

1. All role-holders copy some of the behavior associated with their role from the past behavior of occupants of the same role,
2. At least one role has been vacated and passed on within the institution's history, and
3. There is at least one result promoted by all roles in the institution in a way that is causally inter-enabling, where here are other equally accessible structures of roles that would promote the same result(s).

The BTI captures the key elements that our cases other than Watchdog shared (namely, behavior that clusters into roles, working together to promote some particular result(s)). It also has the three key features that differentiate social institutions: copying (in condition 1), the passing on of roles (condition 2), and alternatives (condition 3). I'll offer some clarificatory comments on various aspects of the account.

As a brief comment about condition 1, I want to emphasize that in some institutions, the only past behavior a role-holder copies may be their own. This can happen in institutions where only some roles have thus far been passed on. We can be causally influenced by our own past behavior, and thereby copy ourselves in the required sense.

Regarding condition 2, when I refer to the "institution's history," I am talking about its history in a broad sense, where things that happened within a social structure that was not yet an institution but would later become one can be considered part of that eventual institution's history. So, there will have been a moment when there was a social structure that satisfied conditions 1 and 3, but not yet condition 2, and then a role was vacated. This was not yet a social institution. But then, when a new individual took on that role, it became an institution. That first instance of the role being vacated and passed on is part of the institution's history, even though it occurred before the social structure met all criteria for institutionality.

Condition 3 requires more extensive discussion. First, a few brief clarifications. The result(s) which the behavior within an institution promotes must be specified in non-institutional terms. Otherwise, we will not have an account of where institutional reality comes from. Moreover, results cannot be disjunctive (so, we cannot say that the National Science Foundation promotes the result of "scientific discovery or artistic achievement"). Individual results must be considered individually, to determine whether each is in fact promoted by the institution.

More should also be said about the notion of "promoting" a result. To promote a result is just to make that result likelier to obtain. The results an institution promotes (in the way condition 3 requires) are a central part of its nature. They are what the institution *does*—its contributions to the world. Participants in an institution may be unaware of the results it promotes; in fact, institutions often have unintended results. For instance, participants in a vast number of institutions, from the Catholic Church to Major League Baseball to the American Philosophical Association, causally inter-enable each other to promote pollution (among other results, of course). This, according to the BTI, is as much a part of their nature as the results that were intended.

Also related to condition 3, I'd like to say a bit more about the notion of alternatives being equally accessible. When an institution is well established, its ways of doing things often become much easier to use than the alternatives, sometimes to the point at which there no longer seem to be real alternatives at all. This feature is part of what makes institutions so powerful, for good or ill. So, when I say that an institution must have equally accessible alternatives, what I mean is equally accessible once we have factored out the influence of the copying that has happened within the institution, or in other words, once we have factored out institutional momentum. This still captures the optionality that characterizes institutions, at least at the time of

their founding.[10]

The final aspect of condition 3 that we'll discuss is its requirement that promotion of the institution's result(s) be done by "all roles in the institution in a way that is causally inter-enabling." We'll simultaneously discuss the nature of roles, as the BTI posits a tight relationship between institutional roles and institutional results. To simplify the discussion, we'll leave conditions 1 and 2 aside for now. So, we'll be considering a pre-institutional social structure that satisfies condition 3, where copying has not occurred and no roles have been passed on.

A pre-institutional social structure, and an initial set of roles within it, emerges when it comes to be the case that there is a set of individuals engaging in behavior that is interconnected in two ways, which we'll discuss in turn. First, the behavior must be interconnected by a relation I will call 'reciprocal exchange of behavior' (REB), which is central to my understanding of roles. When one individual engages in behavior of kind *A* and refrains from behavior of kind *B*, and another individual engages in *B* and refrains from *A*, those two individuals stand in the REB relation (Stotts, 2024). Each individual who is part of the social structure is such that, for each other individual, they either stand in the REB relation to that second individual, or they stand in the REB relation with a third individual who is part of a chain of REBs that ultimately connect to that second individual. This is what is meant by the BTI's statement that the behavior in an institution "clusters into roles," where each role is defined by its REBs (that is, by various behaviors in which holders of that role engage while others abstain, and the behaviors from which its holders abstain while holders of other roles engage in them). For example, imagine a professor (at an emerging university, where roles have not been copied or passed on) who generates syllabi, gives out assignments, and submits grades while others (*i.e.,* the students) refrain from those activities but do attend classes, take notes, and write essays, while the professor refrains from their activities.

