Moving from the Mental to the Behavioral in the Metaphysics of Social Institutions

Megan Henricks Stotts

This version of the article has been accepted for publication, after peer review and is subject to Springer Nature's AM terms of use, but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1007/s11229-024-04532-z.

1. Introduction

Social institutions—such as the government of Canada, the National Football League in the United States, the Japanese monetary system, and the Catholic Church—often seem as real to us as mountains, oceans, and forests. And yet, social institutions seem to be real in an importantly different, more human-dependent way. This distinctive feature of institutional reality motivates the key question behind the metaphysics of social institutions: what, precisely, makes it the case that social institutions exist? Or in other words: what are the metaphysical determinants of institutional reality? Here we are asking not the empirical question of what historical events *caused* particular institutions to exist, but rather the metaphysical question of the kinds of states of affairs *in virtue of which* institutions exist.[1]

One especially influential strand of the contemporary philosophical literature on the metaphysics of social institutions has been the collective acceptance approach, most prominently advocated by John Searle (1995, 2009) and Raimo Tuomela (2002, 2007, 2013). On this sort of view, people's collective acceptance is necessary for the existence of social institutions. Accounts that depart from the collective acceptance approach tend to either preserve some limited role for collective acceptance, or to replace collective acceptance with some other kind of mental state. I will argue that this emphasis on the mental in the metaphysics of social institutions is a mistake. And although my conclusions will be primarily negative, I'll also suggest an alternative approach according to which social institutions and the institutional facts associated with them arise just from our behavior, with no role for mental states.

[1] In keeping with this way of framing the question, I will use the term 'metaphysical determination' only for non-causal metaphysical relations.

Before diving in, some clarification of the paper's scope is in order. In my usage, 'institution' does not apply to a single convention, norm, or rule. It applies only to complex, structured entities such as governments, athletic institutions, and religious organizations.[2] To understand the kind of complexity that I have in mind here, it may be helpful to consider the wide variety of institutional facts associated with these social institutions. For instance, associated with the Catholic Church, there is the institutional fact that a particular man is the pope, that a particular woman is a nun, that another individual is a member, and so on, for a vast number of similar facts. Associated with the Japanese monetary system, we have the fact that some particular piece of metal is a 5-yen coin, the fact that some individual (at a particular time) is a buyer, the fact that another is a seller, and so on. These complex structures that exert so much influence on our lives are quite different in nature from single conventions, norms, or rules, and as a result, their metaphysics should be done separately, in my view.

I also want to acknowledge at the outset that I see institutional reality as a subset of social reality, where the latter also includes single conventions and social norms, but also social facts such as the fact that two people are friends, or that a particular restaurant is popular. I'll ultimately have more to say below about the distinction between institutional reality and social reality more broadly, but in the early sections, we'll discuss only institutional reality.

I'll begin, in Section 2, with a series of troubling counterexamples to collective acceptance views that have accumulated in the literature, arguing that together they make a strong case against the collective acceptance approach. The next two sections will each criticize a different way of responding to the problems that trouble collective acceptance accounts. In Section 3, we'll discuss a response that utilizes Brian Epstein's (2014) pluralism about the metaphysics of institutional facts, according to which collective acceptance still plays a role in the metaphysics of *some* institutional facts.

[2] In this narrower usage of 'institution,' I follow Miller (2001), Searle (2009), and Epstein (2014), in contrast to the broader usage of 'institution' favored by, for instance, Guala and Hindriks (2015) and Schotter (1981).

In Section 4, we'll discuss responses that replace collective acceptance with individual mental states. The problems the responses considered in Sections 3 and 4 face suggest that the way forward is to avoid any appeal to mental phenomena in our account of social institutions, and I'll end (in Section 5) with an overview of how I think that can be done: by seeing institutional reality as arising from behavior that clusters into roles and works together to bring about some result.

2. Collective acceptance accounts and their problems

We'll begin with a discussion of the collective acceptance approach to social institutions and the serious problems it faces. Broadly speaking, collective acceptance is acceptance of some proposition or the existence of some fact or entity within a group, which the members of the group take themselves to do *as* group members, and not just as individuals. If every member of my campus community thinks of some particular tree as the most beautiful tree on campus, and we are all aware that this is a widely held opinion, that is not an instance of collective acceptance because it is an attitude each of us holds only as an individual (*i.e.*, as an 'I') rather than as group members (*i.e.*, as part of a 'we'). On the other hand, if everyone on campus thinks of some particular tree as the Graduation Tree that every graduating senior at our university should touch right after graduation, and we think of ourselves as holding that attitude collectively, as a 'we,' then *that* would be an instance of collective acceptance.

We'll need to go beyond this broad gloss and into the specifics of Searle's and Tuomela's collective acceptance accounts of social institutions to be able to see exactly how the counterexamples we'll discuss tell against them.[3] For Searle, collective acceptance as we've just described it is one of two kinds of collective intentionality that play key roles in institutional reality. Searle (2009, p. 56)

[3] Kirk Ludwig (2017, p. 132) argues for another collective acceptance approach to institutions, but his definition of collective acceptance (in terms of intentions or conditional intentions) is so different from Searle's and Tuomela's that his view is not subject to the criticisms of this section. However, the criticisms in Section 4 do ultimately apply to Ludwig's view (see footnote 19, below).

refers to the phenomenon we're calling 'collective acceptance' as 'full-blown cooperation.' His notion of full-blown cooperation does have the key features of collective acceptance as we've just described it. For instance, Searle argues that full-blown cooperation is irreducible to any kind of individual attitude, in the sense that each individual has in their own mind an attitude of plural form (such as, in Searle's example, "We are cleaning the yard"), which cannot be reduced to any attitude(s) of the 'I' form (pp. 47*ff.*).

For Searle, the key mechanism behind social institutions occurs when people, by means of full-blown cooperation, impose a new status on some object. This involves cooperatively producing a speech act of Declaration (or mental states that have the form of a Declaration) that imposes the status (Searle, 2009, pp. 93, 96). For example, Searle discusses a group of villagers who once had a wall around their village, but the wall has deteriorated so much that it can no longer physically restrict access to the village (p. 94). Once that has happened, the villagers may collectively impose the status of being a boundary on the remnants of the wall, with a Declaration (in speech or thought) of the following form: "We make it the case by Declaration that this row of rubble now has the status 'boundary' and thus is able to perform the function 'limiting rightful access to the village' in the context of the village and its surrounding region" (p. 99).[4]

Although the simplest of these Declarations impose a new status on a single, pre-existing object (such as the remnants of a wall), Searle (2009, pp. 96–97) also discusses cases in which what is collectively declared is that *all* objects meeting some condition will receive a certain status going forward. These Declarations give rise to constitutive rules, such as the rule that any woman who satisfies certain conditions counts as a nun in the context of the Catholic church. As Searle sees it,

[4] I created this sample Declaration by inserting values into the schema Searle (2009, p. 99) provides, but it is worth noting that doing this is not entirely straightforward. Searle himself does not provide examples of filled in versions of his schemata for Declarations, and in particular he does not address the question of what kinds of things are of the right sort to serve as *contexts* for these Declarations. There are also difficulties in knowing how Searle intends us to prize apart the status that has been imposed and the resulting function that the object can perform.

social institutions are "system[s] of constitutive rules" instituted by cooperative Declarations (p. 10). So, for instance, the Catholic church is a system of constitutive rules such as the rule for who counts as a nun, who counts as a priest, which buildings count as churches, who counts as a member, and so on.

Searle (2009, p. 56) contrasts full-blown cooperation with collective recognition, which is "a much weaker form of collective attitudes."[5] Unlike full-blown cooperation, collective recognition *is* reducible to attitudes of the 'I' form, along with mutual belief. Although full-blown cooperation is needed to put in place the constitutive rules that make up an institution and to perform actions within the institution, collective recognition plays a key role in *sustaining* an institution and the statuses imposed within it over time (pp. 57–58).

Tuomela (2013, p. 128) treats collective acceptance as a matter of the formation of a joint attitude within some group of people. For him, collective acceptance can sometimes be in the "I-mode" rather than the "we-mode," where the I-mode is a matter of functioning "as a private person" and the we-mode is a matter of functioning as a member of a group in a strong sense where "members are to be conceptually understood as representatives acting for the group" (pp. 129, 2, 24). I-mode collective acceptance diverges from the broad gloss of collective acceptance I offered above, but Tuomela (2007, p. 198) makes it clear that we-mode collective acceptance is the kind that is necessary for social institutions.

For Tuomela (2013, p. 227), what must be collectively accepted in order to give rise to a social institution is some proposition that "expresses" that institution and "entails the existence of a … social practice … and a norm …, such that the social practice is governed by the norm." His example of a

[5] There is an unfortunate clash between Searle's terminology and mine: Searle (2009, p. 57) sometimes uses the term 'collective acceptance' for collective recognition, rather than for full-blown cooperation. I have chosen to use the term 'collective acceptance' in the main text to refer to what Searle calls 'full-blown cooperation' because I think it is apt, and also because it facilitates discussion of what Searle's and Tuomela's accounts have in common.

proposition that was collectively accepted to give rise to a social institution is "Squirrel pelts are money," where the collective acceptance of that proposition, along with the existence of the relevant norm-governed practice, once made it the case that squirrel pelts were money in Finland (p. 222).

