sciendo

Social Explanation: Structures, Stories, and Ontology. A Reply to Díaz León, Saul, and Sterken

Sally Haslanger
Massachusetts Institute of Technology

DOI: 10.2478/disp-2018-0015 BIBLID [0873-626X (2018) 50; pp.245–273]

Abstract
In response to commentaries by Esa Díaz León, Jennifer Saul, and Rachel Sterken, I develop more fully my views on the role of structure in social and metaphysical explanation. Although I believe that social agency, quite generally, occurs within practices and structures, the relevance of structure depends on the sort of questions we are asking and what interventions we are considering. The emphasis on questions is also relevant in considering metaphysical and meta-metaphysical issues about realism with respect to gender and race. I aim to demonstrate that tools we develop in the context of critical social theory can change the questions we ask, what forms of explanation are called for, and how we do philosophy.

Keywords
Social structure, structural explanation, implicit bias, gender, social ontology.

1 Introduction

It is a great privilege to have an opportunity to engage the wonderful papers by Esa Díaz León, Jennifer Saul, and Rachel Sterken. All of them have raised important challenges to my work and this has helped me think through the issues more deeply. There are many points and arguments that deserve more attention than I can give them here. However, in an effort to provide some unity to my response, I'm going to focus on issues of explanation, since they are raised, in some way or other, by all of them.

Sterken's paper is most explicitly on explanation. She argues that

my account of structural explanation, insofar as it privileges broad and deep structures over local and flexible ones, is implausible. Moreover, attention to more local and flexible explanations provides us better resources for intervention into the particular conditions that constitute and produce injustice.

Saul's paper takes up my criticisms of implicit bias stories and considers the extent to which implicit bias explanations of social injustice are problematically individualistic. She argues that neither structural nor individualistic explanations of social injustice are fully adequate to the phenomena, and that each works best if supplemented by the other. Moreover, effective social change requires attention to both structural/institutional phenomena and psychological ones.

Díaz León's paper invites me to situate my view of gender and race within debates over metaphysical deflationism v. metaphysical realism. One might see this issue also as a matter of metaphysical explanation: what explains the correctness or incorrectness of my accounts of gender and race (and related accounts of other social kinds)? In particular, is my account of gender, for example, correct because it cuts the world at its social joints? Or because it offers an apt interpretation of the terms 'man,' 'woman,' and 'gender' that do justice to our social and political aims? Or are there other possible interpretations of what's at issue?

In what follows, I will start by spelling out in a bit more detail my views about explanation and related views in meta-metaphysics. I'll then turn to more specific arguments raised in the papers by Díaz León, Saul, and Sterken.

2 The erotetic account of explanation

According to the erotetic account of explanation, explanations are answers to questions. Whether an explanation is good or bad, adequate or inadequate, depends in part on the question at issue. I have embraced a particular form of erotetic account sketched by Alan Garfinkel (1981) and Mark Risjord (2000). I find that this view offers us resources to make sense of certain kinds of social explanation, even if it does not provide a theory that works across the board.

On Garfinkel's account, many explanatory projects seek out *difference makers*, and the best way to undertake such a project is to frame

the guiding question in terms of *focus* and *foils*. Suppose you want to know why there is an increasing number of rabbits in a region, say, on the MIT campus. We might ask:

(0) Why is there an increasing number of rabbits on the MIT campus?

But this question is imprecise. It might be further specified in a variety of ways:

(1) Why is there an *increasing* (rather than stable or decreasing) number of rabbits on the MIT campus?

(2) Why is there an increasing number of *rabbits* (rather than rats) on the MIT campus?

(3) Why is there an increasing number of rabbits on the *MIT campus* (rather than in the broader neighborhood)?

In each of these cases (1–3), the answer must point to relevant differences, viz., in the case of (1), between the conditions that produce stability or decrease of rabbits and those that produce increase; in the case of (2), between rabbits and rats; and in (3), between the MIT campus and rest of the neighborhood.

We may also want to ask questions at different levels of generality, and this is reflected in the focus and foils. For example, we might ask,

(4) Why is *Sparky* (rather than Rufus) a good pet for urban apartment life?

(5) Why are Cockapoos (rather than Briards) good pets for urban apartment life?

(6) Why are dogs (rather than wolves) good pets for urban apartment life?