The second way in which the behavior in such a pre-institutional social structure must be interconnected is by causally inter-enabling each other to promote some result(s), in line with condition 3. This is my way of spelling out the intuitive idea, used earlier in the paper, of roles promoting a result "together." It means that there is at least one result that all roles promote, where each role's ability to promote that result has as a partial cause some of the behavior associated with some other role.[11] So, for instance, students' behavior of attending classes is a partial cause of the professor's success in promoting knowledge transmission, as is maintenance staff's behavior of removing debris from the classroom. Only roles that contribute to the pre-institutional social structure's result(s) in this way are roles within that structure.

Now, let's bring conditions 1 and 2 back in, returning to full-fledged institutions. Condition 1 has bearing on the nature of institutional roles, because copying is what individuates institutional roles across time. Specifically, a presently enacted role is the same role as the past role with which it shares the greatest number of causal links formed via copying.[12] This applies even if some behaviors within the role have changed (*cf.* Harré, 1979, p. 99). For example, consider a corporation in which several job duties shift from administrative staff to communications staff. If there is a tie with respect to this causal criterion between two (or more) distinct past roles, the current role counts as a merger of those past roles. This causal way of identifying the same role across time means that different institutions cannot have any of the same roles

10. This discussion about accessibility closely parallels considerations that apply to social conventions (see Stotts, 2023, p. 2126). And Millikan's (2005) notion of copying, to which I appealed above, is also part of her account of social conventions. One might then wonder whether conventions are a necessary part of social institutions. I do think that many conventions exist within social institutions, but I also think that completely non-conventional social institutions are possible, such as, again, an objectively best form of government.

11. This notion of institutions involving individuals inter-enabling each other to promote particular results was partially inspired by Miller's (2001, pp. 57, 181) notion of collective ends.

12. This causal way of individuating roles was partially inspired by Millikan's (2005, pp. 33–34, 60–61) causal individuation of words.

within them. Roles here are institution-specific.[13]

The causal individuation of institutional roles also means that some behavior that does not promote the institution's result(s) can still be part of an institutional role, provided that it stands in the REB relation to other role-holders in the institution and that it was copied from the past behavior of occupants of that role. This allows roles within an institution to have some aspects that do not contribute to the institution's result(s), perhaps due to poor planning. For instance, imagine a certain worker at a corporation is required to write an annual report, which contributes nothing to the institution's results. Writing this report can still be part of their role.

One final remark I'd like to make about roles at this stage is that the roles that actually exist within institutions may be decidedly less fine-grained than everyday talk suggests. Consider empty promotions within a corporation. A corporation may have, for instance, Analyst 1 and Analyst 2 positions on the books, where the job duties and compensation do not actually change when one is "promoted" from Analyst 1 to Analyst 2. The BTI implies that there really aren't two different roles in that scenario; just one, for which participants in the institution mistakenly (or perhaps, deceptively) have two different names.

The BTI specifies the conditions individually necessary and jointly sufficient for an institution to exist, but there are other features that many institutions have. First, there can be behavior by role-holders that is not part of their role, but which is still part of the institution. All behavior performed by role-holders that promotes the institution's results in a way partially causally enabled by the behavior of other role-holders is part of the institution, even if that behavior is not copied or even repeated. In other words, the behavior that *does* cluster into roles determines what the institution's results are, and then other behavior by role-holders that promotes the same results in a way partially enabled by others in the institution is also part of the institution. For instance, if a bank's CEO takes an unusual, one-time action promoting the bank's interests, that behavior may still be part of the institution, even though it is not part of the CEO role.

Many institutions also include what I will call *secondary roles*. The *primary roles* are the ones we have already discussed, defined by role-holders' behavior. In addition to this, entities (singular objects, relations, kinds of objects, and kinds of relations) can come to have secondary roles in an institution by being systematically impacted by the behavior that enacts the primary roles, without themselves engaging in behavior that would make them a holder of a primary role. For example, a particular building takes on the role of *county courthouse* for a particular county because of the court activities that occur within it. Similarly, a *kind* of object, such as balls manufactured at a particular factory, might have the role of *official game ball* within a particular athletic institution. For an example involving relations rather than objects, consider the secondary role of *foul* in an athletic institution, where sev-

13. Nonetheless, two roles in different institutions may be of the same social *kind*. For instance, there is a role called 'coach' within both Major League Baseball and the Women's National Basketball Association. These are two different roles, because they are in two different institutions. But we can still say that they are roles of the same *kind* because they include so many of the same kinds of behavior. I'd like to remain neutral with respect to whether these are robust metaphysical kinds, or whether it is just a matter of entities being "*socially named*" in Millikan's (2014, p. 28) sense (*cf.* Millikan, 2005, pp. 21–22).