Tuomela takes pains to emphasize that collective acceptance need not be enthusiastic. Sometimes collective acceptance takes the form of "going along acceptance," where people come to collectively accept some attitude while feeling, perhaps, coerced into doing so (Tuomela, 2013, p. 137). It may be tempting to look for analogies between going along acceptance and Searle's lighter notion of collective recognition, but the temptation should be resisted. For Tuomela, going along acceptance can still be in the we-mode, although we-mode phenomena based in mere going along acceptance are not the paradigmatic cases (p. 31). So, unlike collective recognition, going along acceptance can still be acceptance *qua* group member, in the we-mode, but it is reluctant or even coerced acceptance.

Now that we have a sense of what collective acceptance is, we're ready to consider a range of counterexamples that have emerged in the literature, which suggest that collective acceptance is not necessary for the existence of social institutions or the institutional facts associated with them.

First, we'll consider an example of a social institution that develops gradually, in the absence of collective acceptance. Alex Viskovatoff (2003), relying on an earlier piece by Stephen Turner (1999), discusses the example of non-coinage monetary systems. A monetary system can emerge gradually, as the exchange of certain objects for goods and services gradually becomes widespread in some group (Viskovatoff, 2003, p. 28; Turner, 1999, p. 225). For example, this may be what happened with respect to the use of shells as money in some human societies that have adopted that practice (*cf.* Turner, 1999, p. 224). In these kinds of cases, there is no need for collective acceptance in order for the institution to come into existence. Such a monetary system could function perfectly well without people collectively imposing a constitutive rule according to which shells of a certain kind have the status of money (as Searle would require), and also without them collectively accepting the proposition "Shells

(of a certain type) are money" (as Tuomela would want). Instead, as Viskovatoff (2003, p. 28) puts it, each individual just has to think, "'I will be able to use this for paying someone else, because members of this community exhibit the behavioral regularity of accepting such objects as payment.'" Thus, collective acceptance does not seem to be necessary for non-coinage monetary systems.[6]

Another kind of counterexample can occur when institutional facts are thought to be natural, not institutional, by the people involved. Edouard Machery (2014, p. 93) makes this point, relying primarily on the example of race as something that was long taken to be entirely natural by the vast majority of people, but which is now widely thought to be socially constructed.[7]

Because facts about race are perhaps best thought of as just social rather than institutional, we'll focus on a different, undeniably institutional example. Imagine a group of people who falsely believe that their sovereign's political power is a natural property—perhaps political authority is believed to be passed genetically from each leader to their offspring (Mäkelä and Ylikoski, 2003, pp. 265–266). Each citizen recognizes that their leader has the authority to govern them, but they recognize this only as individuals, and not as members of a 'we.' Each individual recognizes this fact in the same way in which she recognizes that water is H_2O, or that spring precedes summer (Machery, 2014, p. 98). It wouldn't make sense to these people to collectively accept a constitutive rule or proposition expressing the institution, when it seems to be just an obvious natural fact. Nonetheless, their government is a social institution.

For a third kind of example, we'll consider an adapted version of a case offered by Ignacio Sánchez-Cuenca (2007, pp. 180–181), in which political power is seized by force. Imagine a group in which one person amasses a large amount of resources and creates weapons more powerful than what

[6] Searle (1995, p. 126) discusses this kind of example in relation to an earlier version of his view, saying that "[m]oney gradually evolves in ways that we are not aware of" (*cf.* Tuomela, 2002, pp. 108–111). But he makes it clear that he thinks this evolution still involves implicit or unconscious collective intentionality (Searle, 1995, pp. 47, 126), whereas Viskovatoff's discussion of the case shows that collective intentionality is not needed in any form, conscious or otherwise.
[7] For a contrasting contemporary view of race, see Quayshawn Spencer's (2019) biological view.

the others possess. This person then declares herself to be the leader, demanding that everyone else obey her and surrender all resources to her. No one in the community believes she is their leader; they think of her as just a bully and a thief. But, if everyone in the society does comply with her edicts, it seems as though she has successfully, though unjustly, become their leader.

In this example, it is clear that the group members do not collectively assign a status to their leader, as Searle would require, but it might seem that Tuomela's notion of going along acceptance allows his account to capture this case. It does seem like the people in the example are just "going along" with the leader's demands, and their doing so is what makes her their leader (*cf.* Sánchez-Cuenca, 2007, p. 181). However, as described, the case involves going along with the leader's demands only in a colloquial sense; it does not involve going along acceptance in the we-mode, which is what Tuomela's account of social institutions requires. In this particular case, group members have not been coerced into forming an attitude of acceptance toward the proposition "So-and-so is our leader." Rather, they reject that proposition, but out of fear for their safety they each, *qua* individual, choose to do what the leader asks them to do. Thus, this really is an example of an institutional fact obtaining in the absence of collective acceptance, even in Tuomela's lighter, "going along" sense. The notion of going along acceptance allows Tuomela to capture *some* cases of coercion (as he himself notes (2013, p. 137)), but it does not allow him to capture all cases in which a social institution arises due to coercion.

One might wonder whether some or all of these counterexamples simply fall outside of the scope of Searle's or Tuomela's account. But Searle (2009, p. 91) makes it clear that he intends his account to apply to money and to political institutions, so the counterexamples seem fair. And Tuomela (2013, pp. 224, 222) states that his account is supposed to apply to "standard institutions such as money, property, and marriage," and also mentions "leadership … institutions," making government a natural addition to the list.

So, collective acceptance accounts run into trouble with at least three types of cases: social institutions that develop gradually and with minimal awareness, institutional facts that are mistakenly believed to be natural facts, and certain cases in which force is exerted to give rise to an institutional fact. This makes for a compelling argument against these collective acceptance accounts, in my view.

The lesson drawn by the authors who raised these crucial examples has been that we should identify a different role for mental states in the metaphysics of institutional reality, shifting our focus away from collective acceptance and toward individual mental states, such as intentions and beliefs. After all, these sorts of individual mental states are surely present in the counterexamples we've discussed. Another response to these kinds of counterexamples would be to turn toward pluralism about the metaphysics of social institutions and institutional facts, on the model of Epstein (2014).[8] Under pluralism, we could allow that in some cases, such as modern monetary institutions, collective acceptance *does* play a role in the metaphysics of institutional reality, but we could also accommodate cases in which it does not. Both kinds of responses still give mental phenomena a key role in the metaphysics of social institutions. We'll discuss problems with both responses, beginning with pluralism.

3. Against the pluralist response

Thus far, we've seen that institutional reality can emerge in the absence of collective acceptance, which undercuts collective acceptance accounts of social institutions. As we've just discussed, one possible response to that realization is to follow Epstein (2014) in adopting a pluralist view according to which collective acceptance is involved in the metaphysics of *some* but not all institutional phenomena. In developing his pluralist view, Epstein is not explicitly concerned with trying to preserve a role for

[8] Tuomela (2002, p. 122) himself sometimes sounds a bit like a pluralist. But he makes it clear that he always requires collective acceptance for the kind of "standard" institutions that are the focus of this paper, so he does not endorse the kind of pluralism about all institutions that Epstein favors (Tuomela, 2002, p. 170; 2013, p. 227).

collective acceptance. To the contrary, his aim is to show how collective acceptance plays a much more limited role in institutional reality than others have assumed (pp. 53–54). But his pluralism can in fact serve as a strategy for preserving a key role for collective acceptance in response to the problems discussed in the previous section. We'll ultimately consider two possible versions of pluralism: one that gives collective acceptance a stronger role in the metaphysics of institutions, and another that gives it a weaker but still significant role. We'll see that both forms of pluralism face significant problems.

Before we can understand Epstein's (2014) pluralism, we need to get a grip on his distinction between the *grounds* and *anchors* of institutional facts. The grounds of an institutional fact are the facts that make it the case that the institutional fact obtains, in the sense of being the "metaphysical reason" why the institutional fact obtains (Epstein, 2015, pp. 69–70, 76). So, in Epstein's example, what grounds the fact that a particular piece of paper is a U.S. dollar is that it was produced by the U.S. Bureau of Engraving and Printing (p. 79). But there is a further metaphysical question we can ask: what makes it the case that having been produced by the Bureau of Engraving and Printing grounds being a dollar? This is where anchoring comes in. What anchors the principle (a "frame principle," in Epstein's terminology) that being produced by the Bureau grounds being a dollar is some further fact, such as, perhaps, collective acceptance of the proposition that paper printed by that Bureau is currency (pp. 77, 80).

The kind of pluralism that matters for our purposes, then, is pluralism about *anchors*, since that is where collective acceptance would enter in. Epstein (2014, p. 61) endorses pluralism about anchors when he writes: "I aim to cast doubt on … the claim that there is any single sort of fact that is required for anchoring the social world. Instead, the world of social entities is a diverse one, with a variety of types of facts figuring into determining that constitutive rules are in place for a community." He then offers a list of things that could serve as anchors for different institutional facts, including "[e]xplicit

collective agreement," collective acceptance, "[w]idespread common (but not collective) intentions," "[i]ntentions of one or a few individuals …, with practices spread by mere causal transmission," and just "[p]atterns or regularities in practice" (pp. 61–62). As Epstein notes, these possibilities move from more to less demanding in terms of the presence and complexity of mental states (p. 61).