Notice that even if the question is phrased very specifically (as in (4)), the best answer might be at a higher level of generalization. If Sparky is a Cockapoo and Rufus is a Briard, then the best answer to (4) might be in terms of breed, in which case the question is pushed back to (5).

Garfinkel suggests that there is a kind of trade-off between

questions that seek explanations of very specific events, and those that allow the target phenomenon to be a broader range of similar cases. His example involves a car accident:

> Suppose that I got up one day and went out for a drive. I was doing about 110 when I rounded a bend, around which a truck had stalled. Unable to stop in time, I crashed into the truck. Later, chastising me for the accident, you say, "If you hadn't been speeding, you wouldn't have had that accident." I reply, "Yes, that's true, but then if I hadn't had breakfast, I would have gotten to that spot before the truck stalled, so if I hadn't eaten breakfast, I wouldn't have had the accident. Why don't you blame me for having had breakfast?" (Garfinkel 1981: 30)

He continues:

> [In order to distinguish the cases,] We need something in addition to represent what is really getting explained, something that will account for the fact that my objection somehow misses the point. For not any difference from that very accident is going to count as relevantly different, only certain ones will. And so we need, in addition to the event, a set of perturbations which will count as irrelevant or inessentially different. These irrelevant perturbations determine an equivalence relation,, "differs inessentially from," and the real object of explanation is an equivalence class under this relation. The equivalence relation determines what is going to count as the event's not happening. (Garfinkel 1981: 30–1)

This choice of explanatory target is not arbitrary. Garfinkel argues that "as the equivalence classes become smaller and more numerous, the resulting object, and hence the resulting explanation, becomes less and less stable." (1981: 31) So, for example, *speeding* may be a stable difference maker between cases in which cars traveling around the bend are involved in accidents and cases in which they are not. But *driver recently having breakfast* is not. Likewise, there may be specific facts about Sparky's past that make *him* in particular a good pet as opposed to Rufus, but this explanation will be less stable than one that treats Sparky as a member of a broader kind (generalizing over irrelevant "perturbations"), e.g. the fact that he is a member of the group of dogs that need little exercise, are easily trained, and are friendly with strangers, (whereas Rufus is not) provides the most stable explanation. The point here is that detail does not always strengthen an explanation.

Looking more closely at explanatory questions with this in mind, we can see that even when we have specified a question in terms of focus and foils, there will be background assumptions that are not explicit. For example, if we are considering what makes for a good pet in an urban apartment, the background conditions set constraints, e.g., that neighbors are annoyed by barking, that urban apartments tend to be small and do not provide direct access to outside space, that parks for exercise may be some distance away, that some dogs are more territorial than others. Pointing to these structural constraints may also be relevant to explanation. For example, in order to explain why Cockapoos are better than Briards for urban apartment living, a relevant difference is that Briards, and not Cockapoos, need vigorous daily exercise. But this difference is relevant only given certain background conditions, e.g., that urban apartment living does not usually offer adequate opportunities for vigorous daily exercise that Briards need. On my view, structural explanations often make reference to background facts that clarify what difference between the focus and foils is relevant.

In his recent work on explanation, Brad Skow (2018) argues that structural explanations can be reframed so that they provide causally relevant information. The idea is that structural explanations provide information about why a cause has (or would have) the effect that it does (or why causes of a particular kind have (or would have) the effects of a particular kind). He considers an example of a special house that has a room that is inaccessible from any other room in the house. One might wonder, why it is that whatever pattern of movements one makes in the house, one never gains access to the room. The answer to the question should make reference to the structural features of the house—where there are walls, where there are doors, etc.—that prevent one from entering the room. The question might be:

(7) Why does walking through the house in ordinary ways (using existing doors, stairways, etc.) *never* (v. sometimes) give one access to the secret room?

The answer will make reference to the fact that the existing structure of doors, stairways, wall, etc. prevent access, and in order for there to be an occasion when one could gain access one would have to change the architecture of the house.