    With respect to the question of roles belonging to multiple institutions, one might also wonder whether a role *qua* enacted by a particular individual could ever belong to multiple institutions simultaneously. Consider a delivery worker who regularly delivers textbooks to a particular university. Delivering those books contributes to the university's result of knowledge transmission, in a way that both enables and is enabled by the behavior of other role-holders in the university. Does the delivery worker have a role in the university, or only in the delivery company? Presumably, the delivery worker delivers books to other locations as well and copies their role from others who deliver books to multiple places. So, when we consider the delivery worker's entire role as it is individuated by its causal history, a small percentage of their behavior promotes the university's results, and a much higher percentage (likely nearly all of it) promotes the results of the book delivery company. This means that their role falls within the delivery company, though it is a role that involves contributing to the success of external institutions, including the university. In cases where someone's role genuinely contributes equally to the results of two different institutions, we truly would have a role that is simultaneously within two institutions. Some liaison roles may be like this.

eral different relations (such as grabbing someone's clothing, tripping them, or forcefully bumping them) are all treated as members of a kind. Institutions are not required to include secondary roles, so within the BTI itself, 'role' means just primary roles.

3. Metaphysical Details

With the BTI's main features clarified, I now want to say more about the metaphysical details. There are several important questions here: What is the precise metaphysical relation that generates social institutions? What is the metaphysical status of roles? How do institutions persist through time? Where are institutions located? And, how do institutional facts fit into the picture?

3.1 What is the precise metaphysical relation that generates social institutions?

My contention is that the relation that generates social institutions is *composition*. Flowers have petals and stamens as their parts; social institutions have individuals and behavior as their parts. More specifically, the primary role-holders, concrete behavior enacting the primary roles, and the primary role-holders' other behavior (if any) that promotes the institution's result(s) in a way causally enabled by other role-holders, are all parts of the institution. Secondary role-holders (if any) are better thought of as the institution's tools, beneficiaries, or victims, in my view.

Composition is the right metaphysical relation because it implies that when one encounters someone who is enacting an institutional role, one is bumping (perhaps quite literally) into a *part* of an institution. Seeing institutions as wholes with individuals and behavior as their parts underwrites the idea that, for instance, if a police officer physically places me inside a vehicle, the policing institution itself is placing me inside the vehicle. Institutions are with us in the physical world, and we regularly have interactions with them (that is, with one part of them or another) in which they exhibit physical, causal power. If we appealed to grounding instead of composition, the police officer and their behavior would be not parts of the institution but just an aspect of the institution's many partial grounds. When we encountered them, we would be encountering not a part of the institution, but just something that helps to make it the case that the institution exists. In other words, it would be much less clear that we're having physical interactions with the institution itself.[14] Composition is also, in my judgment, a better option than constitution. Constitution is a one-one relation, and behavior and individuals together do not seem to me to be wholes that exist prior to being made into institutions, which are then made into institutions by an additional step (Bennett, 2017, p. 9).

Allowing both individuals and concrete instances of behavior to be parts of social institutions allows us to avoid a potential puzzle. When it comes to composition, it is commonly held that "two distinct things cannot have all the same parts" (Bennett, 2017, p. 25; *cf.* Cotnoir, 2014, p. 5). But it certainly seems as though there might be a single set of individuals who are all of the primary role-holders in two different institutions, such as perhaps an amateur softball team and a school council, just by coincidence. How is this possible? Because the sports team has as its parts each of the individuals plus all of their various softball-team-related behavior, whereas the school council has as its parts those same individuals plus all of their various school-council-related behavior. So, the two institutions have some of the same parts, which poses no mystery, but they also have parts that differ.[15]

14. I do not mean to suggest that the fact that a particular officer puts me into a police car does *not* ground the fact that the policing institution has put me into the police car. What I am claiming is that these grounding relations between facts, though they do obtain, are not what metaphysically generate the institution.
15. Ruben (1983, pp. 232*ff.*) raises this kind of concern as an argument against seeing social entities of any kind as composed of individual humans, and the BTI is thus not vulnerable to his arguments, due to not portraying individual humans as institutions' *only* parts.