Pluralism about anchors seems like a promising strategy for responding to the counterexamples discussed in Section 2. A pluralist can say, for instance, that the fact that Justin Trudeau is the Prime Minister of Canada is anchored in collective acceptance: Canadians do accept, *qua* group members, the procedure that made him their Prime Minister. But then in the cases we discussed in Section 2 in which collective acceptance was absent, institutions and institutional facts could be anchored just in "widespread common (but not collective)" mental attitudes—for instance, in people's individual expectations when they use their group's form of currency. Thus, pluralism can smoothly accommodate the cases from Section 2, while preserving a key role for collective acceptance in the metaphysics of other institutional facts.

To see why the turn to pluralism is ultimately not a viable strategy, we first need to distinguish two different forms a pluralist view can take, both of which I take to be equally compatible with what Epstein says. The first form of pluralism would claim that, metaphorically speaking, collective acceptance always has metaphysical potency when it comes to anchoring institutional facts, but it is not the only thing that has this kind of metaphysical potency. Speaking literally, this would mean that whenever people collectively accept a frame principle for institutional facts, they succeed in anchoring that frame principle. But, according to this form of pluralism, there can also be instances in which frame principles are successfully anchored where collective acceptance is absent, or where collective acceptance plays only a partial anchoring role alongside some other item(s) from Epstein's list. This form of pluralism, which we'll call *strong pluralism*, attributes a high degree of strength to collective

acceptance: collective acceptance, on this view, always successfully anchors the frame principles for institutional facts that it purports to anchor, when it is present.

The second form of pluralism, which we'll call *weak pluralism*, would instead say that collective acceptance does not always succeed in anchoring whatever frame principle for institutional facts is collectively accepted, perhaps because it is overridden by other factors, but nonetheless collective acceptance *sometimes* does play a complete or partial anchoring role. Here, collective acceptance has a weaker, more limited ability to anchor frame principles for institutional facts, which can be overridden by other factors.[9] As we'll see, both versions of pluralism face problems.

We'll start by raising a problem for strong pluralism, which, again, is the idea that collective acceptance, though it does not play a role in anchoring the frame principles for all institutional facts, does always succeed in anchoring (or partially anchoring) the frame principles it purports to anchor for institutional facts, when it is present. Strong pluralism is subject to counterexamples quite similar to those in Section 2. In this kind of counterexample, there is collective acceptance of a frame principle that ought to imply that a particular institutional fact obtains, and yet that institutional fact fails to obtain.

Consider the following variant of Sánchez-Cuenca's (2007) case, which we'll call *Claudine the Conqueror*, for ease of future reference. Imagine a community whose members collectively accept that a certain electoral procedure determines who their leader is, and they follow the dictates of an individual, Martha, who was selected by that procedure. The community is then forcibly conquered by another individual, Claudine, and Martha is driven out. The members of the community do not accept any procedures that confer the status of leader upon Claudine. They persist in collectively accepting only the procedure that imposed the status of leader upon Martha, having clandestine

[9] Epstein (2016, p. 167) makes it clear that he is open to "heterogeneous anchors" of the sort discussed in this and the preceding paragraph.

conversations expressing their hope that their leader will return and free them from Claudine's grasp. Nonetheless, due to the threat of violence, they comply with all of Claudine's demands. Claudine remains in control for decades, until Martha has died, and then Claudine passes on control of the community in a manner of her own choosing. Yet, for as long as Martha was believed to be alive, the members of the community persisted in collectively accepting the procedure that designated Martha as their leader.

I contend that in this case, despite the fact that collective acceptance continued to designate Martha as the community's leader, Martha was not their leader, once Claudine decisively took control. From that time on, the community members lived their lives in the sway of Claudine's commands, in a way that remained politically stable over time. Claudine clearly never had a moral *right* to lead the community, but nonetheless, she did lead them. Here collective acceptance, though present, fails to anchor the frame principle it purports to anchor. The institutional fact that would have followed from the frame principle that collective acceptance would have anchored in this case, fails to obtain.

We can imagine a similar case involving money. Imagine that all U.S. citizens believe that every U.S. dollar corresponds to a certain quantity of gold, stored in Fort Knox (*cf.* Searle, 1995, p. 47). They collectively accept the principle that gold stored in Fort Knox is money, and they believe that the pieces of paper they exchange are just stand-ins for that actual money. But imagine that in fact, the U.S. dollar is fiat money, with no correspondence to any quantity of gold. The practice in which U.S. citizens are actually engaged is just a practice of exchanging paper bills for goods and services—that is, a practice of using paper bills as money, and not as a stand-in for gold. Despite their collective acceptance of the principle that gold stored in Fort Knox is money, whatever gold actually is stored in Fort Knox is *not* their money. It is just, let's suppose, a store of valuable metal that the U.S. government could trade with other nations in dire circumstances. Their money is in fact just the pieces

of paper. Thus, collective acceptance is unable to anchor the frame principles for institutional facts it purports to anchor, in at least some cases. Strong pluralism fails.

Next, we'll consider weak pluralism, which is the view that collective acceptance does not always succeed in anchoring whatever frame principle for institutional facts is collectively accepted, but nonetheless collective acceptance does play a complete or partial anchoring role with respect to *some* frame principles for institutional facts. One way to respond, for instance, to the counterexamples to strong pluralism would be to say that collective acceptance failed to anchor any frame principle because it was overruled by widespread, long-lasting individual mental states and/or regularities in practice. In Claudine the Conqueror, people continue to accept a frame principle that would imply that Martha was their leader, but they behave unambiguously as though Claudine is their leader, and they also have widespread mental states such as intentions to conform to Claudine's demands, and expectations that she will retaliate if they do not conform. So, the collective acceptance was outweighed. A similar point applies to the case involving gold and U.S. currency. These cases, it seems, leave the weak pluralist with an opening: the weak pluralist can say that although collective acceptance fails to anchor the frame principles it purports to anchor in these cases in which it is outweighed by other factors, it does still play an anchoring role in cases in which it is not outweighed.

Cases in which collective acceptance is not outweighed by anything else can take two forms: cases in which collective acceptance is the only factor present, and cases in which collective acceptance is in agreement with all (or most) other factors. We'll consider both kinds of cases, in turn.

First, we'll consider, or try to consider, cases in which collective acceptance is the *only* item present from Epstein's list. If we could find one such case in which an institutional fact is clearly created, that would prove that collective acceptance must play an anchoring role in at least that case. But, I am unable to think of any natural cases in which only collective acceptance is present. The only kinds of cases I can think of in which there is truly *just* collective acceptance are decidedly science-

fictional. For instance, imagine a group of people who use a well-functioning barter system to directly exchange goods and services. One night, a rogue scientist alters the brains of every group member while they sleep, such that when they wake, they collectively accept that peanuts are a form of currency among them. To keep our case a case of mere collective acceptance, we must imagine that the group members all continue their pre-existing barter practices and never actually use peanuts as currency, nor do they form individual intentions to use them, but nonetheless they collectively accept that peanuts are currency among them and *could* be used.

The problem with this case is that there just does not seem to be a social institution here according to which peanuts are currency; rather, what we have seems more like a situation in which such an institution could easily emerge, if people started acting as a result of this collective acceptance. Collective acceptance all on its own just doesn't seem to be enough. Now of course, my inability to come up with a plausible case in which collective acceptance, in isolation, anchors an institutional fact does not prove that there is no such case. But until we have such a case, we have no reason to think that collective acceptance can anchor a frame principle for institutional facts all on its own.

Next, let's consider potential cases in which collective acceptance is not outweighed because it is in agreement with a sufficient number of other factors (such as individual mental states, and regularities in practice). For instance, we can imagine a variant of Claudine the Conqueror in which Claudine gains powerful weapons and declares herself the leader, and then the others form individual intentions to comply, initiate actual practices of compliance, *and* collectively accept that having conquered them makes Claudine their leader. That sort of case is certainly possible, and I will grant that a substantial number of actual cases take this general form, with collective acceptance in agreement with a variety of other factors.

But when collective acceptance is in agreement with all (or most) other factors, the burden falls on the weak pluralist to make a case that collective acceptance is truly playing an anchoring role,

as opposed to being just an epiphenomenal feature of the case. This is due to pressure from some of the cases we have already discussed. The counterexamples to strong pluralism show us that factors other than collective acceptance can anchor institutional reality not just in the absence of collective acceptance, but in *opposition* to collective acceptance. So, if collective acceptance happens to be present and to agree with these other factors, we have no reason to think that it is part of the metaphysical reason why the institution is the way it is.