To take a more everyday example, suppose we consider a pinball machine and ask why, however one launches the ball, it always ends up back at the start, ready to be launched again, i.e., why does the launching of the ball always have the effect of returning the ball to the launch site? The answer is structural: the table is designed on a slope with a "drain" that leads to the launch. Of course, if we are asking about a particular launching of a ball we might want to consider how or why *it* returned to the launch site—what particular velocities, spins, active bumpers, flipper hits, and such—returned it to the launch. But a story that provides such detail would include a lot of extraneous information that is not necessary: virtually *no matter how* the ball was launched and no matter what bumpers and flippers it hit (providing, of course, that it stayed on the table), it would have returned to the launch, given the structure of the table. Extra detail about the ball and its trajectory also makes the explanation less stable.

Drawing on these considerations, there are several ways in which structure is relevant to explanation (see Haslanger 2016: 116):

First, questions calling for explanations presuppose a structure of options: why *this* rather than *that (or that, or that…)*. The options may be considered at different levels of generality, and at different counterfactual distance from the actual circumstances. Questions can be better or worse depending on the aptness of focus and foils, where aptness is determined partly by the phenomenon under consideration and the background purposes of the inquiry. How do we decide which equivalence relation to use when seeking the explanation? Garfinkel suggests:

> Clearly, there are some pragmatic, practical factors at work. Yet the situation is not completely determined by these factors, for these practical demands must be reconciled with the nature of the phenomena themselves and with the stability demands of good scientific explanation.
>
> So the answer to the question, Are irrelevance-geometries stipulated, or are they "in the world"? is: both! We can stipulate equivalences at will, but the result will be a good explanation or a good piece of science only if the way we are treating things as inessentially different corresponds to the way nature treats things as inessentially different. (Garfinkel 1981: 32)

Second, the options under consideration may be relationally constrained. In individualistic cases, the possibility space of options

consists in the Cartesian product of the possibility spaces of the particular individuals. For example, in a standard Prisoner's Dilemma, the options are either both prisoners defect, both do not, one does, or the other does. The possible outcomes are determined simply by what each prisoner does. But in non-individualistic cases, the options are constrained in ways that rule out certain possibilities. For example, suppose there is only one seat left at a movie we want to see. We can both forego the movie, or one or the other of us could go, but we cannot both go (assuming the background condition that the theater will only sell tickets for available seats, and we cannot get in without a ticket). In such cases, the structure of possibilities constrains what can happen, and this constraint is relevant to explaining what happens; it makes a difference. Again, the structure may be relevant at different levels of generality.

3 Sterken

In my paper "What is a (Social) Structural Explanation?" I consider a case in I which start by asking why there is a longstanding pattern of women being economically disadvantaged relative to men, and relatedly, why women more often than men quit their jobs when they become parents. I consider an example—intended as a paradigm case—in which it is rational for Lisa (rather than Larry) to quit her job when the infant Lulu enters the family, given she, like other women in the context, makes only 75% of what men make. I ask:

(8) Why did *Lisa* (rather than Larry) quit their job?

I suggest that an individualistic or psychological explanation in response to this question fails us in crucial ways. Suppose we answer (8) in this way[1]:

(9) Lisa doesn't like her job very much and really enjoys parenting Lulu.

Of course, this won't even begin to answer the question, i.e., to

[1] Note that this supposition (9) is not part of the example as stated, for part of what needs to be explained is why Lisa (and others like her) is likely to quit even if she would prefer to work.

provide difference makers, unless Lisa's attitudes—on these dimensions—differ from Larry's, i.e.,

(9') Lisa doesn't like her job as much as Larry likes his, and Lisa enjoys parenting Lulu more than Larry does.

This leaves the answer to the question at a wholly individualistic level: it is just the preferences of the individuals that count. There is a background structural assumption, though, that is relevant to the explanation: either Lisa or Larry must quit their job, i.e., their choices are relationally constrained. Otherwise, the mere fact of the differences stated in (9') don't explain why Lisa rather than Larry *quit*. Many couples are differently situated with respect to their job and parenting satisfaction, but they both work.

Garfinkel would have us articulate the background structural constraints as presuppositions of the question. What structural constraints should be included will depend on the actual circumstances. I suggested that the circumstances were these:

(10) Given that there is no affordable quality childcare in their region, that they have no local family support, that Larry makes more than Lisa and that this wage pattern is not likely to change, and that Lulu cannot be left on her own during the day, why did Lisa (rather than Larry) quit their job?