3.2 What is the metaphysical status of roles?

If an institution has concrete individuals and behavior as its parts, one is left wondering how roles fit into the picture, metaphysically speaking. Roles are not good candidates to be additional parts of institutions, because they are not concrete particulars: multiple individuals can hold the same role simultaneously. We'll focus first on primary roles.

My proposal is that primary roles are causally individuated behavioral profiles. Imagine that some institution exists, satisfying all aspects of the BTI. This entails that the concrete behavior within the institution clusters into roles, which means (as stated in Section 2) that there are kinds of behavior in which certain individuals engage, and from which certain others abstain, and *vice versa*. A behavioral profile is a set of kinds of behavior sorted into "engaged-in" and "abstained-from" categories, which can of course be enacted by multiple individuals simultaneously. Behavioral profiles are *causally individuated* in the following sense: a presently enacted profile is the same role as the past profile with which it shares the greatest number of causal links formed via copying (again, see Section 2).

Secondary roles, as we've already discussed, emerge when entities come to be systematically impacted by behavior that enacts primary roles within an institution, without themselves enacting a primary role in that institution. They, too, are causally individuated behavioral profiles, though passive rather than active ones. Specifically, they are sets of kinds of behavior sorted into "done to this entity" and "not done to this entity" categories. There may also be subtler categories identifying what holders of various specific primary roles do, and don't do, to the entity. Passive behavioral profiles are causally individuated, too: a presently enacted profile is the same role as the past profile with which it shares the greatest number of causal links formed via copying (where the copying was done by primary role-holders, of course).

3.3 How do social institutions persist through time?

Given that social institutions are these strange wholes with individuals and behavior as their parts, it is natural to wonder how they persist through time. My answer, building on some helpful comments from Millikan (2017, pp. 18–20), is that institutions persist through time in much the same way persons do. We can think of a person as having different stages: me in 2025, me in 2012, *etc.* These stages are stages of the same person because of certain causal connections among them. Similarly, what makes it the case that the U.S. National Science Foundation (NSF) in 2025 is the same institution as the NSF in 1962 is the presence of certain causal connections: the 2025 NSF role-holders are copying from past behavior that has its causal origins in the behavior of 1962 NSF role-holders.

For institutions more so than for persons, there can be some blurriness. Imagine that in 1997, some especially efficient administrative workers at another institution shared their methods with some of the NSF's administrative staff, such that some 2025 NSF role-holders are causally connected to 1997 role-holders in that other institution. Does that mean that both the 1997 NSF and that other institution in 1997 are somehow past stages of the 2025 NSF? No, it does not. The response here parallels the way we individuated roles: if there are multiple past stages to which a present institution stage is causally connected, the quantity of the causal connections determines which is the current institution's past stage. The 1997 NSF would have vastly more causal connections to the 2025 NSF than would that other institution in 1997. But if there were to be a tie on this metric, that would amount to fusion of multiple past institutions.

As a final comment: I do not require an institution to promote the same result(s) as its past stages, in order for them to be its past stages. It's important to allow that what an institution does can change over time, while it remains the same institution. This means an institution can potentially change so much as to be unrecognizable, but of course persons can do that, too.

3.4 Where are institutions located?

Hindriks (2013) argues that accounts of social institutions face a crucial question: where, according to the account, is any given institution located?[16] The BTI's emphasis on institutions' physicality makes this question especially pressing.

With composition as the BTI's underlying metaphysical relation, answering this question is relatively straightforward. A whole is located wherever its parts are located, so an institution is partially located wherever each of its primary role-holders (and their behavior) is located at any given time. For an institution whose primary role-holders never gather in person (such as, for instance, an international professional association that conducts all activities over the internet), the location will be highly dispersed. But there's nothing surprising about this, in my view. It seems quite natural to say that such an institution is located throughout the world, rather than trying to prioritize certain role-holders' locations as the institution's "real" location.