One could potentially argue for a hierarchical form of weak pluralism here: that regularities in practice and/or individual mental states are the strongest metaphysical determinants for social institutions, with collective acceptance having a lighter metaphysical "force," so to speak. But again, it's incumbent upon the weak pluralist to argue that collective acceptance does have *some* metaphysical force that is being outweighed in cases like our counterexamples to strong pluralism, as opposed to just having *no* metaphysical force. We've also just seen that it's hard to come up with cases in which collective acceptance *alone* anchors an institutional fact. That leaves us without the evidence of collective acceptance's metaphysical potency that such cases could have provided. The claim that collective acceptance does act as a metaphysical determinant when it is in agreement with other factors is the positive claim, and it is what needs defense. As things stand, for any particular case offered in which collective acceptance is in agreement with other factors, we have to wonder whether collective acceptance may be more like the music playing in an elevator: it is there, it is something people may remember as distinctive of their experience of riding the elevator, and yet it plays no role in getting them from one floor of the building to another. It seems, then, that even weak pluralism is foundering.

But here we can consider another strategy for the weak pluralist: identifying a specific, qualitatively distinct contribution that collective acceptance makes to the metaphysics of some institutional facts, when it is present. If there were such an identifiable difference between institutional

facts anchored by collective acceptance in agreement with other factors, versus institutional facts anchored by only those other factors, that would vindicate a kind of weak pluralism.

The most plausible candidate for such a difference, in my view, is a change to the modal features of institutional facts whose frame principles are anchored by collective acceptance. If people in some group just happen, without any attempt at cooperation, to defer in practice to whoever has the most weapons and resources, there will be a fair number of nearby possible worlds in which their practice doesn't happen to converge in this way. But if people collectively accept that whoever has the most weapons and resources is their leader, it may seem that there will be far fewer nearby worlds in which they don't treat the person who satisfies that condition as their leader. In other words, collective acceptance may increase the modal durability of institutional facts.

However, to me these modal features just show that collective acceptance makes it likelier that people would still create similar institutional facts if various small things about the world were otherwise, not that collective acceptance plays a role in making it the case that a particular frame principle is actually anchored, and thus that a particular institutional fact actually obtains. In my view, what we want to know when giving an account of institutional reality is what makes it the case that the institutional facts that actually obtain, actually do. Certainly for some purposes it will be of interest to know how likely we would have been to still anchor the same frame principles had the world been slightly otherwise than it is. But that is not the question we're interested in, when we're just doing the straightforward metaphysics of institutional reality. We want to know what kinds of factors make it the case that our institutional facts actually obtain, and factors that affect what institutional facts *would have* obtained if things were otherwise, just aren't relevant.

Another modal contribution that collective acceptance might seem to make is to generate purely hypothetical institutional facts. For instance, imagine that a group of people use gold as their currency, and they collectively accept that gold is their currency. But they also have a plan for what

they would do if all of their gold were ever to be stolen, and they collectively accept the following: "If all of our gold were stolen, we would use silver as our currency." Their gold is never stolen, so they never actually use silver as their currency. If we take a standard semantics for the counterfactual statement that the group members collectively accept, then to see whether it is true, we would need to consider all of the nearby possible worlds in which the group's gold is stolen. Are those worlds in which they then use silver as their currency?[10] Presumably, many of them will be. Assuming that they remember their previous, collectively accepted back-up plan, they are likelier to use silver than anything else, in nearby worlds in which all of their gold is stolen. So the counterfactual may well come out true, due to the collective acceptance.

However, the important question for our purposes is whether the truth of this collectively accepted counterfactual amounts to the generation of a hypothetical institutional fact. In my view, it does not. Just like it did in the first modal contribution we considered, collective acceptance is just making it the case that people would be *likely* to generate a certain institutional fact in certain non-actual scenarios. The fact that people would be likely to generate a certain fact in non-actual scenarios does not yet generate a fact of any kind.

Moreover, it's worth noting that if the possibility in the antecedent were to become actual—if their gold were actually to be stolen—the group's previous collective acceptance of the counterfactual would not then automatically determine what their currency would be. If they forgot or ignored their previous collective acceptance and began using bronze as their currency, I don't think we'd feel any temptation to say that silver is nonetheless their currency, just because that's what they had previously collectively accepted before their gold was stolen. So even though collective acceptance of the counterfactual may make the counterfactual true and affect what institutional facts people would

[10] This approach to the semantics of counterfactuals comes from Lewis (1973) and Stalnaker (1968).

be likely to generate in certain non-actual scenarios, we still lack a reason to think that collective acceptance metaphysically determines any institutional fact.[11]

A final problem for pluralism, whether strong or weak, is that it implies the possibility of metaphysically inconsistent institutional reality. Thomas Brouwer (2022, p. 24) contends that any kind of two-level metaphysics for social phenomena, such as Epstein's anchoring/grounding framework, allows inconsistencies to be generated. Specifically, in Epstein's framework, people can easily anchor a particular frame principle that, in combination with the obtaining of some particular fact, makes it the case that *A*, and anchor another frame principle that makes it the case that not-*A* (Brouwer, 2022, p. 30). Brouwer (borrowing from Priest (1987/2006)) offers an example involving a group of people who anchor the principle 'all people whose property holdings exceed threshold *X* can vote' and then also the principle 'no women can vote.' They do not consider the possibility of a woman ever coming to exceed the designated property threshold, but then eventually one woman in the group does. This woman, according to Brouwer, both can and cannot vote (p. 30). Brouwer thinks inconsistencies of this kind can also arise within Searle's (2009) and Tuomela's (2013) views, because their views can be interpreted as two-level, but at this stage of the argument we'll focus just on how the point applies to Epstein (Brouwer, 2022, p. 24).

Although the possibility of inconsistency is present in any two-level view of the metaphysics of institutional reality, pluralism greatly amplifies it. If we had a two-level view according to which, for instance, only collective acceptance could anchor frame principles, we might at least hope that some potential inconsistencies would be noticed by those doing the collective accepting. But once frame principles can be anchored by anything from collective acceptance, to individual mental states, to mere regularities in practice, it is easy to imagine inconsistencies proliferating (*cf.* Brouwer, 2022, pp. 29–30).

[11] The two modal strategies for the weak pluralist were suggested by anonymous referees.

For instance, imagine that the leadership of a sport with a high injury rate gathers together and explicitly agrees to a principle that imposes the status of "watchdog" on a particular physician, purporting to give her the authority to stop play at any time when a particular match appears dangerous. But as a matter of fact, the physician almost never attempts to stop play, and when she does, the players do not comply. According to a pluralist picture, it seems that explicit agreement would make it the case that the physician has the role of watchdog within the institution, and regularities in practice would make it the case that the physician does not have the role of watchdog within the institution. The physician both is, and is not, a watchdog within that institution.

Brouwer (2022, pp. 42–43) portrays the possibility of metaphysical inconsistency as an interesting and unproblematic implication of two-level views, but to my mind it seems like a damning objection. While it's certainly true that our attempts at setting up institutional reality sometimes are inconsistent (as in Brouwer's examples (pp. 30–31)), a view according to which such errors actually succeed in creating a metaphysically inconsistent institutional reality seems deeply misguided.

To show why, I first want to note that the kind of inconsistency at issue here is inconsistency *within* a single institution, and not between institutions. For instance, a person could be (and is) Prime Minister with the institution of Canada, and simultaneously not Prime Minister within the institution of the United Kingdom. There is nothing truly inconsistent there, as the two institutions are simply separate. The kind of inconsistency that Brouwer points to, and that I find so troubling, is inconsistency *within* a single institution. In our example, it is within one institution that a particular physician both is, and is not, a watchdog.

The idea of metaphysical inconsistency within a single social institution is untenable. An argument against dialetheism in general would be beyond the scope of this paper, but I do want to say a few words about why dialetheism about institutional facts in particular is misguided. In other words, even if dialetheism is plausible for some kinds of facts (such as, for instance, facts about the application

of vague predicates), it would still be implausible for the domain of institutional facts.[12] This is because institutional facts tend to have significant practical import. For instance, in the women's suffrage example, presumably one of four things will happen. First, people in the relevant group might not notice that any women have reached the property threshold, and thus in practice no women will have the ability to vote. Second, the people administering the election may overlook the principle that no women can vote, and then allow all women above the property threshold to vote. In this second case, women who meet the threshold would have the ability to vote. Or, there's a third possibility: there may be some messy mixture of some women above the property threshold being allowed to vote, and others not, depending on which frame principle the specific people involved are thinking of at the time. And finally, a fourth possibility is that people become aware of the inconsistency in their statutes and begin to treat the issue of whether women who exceed the property threshold can vote as unsettled. But what will *not* happen is that any particular woman on any particular occasion of voting will be treated like she both can and cannot vote. She will either be able to vote, or she will not. In other words, unlike the question of whether someone is bald when they have only a few small patches of hair on their head, questions about whether a particular institutional fact obtains get resolved one way or another in practice, as people proceed with the actions needed to fulfill the institution's practical functions. We are unable to embody these kinds of inconsistencies in our actions. Thus, implying the possibility of inconsistent institutional facts is another problem for pluralism.

So, it seems pluralism is not a viable strategy for preserving a role for collective acceptance in the metaphysics of institutional reality. Whether we take pluralism to mean that collective acceptance *always* plays an anchoring role when it is present, or that it *sometimes* does, problems arise. Moreover,

[12] On dialetheism for vague predicates, see, *e.g.*, Dominic Hyde (1997, p. 645), who mentions the examples of something being both a seedling and not a seedling for a period of time while growing into a tree, and a person being both bearded and not bearded for a period of time while growing a beard.

we've seen that two-level approaches to the metaphysics of social institutions, whether pluralist or not, confront the serious problem of countenancing inconsistent institutional facts.