The question (10) makes explicit that Lisa and Larry's decision is relationally constrained, so it isn't just a matter of individual preferences as (9') suggests. In fact, it would seem that the explanation (9') is, at best, only a minor consideration. I included these background constraints because they are fairly common in the United States, given the lack of state supported childcare and parental leave time, dispersion of families, gender wage differentials, etc. If we frame the question as (10), however, one might argue that the answer can be given in straightforward rational choice terms:[2]

[2] I am not arguing here that it is, all things considered, rational for Lisa to quit. Given the bad consequences for women who become economically dependent on men (their exit options in bad marriages are significantly decreased), and the reinforcement of employer dispositions to treat women as less reliable, it may be more rational for Larry to quit. The rationality of Lisa's and Larry's choices will have to take a variety of further considerations into account. My point here,

(11) It is rational for Lisa (rather than Larry) to quit, given the constraints in the situation (such as differential income potential, lack of affordable childcare, and Lulu's need for care), along with their desires to maximize their income, support Lulu with good quality care, and not deal with various kinds of stress caused by cultural gender norms and the demands on couples balancing work-family responsibilities.

Note that although this would appear to be an individualistic explanation *about Lisa*, it is really a *rationalizing* explanation that offers a story about how the decision for Lisa to quit would be rational for *anyone in their situation*. Moreover, although it does not appear to be a *structural* explanation—what is offered as an explanans is the rationality of Lisa's decision—the rationality of the decision depends on the background constraints. It is exactly those constraints that not only link Lisa's and Larry's decisions, but also situate them differently in the circumstances. In other words, their positions within a socio-economic structure makes the crucial difference between Lisa and Larry (and those like them).

As I read Sterken, she emphasizes two points. First, that if we are to understand Lisa's choice, then we need to get into the details of her situation and consider carefully what background constraints she and Larry actually face. So, as she puts it, "in paradigmatic cases, a local and flexible structure, as opposed to a broad and deep structure, is the best object of explanation" (p. 183). By 'local and flexible structures,' she means structures that are more closely tied to local circumstances, and structures that are more contingent and less modally robust (p. 184–5). Second, it is generally more useful for political purposes to focus on local and flexible structures (p. 190).

Sterken grants that what counts as the best explanation will depend on the question we are asking. She contends, specifically, that if we want to know why Lisa quit, we should give some form of structural explanation, but the better explanation makes reference to local and flexible practices rather than broad structures. In particular, she contends that there is no reason to make reference to gender at all.

---

however, is just that even if we judge that Lisa makes a rational choice to quit, the rationality of the choice—and so the adequacy of the explanation—depends on background structural constraints.

I agree that there are cases where we want to explain why a woman quit her job that have nothing to do with gender, and where we can only plausibly offer a local and flexible explanation; it is also plausible that in some cases we want a very particular explanation of an individual's decision, e.g., why did Lisa quit *at 9:00 am* (v. 9:30 am) *today*. However, note that in setting up the Lisa/Larry/Lulu case, I was specifically trying to answer general questions about ongoing economic inequality along lines of gender and patterns of women (rather than men) quitting their jobs to take on full time childcare. The question at issue is explicitly a question about gender differences. So it is not surprising that I think gender is relevant.

One might argue, however, that in answering the question about why Lisa (v. Larry) quit their job, the question is not about gender. However, Lisa is not an actual particular woman. She is an example that I constructed (actually, that Susan Okin (1989) and Ann Cudd (2006) constructed) to represent a group of women who are structurally situated so that their options are materially and economically limited in a familiar way. So drawing on Garfinkel's language, the object of the explanation was an equivalence class that included only women similarly situated, where similarly situated involved the presuppositions articulated in (10). Why does adding a child to a (heterosexual) family in such circumstances tend to lead to women quitting their jobs rather than men. And I offer an explanation of the pattern that refers to the structural factors mentioned as background constraints. To suggest that not all women face those same structural constraints, or that some women find other ways to deal with the constraints does not undermine the adequacy of the explanation of the pattern that was at issue.

My discussion may be confusing because I sometimes speak of explaining Lisa's decision to quit her job. I acknowledge that I didn't do enough to emphasize that Lisa was functioning as an exemplar of a group (just as we might take Sparky to be an exemplar of Cockapoos in considering (4) and (5)). So the question arises how I would explain Lisa's decision, if we are supposing that we are considering her as an individual rather than as an exemplar. Suppose she is my friend and I am considering why she quit her job. It is plausible that I would want to consider her beliefs, preferences, and other attitudes.