One advantage of treating not just primary role-holders but also their behavior as parts of the institution is that we can distinguish between locations where an institution is merely present and those where it is robustly, behaviorally present. So, for instance, if all of the primary role-holders in a certain institution happen, by coincidence, to go on vacation to the same destination, it would be true that the institution is present at that destination at that time.[17] But there would be key parts of the institution that we'd be able to note had never been present in that location: no (or perhaps trivially few) behavioral parts were present. The role-holders were there, but behavior enacting their roles was not, so the institution was present there only in an attenuated manner.

16. Hindriks (2013, p. 428) actually frames the question as being about organizations, rather than institutions, but he makes clear that he sees organizations as a species of institutions, and the question generalizes.
17. This scenario is adapted from Ruben (1983, p. 225; *cf.* Hindriks, 2013, p. 415).

3.5 How do institutional facts fit into the picture?

The BTI does not mention institutional *facts*. Some accounts of social institutions treat institutional facts as primary with respect to institutions, with the implication that institutions are then built up out of institutional facts.[18] I do not see institutional facts as primary with respect to institutions, but institutional facts are nonetheless part of my picture. The main kinds of institutional facts are facts about the existence of institutions (*e.g.*, *Yamaha Corporation exists*), and facts about the occupants of roles within them (*e.g.*, *Susan is a Catholic nun*, or *This building is our courthouse*). These kinds of institutional facts obtain within all institutions. There may be roles, and social facts about who or what occupies them, prior to the existence of an institution. But only once there is an institution are the roles institutional roles, and only then are facts about them institutional facts.

4. Applying the Theory

The next step is to confirm that the BTI deals adequately with Racist Courts, Disputed Government, and Watchdog. In Racist Courts, behavior clusters into roles, such as *attorney* and *judge*. Occupants of those roles copy each other, and the roles have been vacated and passed on. All of the roles together inter-enable each other to promote certain results, including over-incarceration of a certain racialized group, that could have been promoted in other ways. The injustice against the over-incarcerated group is one of the institution's results—a defining aspect of its nature—regardless of anyone's intentions or beliefs. So far, so good. In Disputed Government, there is behavior clustering into roles that have been copied and passed on, leading to inter-enabled promotion of certain results that could have been promoted in other ways, and on the basis of that behavior, Candidate A has the *leader* role and Candidate B does not. The government has one leader, despite all of the disagreement. And in Watchdog, there is no watchdog institution because the requisite behavior is missing, regardless of people's

18. *E.g.*, Epstein (2014) and Searle (2009).

beliefs.

Let's also consider how the BTI applies to a straightforward, paradigmatic example of an institution: Harvard University. Harvard promotes the results of knowledge generation and transmission. At Harvard, there is the role of *student*, which involves attending classes and submitting assignments, among other things. Occupants of the *student* role abstain from other behavior that promotes knowledge transmission, such as creating and grading assignments. The *instructor* role, on the other hand, involves engaging in some of the result-promoting behavior from which students abstain, and abstaining from some of the result-promoting behavior in which students engage. There are many other roles, such as administrative roles, alumni roles, the *president* role, various food service staff roles, and various facilities staff roles. Importantly (for condition 1), these roles are all copied: students are influenced by other current or past students, instructors influence each other, and so on. And, as condition 2 requires, these roles have certainly been vacated and passed on in Harvard's history. Finally, in line with condition 3, there are equally accessible alternatives (such as an entirely different leadership structure for the university), and the roles promote Harvard's results in an inter-enabling way. That is, the actions of faculty, students, and staff enhance each other's ability to promote knowledge generation and transmission.[19]

19. One might wonder how the BTI would apply to a university that separates its teaching and research missions. Consider a university with a single overall leadership structure and a shared staff of workers in facilities, food service, and so on, but with two separate wings: a teaching wing, and a research wing. The teaching wing pursues the result of knowledge transmission, while the research wing pursues the result of knowledge generation. Neither of these results is such that all roles in the overall institution promote it. What this means is that the result that the university as a whole promotes will be somewhat vague: just knowledge, because that is what all roles do in fact promote, together.

5. Comparison to Other Views

I'll discuss two main ways in which the BTI differs from other accounts of social institutions: it eschews mental phenomena, and it precludes inconsistency.