4. Against the turn to individual mental states

Next, we'll criticize the strategy of replacing collective acceptance with individual mental states in response to the counterexamples to the collective acceptance approach, which is what the originators of those counterexamples suggest. Sánchez-Cuenca (2007, p. 181) recommends appealing to individuals' *expectations*. On this view, part of what makes it the case that a person who seized control by force is a group's leader is group-members' individual expectations about the consequences of failing to comply with the new government. Similarly, both Turner (1999, p. 225) and Viskovatoff (2003, p. 28) think individual *intentions* to make or accept payments using certain objects and *beliefs* that others will accept those objects as payment play a crucial role in non-coinage monetary systems. Mäkelä and Ylikoski (2003, p. 265) also give a key role to beliefs. And in a similar vein, Francesco Guala (2014, pp. 62–66) alludes to counterexamples to collective acceptance accounts that are similar to the ones we discussed in Section 2, and then endorses an approach that appeals to individual beliefs. Machery (2014, pp. 88, 98–99) takes a slightly different approach, expressing openness to the idea that a necessary condition for social institutions may be our general ability to have some as-yet-unspecified kind of mental state (rather than requiring that a token of some particular kind of mental state be present in every single case).

As I mentioned in Section 2, turning toward individual mental states seems promising because although collective acceptance is absent from the Section 2 counterexamples, individual mental states are surely present. In a gradually emerging monetary system, people do seem to have individual beliefs, intentions, and expectations related to the use of their currency. Similarly, people who think their leader's political power is a natural fact still have individual beliefs, intentions, and expectations related

to that power. And when political power is seized by force, even though no one in the group may have the individual belief that the conqueror is their leader, they do have all sorts of individual beliefs, intentions, and expectations related to the conqueror's power and their compliance with her demands. Nonetheless, I'll argue that individual mental phenomena do not play a role in the metaphysics of institutional reality.

Note that here we'll be considering just single-level metaphysical accounts of social institutions, to try to preclude the possibility of inconsistent institutional facts brought out at the end of Section 3. So, we're considering the possibility that people's expectations, intentions, or beliefs are direct metaphysical determinants of institutional facts and entities. A particular government exists because the right people have the right expectations, beliefs, and/or intentions. A particular piece of paper is money because people believe others will accept it as payment (or because it is a token of a type such that people believe others will accept tokens of that type as payment[13]). Moving away from a two-level approach means that we will no longer use Epstein's notion of anchoring. I will remain neutral about exactly what the metaphysical relation is supposed to be between mental states and the institutional facts and entities they give rise to, because my criticisms apply regardless of what the relation is. But we can take it to be some kind of metaphysical determination relation, such as grounding.

I will offer two key reasons to think that the individual mental states our authors suggest are unsuited to being metaphysical determinants of institutional reality. The first, the *location problem*, applies to any portrayal of individual mental states as metaphysical determinants of social institutions, considered as entities. The second, the *epistemic privacy problem*, applies to portrayals of individual mental states as metaphysical determinants of either whole institutional entities, or of individual institutional facts.

[13] This shift to talking about types may actually risk smuggling in an implicit two-level view, which would permit inconsistency again. If so, that would be a third problem for the strategy of shifting to individual mental states, in addition to the two I will discuss below.

We'll begin with the location problem: taking the physicality of institutions seriously reveals that mental states are in the wrong location to give rise to social institutions. First, let me say a bit to establish the physicality of social institutions. In recent discussions of institutional reality, there has been so much emphasis on what participants are thinking, or on what they have collectively accepted, that I think we sometimes lose sight of the physicality of living in a world heavily populated by institutions. We have very physical run-ins with social institutions on a regular basis. For instance, if I try to drive across an international border without permission, an institution will prevent me from doing so. Another institution might stop me from entering a building over which it has control, via a locked door or a security guard. The institution that employs me might physically allow me to park my car in a certain place, by means of an electronic transponder that a participant in the institution has given me. It also lets me into a certain building, and a certain office, by means of keys I have been given, and prevents me from accessing other offices using locks in which my keys do not fit. The institution physically enables me to speak to 100 students at a certain time each week, by giving me access to a certain classroom and microphone. Another institution, perhaps a shopping mall, might let me access a particular building, by leaving the doors unlocked. A store inside the mall physically permits me to leave with one item, after I give a participant in that institution some money. If I attempt to take an item without paying, another institution might physically remove me and physically force me to appear in court. A corporation physically enables me to talk to my parents, via a mobile phone and cellular networks. There is, in fact, a great deal of physicality to social institutions. Social institutions are things we "run into" in the world. They physically stop us from doing some things, physically do certain things to us, and physically enable us or allow us to do other things.[14]

[14] This paragraph parallels Searle's (2009, pp. 90–91) description of how surrounded we are by institutional reality, though with a shift to emphasizing its physicality.

Mental states may be physical, too, if materialism about the mind is right. Or, if mental states are not physical themselves, then they must be at least partially metaphysically determined by physical things (namely, brain states). But the physical stuff that is identical to, or that metaphysically determines, mental states is *not* the kind of physical stuff we run into in the world. When I go to work, I run into parts of my institution that have direct physical effects on me. I do not run into people's brain states, because their brain states are located inside their heads. In other words, mental states cannot give rise to institutions because, even if they too are physical, they are simply in the wrong location.[15]

To see why location matters, it will be helpful to contrast metaphysical determination with causation. It is close to being a consensus view in philosophy that "[c]ausal determination is horizontal, and noncausal building is vertical" (Bennett, 2017, p. 67).[16] This implies that a cause and its effect do not need to be co-located. A causal impact can originate at one location and be felt at another. But when it comes to metaphysical determination, the relation is vertical. And when we are talking about physical objects, broadly construed, the verticality of metaphysical determination implies a need for shared location. For instance, the arrangement of atoms that gives rise to my car must be in the same location as my car, unlike the engineers and factory workers who merely caused my car to exist. The arrangement of cheerleaders that gives rise to a pyramid must be in the same location as the pyramid, unlike the coach who trained them. Similarly, the physical material that gives rise to an institution must be in the same location as that institution, and the individual mental states of members of the group

[15] One might wonder what implications Clark and Chalmers's (1998) extended mind thesis has for this argument. On that view, not all aspects of mental states are located inside people's heads. If an electronic transponder could be seen as itself part of someone's mental state, such as perhaps an intention to let me into a certain parking lot, then that aspect of that intention *would* be in the right location to be a metaphysical determinant of an institution. But unlike, for instance, a "to do" list, an electronic transponder does not fulfil the functional role of an intention. A "to do" list fulfils the function of an intention by making the haver of an intention likelier to do certain actions in the future. A transponder does no such thing for the person who issues it; instead, it makes further action on their part unnecessary.

[16] Bennett goes on to present a subtler picture of the relationship between causation and other building relations, as she calls them, but she does seem to retain the basic idea that causation is horizontal (2017, p. 69).

in which an institution obtains do not satisfy that criterion. The mental states, and the brain states to which they are identical or by which they are metaphysically determined, are located inside people's heads. But the institutions that I run into when I attempt to cross an international border without permission, or enter a shopping mall in the middle of the night, are clearly not inside anyone's head.

One might respond here that although I've shown that institutions do have physical aspects, and that mental states are not metaphysical determinants of those physical aspects of institutions, there is still room to say that institutions have mental aspects as well, which in fact are plausibly seen as located inside participants' heads, and thus can be metaphysically determined by mental states. We might then see social institutions as "metaphysical hybrids," with mental aspects that are metaphysically determined by mental states, and physical aspects that are metaphysically determined by physical phenomena located outside of anyone's head (Stainton, 2014). This move may sound like a turn back to pluralism, but it differs from the kinds of pluralism discussed in Section 3 in that it is not a two-level view. Instead, the idea is that both mental states and physical phenomena outside of anyone's head can act as direct metaphysical determinants of institutions and institutional facts.

However, even without returning to two-level metaphysics, this hybrid entities approach raises the specter of metaphysical inconsistency once more. If both physical phenomena outside of anyone's head, and mental states, could directly metaphysically determine aspects of social institutions, we could end up with cases in which people's mental states supposedly generate an institutional fact *A*, and physical happenings outside of anyone's head supposedly generate an institutional fact not-*A*. For instance, returning to our previous example of institutional inconsistency, members of a sports league might all individually believe that a certain physician has the status of watchdog in their institution, meaning that she has the role of stopping play whenever injury is too likely. And yet the physician never stops the game, or if she attempts to do so, the players always ignore her. According to the

mental aspect of that institution, the physician is a watchdog; according to the physical aspects outside of anyone's head, the physician is not a watchdog.

At this stage, it may be helpful to reorient ourselves in the dialectic. At the beginning of our discussion of the location problem, we established that institutions have physical aspects. Mental states, even if they too are physical, were shown to be in the wrong location to give rise to those physical aspects of institutions. We then considered the possibility that institutions might have mental aspects, in addition to their physical aspects, and that mental states could act as metaphysical determinants of the mental aspects of institutions. But that has led us back to the possibility of metaphysical inconsistency. So, we can now conclude that social institutions do not have mental aspects in addition to their physical ones, and thus that mental states are not metaphysical determinants of any aspects of social institutions. This is not to deny that people tend to have certain typical mental states when they participate in social institutions. Rather, it is just to deny that any such mental states are among institutions' metaphysical determinants.