In my paper, however, I distinguish structural explanations from

more typical causal explanations concerned with events. Sterken says that "what [Lisa's] family and the members of her family do is in large part determined by the decisions and deliberations of the family system" (p. 187). *Of course* that is true, if we are looking for the event that caused Lisa to quit. Lisa decided to quit (perhaps after deliberations with Larry), and acted on that decision. No one denies this, e.g., Garfinkel never claims that the event of Garfinkel assigning an A to Mary is caused by anything other than his decision to assign Mary an A. The claim that Lisa quit because she decided to is true, but answers the question: What event caused Lisa to quit her job? This is not the question under consideration.

From what I can tell, however, Sterken recognizes that there is a question other than the particular causal question that calls for a structural explanation, but maintains that it should be answered in terms of more local and flexible structures, rather than broad and deep ones. We should look *not* at gender wage gap, the systematic lack of childcare opportunities, and the tendency for families to disperse, but at Lisa's employer and other more immediate institutional factors that limit her. It is a problem that *her employer* does not offer subsidized parental leave. It is a problem that there isn't a parenting co-op *in her neighborhood*. These are what really explain why she quit her job and are also sites where we can push for change.

I don't deny that there are local structures that explain some of the patterns, and perhaps do so more effectively. As I mentioned above, however, I was considering a kind of case in which the employment and childcare constraints are widespread and not due to particular bosses or neighbors. If we change the pattern in question and the problem really is the company Lisa works for and other local constraints, then it makes sense to put pressure on the company to promote paid parental leave. More generally, I agree that local activism is tremendously important. But I also think that whether this is a good political strategy will depend a lot on the broader structures. Companies are often unwilling to provide benefits to their employees because they feel pressure to be competitive; in the case of parenting policies, coverage for parental leave is costly and difficult to manage (because it requires a reserve pool of skilled employees that can step in for six months). Often the only way to promote change under capitalism is to rely on regulation that insures that particular

companies aren't bearing the burden of doing the right thing and so can remain competitive. Seeing that there is a broad and deep effect of the lack of parenting leave on women's opportunities can be crucial for legislative reform. The broad and deep pattern may need a broad and deep solution. There is (and has long been) a broad and deep pattern in the United States; the pattern in other countries deserves its own discussion. The goal of my paper was never to generalize about what sorts of structural explanations we need (local/broad, flexible/deep), for that depends entirely on the phenomenon we are attempting to understand.

In short, I don't deny that there are many different kinds of case: our questions can vary, the background conditions can vary, the degree of specificity of the pattern can vary, the strategic implications can vary. Sterken's paper demonstrates effectively that my discussion did not adequately highlight the range of relevant variations. Broad and deep structural explanations (whether in terms of gender or not) should not be all we seek. But neither should we focus solely or primarily on local and flexible structural explanations. Where there are broad and deep patterns, or local and flexible ones, we need to understand them and address them politically.

4 Saul

In my paper, "Social Structure, Narrative, and Explanation" (2015), I argue that implicit bias stories are overly individualistic and do not provide an adequate basis to promote social justice. On my account, injustice is largely a structural problem, and should be addressed on a structural level. In order to achieve justice, we need new social and economic practices, new policies, new institutions and laws, a redistribution of resources (things of +/– value), and changes in social meanings. These different elements of social life form a homeostatic system that tends to correct itself unless there are broad and multi-pronged challenges to it (Haslanger 2017); bad attitudes (even implicit ones) are not the linchpin that maintains such practices, policies, and distributions. Moreover, because individuals depend crucially on coordination in order to manage life in complex societies, changing minds, without also changing the social conditions of coordination, will be hard to accomplish and very hard to maintain.

As long as unjust practices structure our social environment, we will be drawn into them and fluently participate in them, like it or not.