First, the avoidance of mental phenomena. It should be plain at this point how deeply the BTI differs from collective acceptance accounts such as Searle's (1995, 2009), Tuomela's (2002, 2007, 2013), Thomasson's (2003), and Ludwig's (2017).[20] It differs just as starkly from accounts that appeal to mental states other than collective acceptance, such as Miller's (2001, 2003), Guala's (2016), Torrengo's (2017), and Hindriks's (2023).[21] It is worth mentioning that Guala (2014, p. 65) and Miller (2003, pp. 241–242) agree with my emphasis on what actually happens in institutions, as opposed to what we collectively accept as true about our institutions. But they still go on to offer accounts

20. One similarity between Ludwig's (2017) account and the BTI is an emphasis on roles. But Ludwig thinks that roles are imposed by collective acceptance, so roles for him are quite different from the BTI's notion of reciprocal exchanges of behavior (pp. 5–6). Tuomela also thinks that social institutions involve positions or roles, but he sees roles (other than just the role of membership in the group that creates an institution) as having do with some members of a group being authorized to act on behalf of the others (2013, pp. 249, 278; 2002, p. 199; *cf.* 2013, p. 292). Harré (1979, p. 98), too, argues for a role-based account: for him, an institution is "an interlocking double-structure of persons-as-role-holders or office-bearers and the like, and of social practices involving both expressive and practical aims and outcomes." But Harré's account also brings in mental states, in at least three ways: as part of how he understands social practices (pp. 54, 98), as what keeps an institution in existence when it is "latent" (pp. 38, 99–100), and as how he demarcates the "bounds" of institutions (p. 99).

21. It's worth noting that Miller's (2003, p. 234) view of organizations (not institutions) bears interesting similarities to the BTI. He sees an organization as "an (embodied) formal structure of interlocking roles," where "organisations are individuated by . . . their characteristic functions or ends." Organizations that do have the right "normative dimension" are also institutions, for him (p. 235). The BTI differs from Miller's account in several important ways: by seeing interlocking roles as necessary for all institutions, by eschewing the kinds of mental states Miller builds into his account, and by focusing on the results the behavior within an institution actually promotes rather than the ends participants take themselves to be pursuing (on this last point, see especially Miller (2003, p. 245)).

of social institutions that, unlike the BTI, appeal to mental phenomena. In eschewing mental states, the BTI also differs importantly from Epstein's (2014) pluralist approach to social institutions. According to Epstein's approach, institutional facts can obtain in ways not metaphysically dependent on mental phenomena, but mental phenomena of various kinds *can* give rise to institutional facts.

By eschewing mental states in the BTI, I am *not* eschewing mental states in the broader explanatory story of the behavior that gives rise to social institutions. Intentions (conscious or otherwise) are often present, guiding the behavior of primary role-holders in institutions. My point is just that these intentions and other mental states do not metaphysically determine the institution's existence and nature.[22] They are part of the causal story of why the institution exists, and not part of the metaphysical story of what generates the institution. This means that, in principle, non-conscious machines could have a social institution, provided that they were sophisticated enough to satisfy the BTI (including having the ability to copy each other). But I also want to acknowledge that for the purposes of certain conversations within political philosophy or the social sciences, we might be concerned only with the kinds of institutions that do have intentions and other mental states underlying the behavior (a criterion that, I assume, most actual institutions satisfy). The BTI is not intended to undermine these conversations, but just to show that when it comes to the project of doing the metaphysics of social institutions—that is, the project of understanding what makes it the case that they exist—those mental phenomena are not relevant.

The second feature of the BTI that I'd like to highlight is its preclusion of inconsistency in institutional reality. Brouwer (2022, pp. 24, 30) argues persuasively that Epstein's approach to social ontology, as well as Searle's and Tuomela's, implies the possibility of inconsistent institutional facts—that is, of situations in which some fact *A* and another fact not-*A* obtain simultaneously, within a single institution. I won't delve into the details of how inconsistency can arise for other theorists, but one key feature that allows it to arise for each of them is that they hold that institutional facts can be generated one by one (p. 38).

On the other hand, according to the BTI, institutional facts about which individuals hold which roles are determined by all of the behavior within an institution across a given span of time, taken together. That behavior as a whole will either cluster into roles that work together to promote some result(s), or it will not. And any given individual will either satisfy the criteria for being a primary role-holder within a given institution, or they will not. That is: they will either behave in ways that link into the web of behavior by role-holders in the institution that is interconnected by REB relations and inter-enabled promotion of the institution's result(s), with the right history of copying, or they will not. It is not possible for someone to simultaneously satisfy, and not satisfy, these requirements with respect to the same role in a given institution. Similarly, a given entity that does not hold a primary role in a given institution either will or will not be systematically impacted by the behavior of primary role-holders, and thus either will or will not have a secondary role.