Before moving on to the epistemic privacy problem, it is worthwhile to consider Machery's (2014, pp. 88, 98–99) idea that even if concrete instances of individual mental states do not metaphysically determine social institutions, our ability to have certain kinds of individual mental states, *in general*, is necessary for social institutions to exist. Ultimately, I think this is right. Without our ability to form complex beliefs about who is running for office and complex intentions about what needs to be done to vote, it seems unlikely that democracies could ever exist. But we can draw from the preceding discussion to see that this fact does not make those individual mental states metaphysical determinants of institutional phenomena (*cf.* Guala, 2014, p. 63). These individual mental states may be *causally* necessary for the existence of institutional phenomena, bearing a horizontal relationship to them, but that does not make them *metaphysical* determinants of those phenomena, where the relationship would have to be vertical. Many other states of affairs, such as having a planet with

sufficient oxygen in its atmosphere, might be causally necessary for the existence of our institutions. But that does not make these myriad factors into partial metaphysical determinants of our institutions.

Next, let's consider the epistemic privacy problem, which applies to portrayals of individual mental states as metaphysical determinants of individual institutional facts as well as of whole institutional entities. Stated briefly, the problem here is that institutional entities and facts are epistemically public, which means they cannot be metaphysically determined by the individual mental states to which our authors appeal, due to those mental states being epistemically private. In my view, epistemically public phenomena are those such that "all minds have equal access to them, *ceteris paribus*," and epistemically private phenomena are those such that "a single mind has privileged access to them, *ceteris paribus*" (Stotts, 2020, p. 197). Facts about the age and width of the Grand Canyon, for instance, are epistemically public: any mind can have equal access to them, provided that the person is appropriately situated, where being appropriately situated may require a particular physical location, particular tools, or other knowledge. On the other hand, your beliefs, expectations, and intentions are epistemically private because you have a kind of privileged epistemic access to them: you can access them in a direct manner that no one else can duplicate.

Institutional phenomena show clear signs of being epistemically public. For example, anyone who is appropriately situated (that is, anyone who is in the geographical area corresponding to Canada's territory, or perhaps in possession of a book about Canada) has equal epistemic access to Canadian institutional phenomena. But the beliefs, expectations, and intentions to which our various authors have appealed are again epistemically *private* because the minds in which they exist have privileged epistemic access to them. My suggestion is that this difference means that these individual mental states are unable to be metaphysical determinants of institutions and institutional facts. To see why, I will draw on Jonathan Schaffer's (2016, p. 95) notion of 'inherited reality' that obtains between

a grounded entity and whatever grounds it.[17] A grounded entity does not get its reality on its own; it inherits reality from an entity that is real independently of it. This basic idea generalizes well to relations of metaphysical determination more broadly: whenever something makes it the case that something else exists, it's natural to describe the second entity as having inherited its reality (perhaps partially) from the first.

Ceteris paribus, when you inherit something from someone else, and that other person's possession of it was restricted in a certain way, your possession of it will be subject to the same restriction. For instance, if I own a piece of land with only the surface rights, it is not possible for someone who inherits the land from me to inherit it with the mineral rights, absent some other source of the mineral rights. Similarly, beliefs, expectations, and intentions are subject to a certain kind of restriction: they can be known in a privileged way only by the person whose mental states they are. If institutions and institutional facts inherited their reality from these individual mental states, then they would possess that reality under the same restriction: they would be knowable in a privileged way only to certain individuals (or at least, certain aspects of their reality would have that restriction, if the mental states are not their only metaphysical determinants). In other words, if institutions and institutional facts inherited their reality from epistemically private mental states, institutions and institutional facts would be epistemically private, too. And so, because institutions and institutional facts are not epistemically private, they must not be metaphysically determined by epistemically private mental states.

In case the inheritance metaphor seems stretched to the breaking point, let me put the point literally: an entity or fact can only ground or otherwise metaphysically determine other entities or facts that have the same kinds of limitations its own reality has. In this particular case, I have described

[17] Schaffer (2016, p. 95) also notes that this idea of inherited reality marks an important difference between causation and grounding.

those limitations epistemically: in terms of who can know it in a privileged way. But importantly, this epistemic feature is going to have some kind of metaphysical basis. It's not a mere accident that only I can know my beliefs, expectations, and intentions in a direct way. Rather, it is the very nature of these mental states—that they are *mine* and not someone else's—that makes them knowable to me in a privileged way. The epistemic upshot of this difference is what is easiest for us to identify, because it is what we experience, but really that epistemic upshot is a sign of a metaphysical feature of these mental states. And we can be sure that it is a metaphysical feature that institutional reality does not have, because it lacks the corresponding epistemic privacy property.

With the epistemic privacy problem laid before us now, let's consider two different objections to it, both of which have to do with ways in which the relevant mental states may not be quite as private as I've made them seem.[18] First, one might object that I framed epistemic privacy in terms of a single person's mental states, but our authors appeal to facts about individual mental states across an entire group. Even if one has privileged epistemic access to one's own belief that Justin Trudeau is the Prime Minister of Canada, what about the fact that *Canadians* believe that Justin Trudeau is the Prime Minister of Canada? Is that kind of group-level fact about belief really so different from the kinds of facts about social institutions that I've characterized as epistemically public, such as the fact that Justin Trudeau is the Prime Minister of Canada?

But as I see it, even individual mental states aggregated at the level of a group are importantly different from institutional phenomena. A Canadian person and, for instance, a Brazilian person both have equal epistemic access to the fact that Justin Trudeau is the Prime Minister of Canada. By reading a book about Canada or just casually overhearing a news report, a Brazilian person can know this institutional fact in the same way a Canadian person can. But when it comes to the fact that Canadians *believe* that Justin Trudeau is the Prime Minister of Canada, the situation is different. A Brazilian person

[18] Both of these objections were inspired by comments from anonymous referees.

can certainly come to know that Canadians hold this belief. But there's a difference in the kind of knowledge a non-Canadian versus a Canadian can have here. If I'm Canadian, and I believe that Justin Trudeau is Prime Minister of Canada, I can have a kind of access to the fact that Canadians believe he is the Prime Minister which a non-Canadian could never have, because my own belief to which I have privileged epistemic access is part of the overall state of affairs of Canadians holding that belief. I have privileged epistemic access to a part of that state of affairs, and so I have a degree of privileged epistemic access to the overall state of affairs.

Now of course, when it comes to institutions, it's true that those who live and act within them know them better than those who only read about them in books. But this is no different than the way in which someone who has hiked through the Grand Canyon many times knows it better than someone who has only seen photographs. The point is that anyone can move to Canada and become as familiar with Canadian institutions as a Canadian, over time, just as anyone can travel to the Grand Canyon and get to know its trails. But a non-Canadian cannot gain the kind of access Canadians have to propositions about what Canadians believe, due to the epistemic privacy of beliefs. So, even when we explicitly acknowledge that the mental states in question are mental states of a group of people and not of just an individual, they remain epistemically private and thus unsuited to be metaphysical determinants of institutions and institutional facts.

Second, one might wonder whether adding a common knowledge requirement to an account of social institutions that appeals to individual mental states could avoid the epistemic privacy problem. This would mean that participants in a given institution would be required to have additional individual mental states amounting to knowledge of each other's relevant expectations, beliefs, or intentions. They would also be required to know that others know about the relevant expectations, beliefs, or intentions, and so on (Lewis, 1975, p. 6; Miller, 2001, p. 59).

I will grant that adding a common knowledge requirement would make those individual mental states public in a very real sense. But although it would make them public in the sense of their being known to everyone, it would not make them epistemically public in my sense. The individual mental states that supposedly give rise to social institutions would still, like all mental states, be known in a direct, privileged way only to those whose mental states they are. They could be known to all of the others, but that would not make them known in that privileged *way* to all of the others. And in having this feature, they would still differ from institutional phenomena, leading once again to the epistemic privacy problem.

I've offered two reasons to think that the individual mental states proposed as replacements for collective acceptance are unsuited to being metaphysical determinants of institutional reality. First, the location problem: taking the physicality of institutions seriously reveals that individual mental states are in the wrong location to give rise to them. And second, the epistemic privacy problem: institutional phenomena are epistemically public, which means that the beliefs, expectations, and intentions to which our authors appeal cannot be their metaphysical determinants, due to the latter being epistemically private.[19]

5. The way forward

Looking back at Sections 2–4, we have found strong reasons against straightforward collective acceptance accounts of institutional reality, against pluralism that preserves a role for collective acceptance, and against approaches that replace collective acceptance with individual mental states. Given these roadblocks, it might seem that social ontology is in trouble. What is the way forward?

[19] As noted previously (footnote 3), Ludwig proposes a collective acceptance approach to social institutions, with collective acceptance defined in terms of intentions or conditional intentions. Whether the intentions to which he appeals are ultimately individual in the sense of the individual mental state views we've discussed in this section is not a simple interpretive question, but it is certainly clear that Ludwig's account is still an account in terms of mental states (2017, pp. 21–35, 132). Due to that fact alone, both the location problem and the epistemic privacy problem apply to Ludwig's view as well.