The target of my paper was implicit bias *explanations* of injustice. I was interested in a certain form of implicit bias explanation which presumes explanatory individualism. Roughly, explanatory individualism is the view that explanations in the social domain should (ultimately) be stated in terms of individuals and their properties. Explanatory individualism is associated with projects in the social sciences that demand "microfoundations" of social phenomena. (See Epstein 2009: 188; Jackson and Pettit 1992.) Structural explanations are mere placeholders for the "real" explanation at the lower level. I relied on Charles Tilly's (2002) critique of "standard stories" as a way to challenge a background individualism that I found in much—but not all—discussion of implicit bias. It was not my intention to argue that psychological states of individuals have nothing to do with injustice. Rather, I was trying to highlight factors that create and sustain injustice that individualistic explanations tend to ignore. As Tilly points out:

> …whatever else we have learned about inequality, social scientists have made clear that a great deal of social inequality results from indirect, unintended, collective, and environmentally mediated effects that fit very badly into standard stories. (Tilly 2002: 28)

Saul and I agree that existing injustice is not simply the result of bad attitudes and structural change is needed to achieve social justice. She argues, however, that I have misrepresented implicit bias stories—and broader implicit bias interventions—as overly individualistic. In particular, she demonstrates how, in the best cases, such stories provide a crucial link between individual attitudes and structures. On her view,

> Implicit bias training, then, forces one to make a move from the individual to the structural, and this is to my mind one of its great strengths. It can provide the bridge that helps people to move—perhaps without even seeing that this is happening—away from the individualistic view of the world with which they start. (p. 230–1)

She offers some criteria for judging when and how implicit bias training can accomplish this (p. 241):

(1) The story situates implicit bias as a result of and contributor to broader structural injustice, and does not underrate the importance of combatting structural injustice.

(2) The story is one on which seeking progress toward social justice is possible.

(3) The story is one on which seeking progress toward social justice is desirable.

(4) The story motivates action (collective or individual) toward social justice.

(5) The story offers a road-map for such action.

I find Saul's arguments compelling and am happy to grant that there are excellent and valuable forms of implicit bias training that include "standard" or "nouveau" stories only as one part of a broader explanation of injustice. I also agree that narratives are crucial for motivating individuals to take action. Because Saul clearly rejects explanatory individualism—which was my main target—I see us as allies in the effort to promote social justice. My sense is that our disagreements are mainly a matter of emphasis: We are looking at different parts of big systems.

It might be useful, however, to think further about how individuals, individual bias, and individual responsibility/accountability fits into the big picture. For example, Robin Zheng (2018) has recently suggested that my paper not only poses challenges for explanatory individualism; it also takes aim at normative individualism, i.e., the view that the source of moral wrong and site of moral responsibility lies in the actions and attitudes of individuals. Zheng argues that my move away from normative individualism is a mistake; more specifically, a conception of moral responsibility—as (individual) *accountability*—is necessary in order to undertake social change at all. Responsibility-as-accountability is forward-looking, rather than backward-looking, and calls upon us to act in ways that address injustice.

Both Zheng and Saul (e.g., p. 233) emphasize the importance of accountability to motivate people to behave more justly, and by doing so they further illuminate the bridge between individual and

structural approaches to injustice. If we are to promote social justice, certainly we must address issues of moral motivation and moral agency that occur at the level of individuals. As Zheng puts it:

> …theories of moral responsibility provide guidance in cases where others fail to take up their share of the collective burden or make mistakes in doing so—both of which are unavoidable on the long hard road to justice. While the work of structural change is collective, it is always particular individuals, their actions and attitudes, that we confront in the classroom, in the meeting hall, and on the streets. A theory of individual moral responsibility for structural injustice thus takes seriously the interpersonal relationships between persons that are key to the actual day-to-day work of contestation, organization, and activism (Zheng 2018: 10).

However, as I understand the theoretical commitment to individualism (ontological, explanatory, or normative), the claim is that the *real* work in the domain in question happens at the individual level, not just that some of it does. Ontological individualism maintains that the social is wholly reducible to, supervenes on, or is grounded in (depending on the author and the period of research) individuals and their attitudes; explanatory individualism claims that adequate explanations of social phenomena should (ultimately) be stated in terms of individuals and their attitudes; and normative individualism is the view that the fundamental site of normativity—the source of badness or wrongness facts—lies in individuals and their attitudes. All of these forms of individualism allow that there are derivative facts, explanations, and evaluations, at higher levels. For example, even an ontological individualist could allow that there are clubs and committees (made up of individuals); likewise, an explanatory individualist could allow that an increase in unemployment explains an increase in crime (as long as this can be spelled out in terms of the effects of unemployment on the dispositions of individuals to commit crimes). Normative individualists can also maintain that there are normative facts that concern social or collective phenomena, e.g., a normative individualist could allow that there are bad birthday parties, but the badness of a party would be explicated in terms of, say, the unhappiness of the participants. Similarly, a group (or institution, or corporation) might be morally responsible for a state of affairs, but the group's moral responsibility would be explicable in

terms of the moral responsibility of its members.