Now, there may be genuinely murky cases, such as when an entity is just beginning to manifest a passive behavioral profile and take on a secondary role. These may be cases of indeterminacy, but they certainly aren't cases in which someone or something both has, and does not

22. There are two assumptions one might make that might lead to uncertainty about whether this claim is compatible with the BTI. If we assume that composition is transitive (Assumption 1), then the fact that primary role-holders are parts of institutions would imply that their brains (which are parts of the role-holders) are parts of those institutions (Bennett, 2017, p. 46). And then if we assume, more controversially, that mental states are identical to brain states (Assumption 2), it might seem that primary role-holders' mental states *are* in fact parts of institutions. However, I don't think this follows. Although each primary role-holder's brain is a part of them and would thus be a part of the institution under Assumption 1, a particular *state* that obtains in a brain at a given time is not a *part* of that brain. A brain state is a state of affairs within a brain, not a part of the brain. So although brains (and their parts) would be parts of institutions under Assumptions 1 and 2, mental states would still not be parts of institutions.

have, a given role.[23]

One further question that may arise about how the BTI compares to other accounts of institutions is the following: what kind of normativity, if any, do institutions have, according to the BTI? It is common for authors writing on social institutions to build some kind of inherent normativity into their account (Miller, 2003, p. 243; Searle, 2009, p. 91; Tuomela, 2013, pp. 227, 43; Guala, 2016, Ch. 6). The BTI does not explicitly build in normativity, but the question of whether subtler normativity may nonetheless be present is a complex one that I will leave for another occasion. Whether institutions have inherent normativity or not, I want to emphasize that we can (and should) normatively evaluate them. Just as a powerful machine someone creates can be normatively evaluated despite not itself being inherently normative, so could normatively inert institutions be evaluated with respect to independent moral considerations.

6. Conclusion

We began with the idea that social institutions' existence and nature are "up to us" in a deep, important way. Institutions really are "up to us" in that their existence and nature are determined by and dependent on our behavior, and also in that each of us can control our behavior. But still, there are two important ways in which institutions are not "up to us." First, their existence is not determined by or dependent on what we believe, decide, or agree upon. Second, their nature is determined by the results that our behavior actually promotes and not by results at which we intentionally aim. Social institutions, though deeply dependent on our behavior, can slip out of our grip. The right account of social institutions must, as I hope the BTI does, reckon fully with this reality.

23. This way of avoiding inconsistency is similar to a solution Brouwer (2022, pp. 38*ff*.) considers on Epstein's behalf. The concerns Brouwer raises about that potential solution do not apply to the version just suggested because the BTI *begins* by not allowing single institutional facts to be generated individually, rather than allowing that kind of individual generation and then later attempting to impose a requirement of consistency.

Acknowledgements

For helpful feedback and discussion, I am grateful to Graham Hubbs, Stefan Sciaraffa, Julian Sheldon, and audiences at McMaster University, the Social Ontology 2022/Collective Intentionality XIII conference, the 2023 meeting of the Pacific Division of the American Philosophical Association, MillikanFest, Toronto Metropolitan University, the 2024 meeting of the Western Canadian Philosophical Association, and the 2025 meeting of the Canadian Philosophical Association. For research assistance, I am grateful to Siddharth Raman and Gary Spero. I am also grateful to the students in my Fall 2018, Fall 2022, and Fall 2023 graduate seminars at McMaster University for helpful discussion on topics related to this article, and to several anonymous referees.

Funding information

This article draws on research supported by the Social Sciences and Humanities Research Council of Canada (Grant Number: 430-2019-00092).

References

Bennett, Karen (2017). *Making Things Up*. Oxford University Press.

Brouwer, Thomas N. P. A. (2022). Social inconsistency. *Ergo: An Open Access Journal of Philosophy*, 9(2), 18–46. DOI: 10.3998/ergo.2258

Cotnoir, Aaron J. (2014). Composition as identity: Framing the debate. In A. J. Cotnoir & D. L. M. Baxter (Eds.), *Composition as Identity* (3–23). Oxford University Press.