My suggestion is that we should develop an account of social institutions and institutional facts in terms of just observable behavior, with no role for mental states.[20] My aim in this final section is not to actually develop a behavioral account of social institutions, but just to show that doing so is a promising way to proceed, in light of the problems that emerged in the preceding sections for accounts based in mental phenomena.

At this stage, one might wonder about the possibility of turning toward an equilibrium approach to social institutions, especially because such approaches are often described as having a behavioral focus (*e.g.*, Sánchez-Cuenca, 2007, p. 186; Guala and Hindriks, 2015, p. 178; Hindriks, 2022, pp. 355–356). According to the equilibrium approach, the social situations in which institutions arise can be modeled as certain kinds of games in the sense of game theory, and institutions themselves are equilibria (or phenomena closely related to equilibria) in those games, which means that "individuals have no incentive to deviate from the pattern unilaterally" (Guala, 2016, p. xxiii; *cf.* Hindriks, 2022, pp. 355–356).

To a certain extent, I do see the equilibrium approach as compatible with the kind of approach I favor, if we treat it as just offering explanations of *why* we have social institutions (*cf.* Miller, 2003, p. 245). That kind of inquiry, about the origins and motivation of an institution, is interesting and important, but it is also separate from what I take to be the fundamental project of doing the metaphysics of social institutions, where we aim to identify the kinds of facts *in virtue of which* institutions exist. I also think that it would be a mistake to commit to the equilibrium approach as an explanation of where *all* social institutions come from. It's important to leave it open, as a theoretical possibility, that the nature and existence of some of our social institutions might not be explicable in terms of a pre-existing structure of incentives. We are deeply social beings, who sometimes just glom

[20] This proposal is somewhat similar to the "patterns or regularities in practice" item that Epstein includes in his list of possible kinds of anchors for institutional facts, since that item contrasts with the others in his list by making no explicit mention of mental phenomena. A major difference is, of course, that I favor a unitary rather than pluralist account.

together for no identifiable reason. It should at least be a theoretical possibility that some of our institutions are not equilibria in any sort of game.

And insofar as equilibrium theories do answer the fundamental metaphysical question about social institutions, the sense in which these theories are behavioral is quite different from the sense in which the kind of approach I favor is behavioral. Equilibrium theories focus on behavior in the sense that they try to explain regularities in social behavior. But they are not behavioral in the sense of appealing only to observable behavior in their picture of the metaphysical determinants of social institutions. Schotter (1981, p. 11), for instance, includes a vast number of mental states among the metaphysical determinants of social institutions: what makes it the case that a social institution exists, for Schotter, is not merely that people in some group conform to a certain behavioral regularity, but also that conformity to that regularity is common knowledge among them, that group members expect each other to conform, and so on. And Guala (2016, pp. 54–55), though he eschews some of the mental complexity that Schotter favors, ultimately still builds mental phenomena into his account (*cf.* Guala and Hindriks, 2015, p. 185). In a more recent account, Frank Hindriks (2023, pp. 1381–1382) appeals to mental states in an equilibrium theory as well. So, although the goal of equilibrium theories is to explain behavioral regularities, the theories themselves tend to appeal to mental phenomena.

To see why a behavioral account of institutional reality in my sense (that is, one that appeals only to observable behavior in the theory itself) is plausible, it will be helpful to temporarily widen our scope to encompass social reality more broadly. As R. A. Wilson (2007, pp. 148–149) points out, social reality includes not just institutions but also social groups, friendships, and crucially, social structures among non-human animals. Wilson mentions schooling and flocking among fish and birds, as well as social structures among insects. We'll focus on social structures that emerge among insects (specifically, ant colonies) because this kind of social phenomenon bears the closest analogy to human social

institutions, as will become clear below.[21] Associated with an ant colony, there are social facts, such as the fact that some particular ant is the queen. To attribute mental states to ants would be to veer decisively into the realm of analogy. So, because a robust social structure exists among ants in the absence of mental phenomena, it seems that the only thing left to give rise to it is the ants' behavior (*cf.* Wilson, 2007, pp. 148–149). But what, specifically, makes it the case that there is an ant colony, rather than just a bunch of individual ants engaging in behavior near each other? Understanding what makes it the case that social structures exist among beings that lack mental states can aid us in seeing how social institutions can arise among humans in a way that is not metaphysically determined by mental states.

The first step is to notice that the ants in a colony tend to repeat certain behaviors and abstain from others in a way that is structured: certain ants repeatedly engage in behavior of kind *A* while abstaining from behavior of kind *B*, whereas other ants repeatedly engage in behavior of kind *B* while abstaining from behavior of kind *A*. And in fact, this is the case for not just single kinds of behavior, but for little clusters of kinds of behavior. There is one ant (the queen) that remains in the nest and produces eggs. Among the remaining ants (the worker ants), certain ants repeat kinds of behavior related to foraging for food; others repeat kinds of behavior related to defending the nest; others repeat kinds of behavior related to nurturing the larvae. And, ants abstain from the kinds of behavior in the other clusters. Furthermore, different individual ants engage in these clusters of behavior at different times: when older worker ants die (or move on from nurturing behavior to foraging behavior, as happens among some ants), younger ants step in to engage in behavior in the same clusters. This gives rise to what we can identify as different *roles* that various ants can play: the queen versus the

[21] Epstein (2015, p. 163) would likely dispute the characterization of ant colonies as social, but my hope is that the discussion that follows will convincingly display the fruitfulness of bringing ant colonies and human institutions under the same conceptual umbrella.

workers, and then among the workers, the foragers versus the soldiers versus the nurses (Gullan and Cranston, 2014, p. 337).

These repeated behaviors are not interconnected merely in the sense that they give rise to roles, but also in the sense that they work together to promote some particular result: the survival and reproduction of that particular group of ants and their descendants.[22] Contrast the ants with several alligators that happen to live near each other. The alligators engage in a lot of similar behavior: lurking underwater, lunging toward prey, and sleeping. And they also each have their own territory, so we can discern something similar to roles in the alligators' tendency to move around in one geographic area and not in other adjacent ones. But although each alligator's behavior promotes *that* alligator's survival, the alligators' behavior does not collectively promote any result. Each alligator could engage in the same behavior and achieve the same result, even if the other alligators suddenly disappeared. That is not the case for ants: an ant that filled the soldier role while no other ants were filling the forager role would not be able to achieve the result of keeping even itself alive. The result(s) which the individuals' behavior collectively brings about makes the difference between organisms that simply engage in similar behavior in proximity to each other, and genuine social structures that emerge within a social species.[23]

So, there seem to be three main features that make it the case that an ant colony exists: repeated kinds of behavior, interconnection among those repeated kinds of behavior that allows them to cluster into roles, and one or more results that all of the interconnected behavior promotes together. Or, to make matters more concrete, an ant colony is a social structure among ants characterized by the fact that the repeated behaviors related to producing eggs, gathering food, defending the nest, and

[22] I should note that I use 'group' to mean just a collection of individuals. It should not be taken to mean something like a genuine social group in the sense of, *e.g.*, Epstein (2015, Part 2).
[23] This emphasis on the importance of the result at which behavior aims is inspired by Seumas Miller's (2001, p. 181) notion of a collective end and its role in social institutions.

nurturing the young are structured into clusters that give rise to the roles of queen, foragers, soldiers, and nurses, all of which promotes the survival and reproduction of that group of ants and their descendants.

Turning back toward institutional reality now, we can begin by noting that from the outside, human social institutions aren't all that different from ant colonies. For instance, if we consider an athletic institution such as Major League Baseball, we can observe repeated kinds of behavior such as throwing balls, swinging bats when balls are thrown, running from one white mound on the ground to another, producing certain sounds and gestures in between instances of throwing, hitting, or running, and sitting in a stadium while yelling and drinking beer. The first three behaviors are part of the cluster that creates the role of player (with more specific sub-clusters for pitchers, outfielders, *etc.*), the fourth is part of a cluster that creates the role of umpire, and the fifth is part of a cluster that creates the role of fan. The results that all of this behavior promotes are athletic achievement and entertainment. It seems quite plausible that this structure—these connections among individual humans' behavior and the results it all promotes—is what makes it the case that the institution of Major League Baseball exists, just as the connections among individual ants' behavior and the results it all promotes are what make it the case that an ant colony exists.[24]

This being said, I do think it would be a mistake to go too far in assimilating human institutional reality to the kind of social reality we find among much simpler organisms. Human behavior (and thus, our institutional reality) displays a degree of variety, complexity, and flexibility not present among ants (*cf.* Guala, 2016, p. xix). Ants do not have multiple options for social structures available to them; they can live only in colonies structured around a queen, with the roles of foragers, nurses, and soldiers. We do not find alternatives analogous to republics, constitutional monarchies,

[24] Ludwig (2017, p. 5), too, gives roles a central place in his understanding of social institutions. But Ludwig's understanding of roles is very different from what I am suggesting here (most notably in being intention-dependent) (pp. 138–139).

and dictatorships among ants. And of course, the roles and kinds of behavior involved in an institution such as the U.S. government are vastly more numerous and complexly interrelated than those in an ant colony, and there is sometimes much more ability for individuals to change roles. Because of these significant differences, we do need to differentiate social institutions from other kinds of social structures. But interestingly, all of these differences seem to be observable differences in the *behavior* that gives rise to institutions versus mere social structures: human institutional behavior is (observably) more complex, (observably) more prone to change over time, and (observably) more varied across different groups. So, the strategy of developing a behavioral account of institutional reality seems highly promising, due to both its ability to build upon the behavioral metaphysics of non-human social structures such as ant colonies and its ability to differentiate institutional reality from those non-human social structures.