Although I suggested in my paper that I object to a normative *focus* on individuals, I didn't intend to suggest that we should reject moral claims on individuals, or that the fundamental normative facts are structural. I do not have a view about the sources of normativity. I am inclined to reject normative individualism, but I do not have an alternative account of normativity. My intention was to call attention to the tendency within ethics to frame moral questions as concerned with what *I*, as an individual, should do, what *I,* as an individual am morally responsible for*,* and whether *I*, as an individual, have moral worth. These are legitimate and important questions, but they aren't the only ones or even the most important ones we should be considering when we want to address injustice.[3]

Injustice is a distributive and relational matter; it concerns our collective arrangements and the structure of our social lives, not just what I do to you, or owe to you, and you to me. We are looking for a system of laws, policies, norms, and practices that facilitate a morally acceptable form of cooperation. As a consequence, responsibility (both backward-looking and forward-looking) may need to attend to groups, i.e., those who occupy a particular social position, and to institutions. Capitalists (and consumers) may bear a special collective responsibility for the exploitation of workers, white people may bear a special collective responsibility for certain racist phenomena, (etc.); collective action may be needed to discharge the responsibility. In such cases, what *I* should do as a member of the group or institution is not always the sort of thing that can be determined unilaterally, and it is likely that the distribution of responsibility across a collective will depend enormously on details of history, circumstances, ability, and opportunity. For example, a corporation may bear responsibility for polluting a river; if so, plausibly the responsibility of members of the corporation derives from their role in the corporation. (Were you the owner? The environmental protection officer? The janitor?) This suggests that *at least in some cases*, the responsibility of a collective/group for a moral wrong is not just derived from the responsibility

[3] I do not think we need a theory of the sources of normativity to address injustice—at least the gross forms of injustice that we are currently living with. But we do need to look past what an individual can do on her own.

of its members. Rather, the responsibility of the members—both causally and in terms of accountability—is derived from the institutional responsibility and the individual's role in the institution. Once we have determined who is responsible and for what, then we need to educate and motivate people to fulfill their responsibility. This is hard and important work that I did not sufficiently emphasize. But it presupposes that the primary explanatory and normative work has been done, so it would be wrong to think of individual responsibility as always *more* important or *more* fundamental.

The different emphases in our discussion may become clearer if we distinguish several sorts of questions that are relevant for promoting social justice (as before, I will focus on the US context for examples):

I Questions about social explanation:

(a) Why are men paid more than women for the same work? Why is women's work devalued?

(b) Why are unarmed Black men more often confronted with deadly force than unarmed White men?

(c) Why is there a disproportionate representation of LGBTQ individuals among homeless youth?

(d) Why is there a racial achievement gap in educational attainment?

(e) Why are recent Latinx immigrants exploited and socially marginalized?

II Questions about normative/moral evaluation and responsibility:

(a) Where does the moral wrong of racism, sexism, homophobia, transphobia, ethnocentrism, and economic exploitation lie?

(b) Who is responsible for the devaluation of women's work, the racial achievement gap, the exploitation of immigrants, the violence against LGBTQ individuals, women, and members of non-white racial groups?

(c) What are permissible methods for changing (blaming, calling out, holding accountable) people's behavior?

III Questions about human psychology:

(a) How are humans recruited into social practices? What cognitive mechanisms enable them to become fluent in them? What disrupts such fluency?

(b) Why do humans willingly engage in social practices that are at odds with their explicit values and moral principles? Why are humans invested in social practices that do not promote their self-interest?

(c) How do people's minds, attitudes, emotional responses, perceptual frames, etc. change? (Does implicit bias training actually work?)

(d) How do individuals develop oppositional consciousness? How do we motivate individuals to engage in efforts to promote justice (even if it isn't in their immediate self-interest)? (Does implicit bias training enhance moral motivation?

(e) Is moral motivation sustainable under conditions of broad and deep injustice? What promotes longstanding willingness to exercise agency in pursuit of justice?)