Epstein, Brian (2014). Social objects without intentions. In Anita Konzelmann Ziv & Hans Bernhard Schmid (Eds.), *Institutions, Emotions, and Group Agents: Contributions to Social Ontology* (53–68). Springer.

Epstein, Brian (2015). *The Ant Trap: Rebuilding the Foundations of the Social Sciences*. Oxford University Press.

Gordon, Deborah M. (2010). *Ant Encounters: Interaction Networks and Colony Behavior*. Princeton University Press.

Guala, Francesco (2014). On the nature of social kinds. In Mattia Gallotti & John Michael (Eds.), *Perspectives on Social Ontology and Social Cognition* (57–68). Springer.

Guala, Francesco (2016). *Understanding Institutions: The Science and Philosophy of Living Together*. Princeton University Press.

Gullan, P. J., & P. S. Cranston (2014). *The Insects: An Outline of Entomology* (5th ed.) John Wiley & Sons.

Harré, Rom (1979). *Social Being: A Theory for Social Psychology*. Basil Blackwell.

Hindriks, Frank (2013). The location problem in social ontology. *Synthese*, 190(3), 413–437. DOI: 10.1007/s11229-011-0036-0

Hindriks, Frank (2023). Rules, equilibria, and virtual control: how to explain persistence, resilience and fragility. *Erkenntnis*, 88, 1367–1389. DOI: 10.1007/s10670-021-00406-9

Ludwig, Kirk (2017). *From Plural to Institutional Agency: Collective Action II*. Oxford University Press.

Miller, Seumas (2001). *Social Action: A Teleological Account*. Cambridge University Press.

Miller, Seumas (2003). Social institutions. In Matti Sintonen, Petri Ylikoski, & Kaarlo Miller (Eds.), *Realism in Action: Essays in the Philosophy of the Social Sciences* (233–250). Springer.

Millikan, Ruth Garrett (2005). *Language: A Biological Model*. Oxford University Press.

Millikan, Ruth Garrett (2014). Deflating socially constructed objects: What thoughts do to the world. In Mattia Gallotti & John Michael (Eds.), *Perspectives on Social Ontology and Social Cognition* (27–39). Springer. DOI: 10.1007/978-94-017-9147-2_3

Millikan, Ruth Garrett (2017). *Beyond Concepts: Unicepts, Language, and Natural Information*. Oxford University Press.

Ruben, David-Hillel (1983). Social wholes and parts. *Mind*, 92(366), 219–238. DOI: 10.1093/mind/XCII.366.219

Searle, John R. (1995). *The Construction of Social Reality*. Free Press.

Searle, John R. (2009). *Making the Social World: The Structure of Human Civilization*. Oxford University Press.

Stotts, Megan Henricks (2017). Walking the tightrope: Unrecognized conventions and arbitrariness. *Inquiry*, 60(8), 867–887. DOI: 10.1080/0020174X.2017.1285995

Stotts, Megan Henricks (2023). Conventions without knowledge of conformity. *Philosophical Studies*, 180, 2105–2127. DOI: 10.1007/s11098-023-01957-z

Stotts, Megan Henricks (2024). Moving from the mental to the behavioral in the metaphysics of social institutions. *Synthese*, 203, Article 123. DOI: 10.1007/s11229-024-04532-z

Thomasson, Amie L. (2003). Foundations for a social ontology. *ProtoSociology*, 18(19), 269–290.

Torrengo, Giuliano (2017). Institutional externalism. *Philosophy of the Social Sciences*, 47(1), 67–85. DOI: 10.1177/0048393116670010

Tuomela, Raimo (2002). *The Philosophy of Social Practices: A Collective Acceptance View*. Cambridge University Press.

Tuomela, Raimo (2007). *The Philosophy of Sociality: The Shared Point of View*. Oxford University Press.

Tuomela, Raimo (2013). *Social Ontology: Collective Intentionality and Group Agents*. Oxford University Press.

Turner, Stephen (1994). *The Social Theory of Practices: Tradition, Tacit Knowledge, and Presuppositions*. Polity Press.

Wilson, Robert A. (2007). Social reality and institutional facts: Sociality within and without intentionality. In Savas L. Tsohatzidis (Ed.), *Intentional acts and institutional facts: Essays on John Searle's social ontology* (139–153). Springer.