These suggestive comments are a long way away from an actual account of social institutions in behavioral terms. For instance, discussion is needed of how not just people but also objects (such as baseballs, or pieces of paper used as money) can come to have roles within institutions, as a result of the kinds of behavior in which they are repeatedly and systematically involved. And, the differences between social institutions and simpler social structures need much more clarification.

It will also be important to ensure that the account, in its final form, does not allow for the possibility of inconsistent institutional facts. This means it should not be a two-level view, but also we'll need to make sure inconsistency isn't permitted in some other way. It will be important, I think, to formulate the account in such a way that what roles various individuals have in each institution is determined by something like a preponderance of behavior. Who or what has which roles in an institution at a given time will be determined by all of the relatively recent behavior within the institution taken collectively; smaller subsets of behavior within the institution will not have separate power to give rise to institutional facts. Thus, either the behavior within the institution, as a whole,

will give an individual a certain role, or it will not. There will not be room for some people's behavior to give someone or something a role and other people's behavior to simultaneously imply that it does not have that role, within a single institution. Again, the details need much more working out here.

Nonetheless, we can see that turning toward behavior provides a promising way forward in light of the various cases and problems we've discussed. Monetary and political institutions, like all other institutions, will be seen as existing in virtue of behavior that promotes some characteristic result(s) and clusters into roles. For a monetary institution, whether coin-based or not, we will have roles such as buyer and seller, comprising behaviors that involve some kind of object being exchanged for goods and services. All of that behavior will promote, presumably, the result of efficient trading. And for a political institution, whether rightful or based just on fear of force, we will have roles such as leader and subject or citizen, again comprising distinctive behavioral clusters. This will all promote some result(s), such as social stability or perhaps the wellbeing of citizens, depending on the particular institution. If participants in these institutions have beliefs, collective acceptance, or other mental states according to which the nature of their institution differs from the institution their behavior actually makes, those mental states (like all others) will simply be metaphysically irrelevant. And importantly, a behavioral approach will avoid the location problem because behavior is out in the world where institutions operate and not inside anyone's head. It will also avoid the epistemic privacy problem because observable behavior is epistemically public: all minds can have the same epistemic access to it.

One final concern I'd like to address is the concern that what I'm proposing here amounts to a form of behaviorism.[25] To the contrary, I think that our accounts of some other social phenomena, such as social norms, must make reference to mental phenomena. And I am not committed to (nor am I inclined to endorse) any general claims about mental states being in some way secondary to

[25] For instance, see Skinner 1953, p. 35, for claims about behavior and explanation that sound similar to some of mine.

behavior. I think that human behavior is typically caused by mental states, and that mental states are fit objects of scientific and philosophical inquiry (notwithstanding the fact that there is a difference in kind between the sort of knowledge we can have of them scientifically, and the first-person knowledge one can have of one's own mental states). My contention is just that one particular social phenomenon, social institutions, is not metaphysically determined by mental states, real and important though those mental states are.

Working out the details of a behavioral account of social institutions is something I look forward to pursuing in future work. For the present, I hope to have shown that the problems that accompany collective acceptance views, pluralist views that preserve a role for collective acceptance, and accounts that appeal to individual mental states reveal that the most promising way forward is in a behavioral direction.

Acknowledgements

For helpful feedback on past versions of this article, I am grateful to Graham Hubbs, Stefan Sciaraffa, several anonymous referees, and audiences at Vassar College, the University of Idaho, McMaster University, and the Social Ontology 2018 conference. For research assistance, I am grateful to Alex Bryant, Michaela Murphy, Siddharth Raman, and Gary Spero. I am also grateful to the students in my Fall 2018 and Fall 2022 graduate seminars at McMaster University, as well as Brent Odland, for helpful discussion on topics related to this article. This article draws on research supported by the Social Sciences and Humanities Research Council of Canada.

References

Bennett, K. (2017). *Making things up*. Oxford: Oxford University Press.

Brouwer, T. N. P. A. (2022). Social inconsistency. *Ergo*, *9*(2), 19–46.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis*, *58*(1), 7–19.

Epstein, B. (2014). Social objects without intentions. In A. K. Ziv & H. B. Schmid (Eds.), *Institutions, emotions, and group agents: Contributions to social ontology* (pp. 53–68). Dordrecht: Springer.

Epstein, B. (2015). *The ant trap: Rebuilding the foundations of the social sciences*. New York: Oxford University Press.

Epstein, B. (2016). Replies to Guala and Gallotti. *Journal of Social Ontology,* 2(1), 159–172.

Guala, F. (2014). On the nature of social kinds. In M. Gallotti & J. Michael (Eds.), *Perspectives on social ontology and social cognition* (pp. 57–68). Dordrecht: Springer.

Guala, F. (2016). *Understanding institutions: The science and philosophy of living together*. Princeton: Princeton University Press.

Guala, F., & Hindriks, F. (2015). A unified social ontology. *The Philosophical Quarterly*, *65*(259), 177–201.

Gullan, P. J., & Cranston, P. S. (2014). Insect societies. In *The insects: An outline of entomology* (pp. 322–353). Chichester: John Wiley & Sons.

Hindriks, F. (2022). Institutions and their strength. *Economics and Philosophy*, *38*, 354–371.

Hindriks, F. (2023). Rules, equilibria, and virtual control: How to explain persistence, resilience and fragility. *Erkenntnis*, *88*, 1367–1389.

Hyde, D. (1997). From heaps and gaps to heaps of gluts. *Mind*, *106*(144), 641–660.

Lewis, D. (1973). *Counterfactuals*. Malden: Blackwell Publishers.

Lewis, D. (1975). Languages and language. In K. Gunderson (Ed.), *Language, Mind and Knowledge: Minnesota Studies in the Philosophy of Science, Vol 7* (pp. 3–35). Minneapolis: University of Minnesota Press.

Ludwig, K. (2017). *Collective action: Volume 2: From plural to institutional agency*. Oxford: Oxford University Press.

Machery, E. (2014). Social ontology and the objection from reification. In M. Gallotti & J. Michael (Eds.), *Perspectives on social ontology and social cognition* (pp. 87–100). Dordrecht: Springer.

Mäkelä, P., & Ylikoski, P. (2003). Others will do it: Social reality by opportunists. In M. Sintonen, P. Ylikoski, & K. Miller (Eds.), *Realism in action: Essays in the philosophy of the social sciences* (pp. 259–268). Dordrecht: Kluwer Academic Publishers.

Miller, S. (2001). *Social action: A teleological account*. Cambridge: Cambridge University Press.

Miller, S. (2003). Social institutions. In M. Sintonen, P. Ylikoski, & K. Miller (Eds.), *Realism in action: Essays in the philosophy of the social sciences* (pp. 233–250). Dordrecht: Kluwer Academic Publishers.

Priest, G. (1987/2006). *In contradiction: A study of the transconsistent* (2nd ed.). Oxford: Oxford University Press.

Sánchez-Cuenca, I. (2007). A behavioural critique of Searle's theory of institutions. In S. Tsohatzidis (Ed.), *Intentional acts and institutional facts: Essays on John Searle's social ontology* (pp. 175–190). Dordrecht: Springer.

Schaffer, J. (2016). Grounding in the image of causation. *Philosophical Studies*, *173*, 49–100.

Schotter, A. (1981). *The economic theory of institutions*. Cambridge: Cambridge University Press.

Searle, J. (1995). *The construction of social reality*. New York: The Free Press.

Searle, J. (2009). *Making the social world: The structure of human civilization*. New York: Oxford University Press.

Spencer, Q. (2019). How to be a biological racial realist. In *What is race? Four philosophical views* (pp. 73–110). New York: Oxford University Press.

Stainton, R. J. (2014). Philosophy of linguistics. *Oxford Handbooks Online*.

Stalnaker, R. C. (1968) A theory of conditionals. In N. Rescher (Ed.), *Studies in Logical Theory* (pp. 98–112). Oxford: Blackwell Publishers.

Stotts, M. H. (2020). Toward a sharp semantics/pragmatics distinction. *Synthese*, 197(1), 185–208.

Tuomela, R. (2002). *The philosophy of social practices: A collective acceptance view*. Cambridge: Cambridge University Press.

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

Tuomela, R. (2013). *Social ontology: Collective intentionality and group agents*. New York: Oxford University Press.

Turner, S. P. (1999). Review: Searle's social reality. *History and Theory*, 38(2), 211–231.

Viskovatoff, A. (2003). Searle, rationality, and social reality. In D. Koepsell & L. S. Moss (Eds.), *John Searle's ideas about social reality: Extensions, criticisms and reconstructions* (pp. 7–44). Malden: Blackwell.

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