IV Questions about policy:

(a) What laws or policies might be implemented to promote greater justice, i.e., to prevent patterns of exploitation, marginalization, systematic violence, powerlessness, and ethnocentrism against subordinated groups? (On these faces of oppression, see Young 1990.)

(b) What policies and practices are effective in motivating and recruiting individuals to be more engaged in working for social justice?

(c) When and how are social movements effective?

In my paper, I was focused on questions about social explanation and moral evaluation. If, as Tilly suggests, some patterns of injustice are "indirect, unintended, collective, and environmentally mediated," then an explanatory focus on human agency and a normative focus on backward-looking moral responsibility is not always helpful. As Zheng argues, however, that is compatible with holding individuals morally accountable for rectifying injustices that they (alone) did not cause. And if we are going to hold people accountable, then, as Saul argues, we should be engaged in practices that motivate and train them to change themselves and their communities accordingly.

This seems right to me, as far as it goes. And I support (and even lead) implicit bias training of the sort Saul recommends. However, I am not optimistic that a focus on implicit bias makes a substantial difference on broad and deep patterns of injustice, and I am more committed to and have more hope for radical forms of consciousness raising and grass roots organizing, strategic intervention of elites (including nudges and material changes), direct action, and cultural intervention. Both Zheng and Saul are right that implicit bias training can be motivating and that structural explanations can be demotivating. But there is no need, I think, to take an either-or stance. Where there is evidence that implicit bias training works to promote greater justice, let's use it. But at the same time, we should recognize that moral motivation tends to be weak in the face other motivations, e.g., for cooperation, economic stability, social status, and identity affirmation. Changing material and cultural conditions so that moral motivation converges with these others strikes me as a strategy worth pursuing.

5 Díaz León

Esa Díaz León's paper considers the metaphysical commitments of my approach to gender and race and argues that, contrary to Elizabeth Barnes' interpretation (2014, 2017), my view is compatible with metaphysical deflationism. This is a tricky question, for I have said different things at different times about what I was aiming to achieve. In my early work (Haslanger 2000), I suggested I was aim-

ing to give an analytical (or ameliorative) account of our concept of race; in later work (Haslanger 2006, 2010), I suggested that an ameliorative account might be achieved by undertaking a descriptive analysis that aimed to identify the social kind our term in fact picks out. The shift in terminology from 'analytic' to 'ameliorative' was itself a potential source of confusion, and the shift from characterizing my view as ameliorative to descriptive (or both) was also confusing. Whether or how a descriptive approach can be ameliorative is an open question. Moreover, given background disagreements and unclarity (both in my work and the work of others) about the use of terms like 'concept,' 'kind,' and 'meaning' it is no wonder that there are different interpretations of my view.

There seem to be three questions that Díaz León takes to be potentially at issue between the deflationist and the realist such as Barnes:

(1) Do the terms 'gender' and 'race' carve the world at the joints?

According to Barnes, the realist says 'yes,' and the deflationist says 'no.' As Barnes represents the deflationist, there are no (intrinsically?) privileged joints in nature.

According to Díaz León, however, there are privileged ways of carving the world, relative to our purposes. If we want to know what gender is, then the privileged joint, i.e., the joint that best answers the question, will be a social joint rather than a biological one. There are privileged joints, but privileging happens relative to us.

(2) Are our discussions about gender and race, primarily, discussions about our concepts or about the world?

According to Barnes, the realist says that the debate is about what there is in the world, and the deflationist says that the debate is about what concept(s) we should use.

According to Díaz Leon, the deflationist is talking both about our concepts and what is in the world. On her view, we need to craft our concepts to capture what is important. In the case of gender, say, we explicate the concept of *gender* to capture what in the world we need to be talking about—given our interests or purposes, some objective facts are the proper object of our attention; and the facts may contribute to our understanding of what reasonable purposes we might have. (See Anderson 1995.) Once we've sorted out our interests and

the relevant parts of the world, speech using the concept in question is true or false depending on whether the concept applies.

(3) What explains why trans women are women?

According to Barnes, the realist view is that trans women are women because they are members of the kind women. This kind is not "up to us." But, she suggests, the deflationist must say that trans women are women because it is politically valuable to have a concept of women that includes them. The worry is that this does not give us an ontologically adequate explanation of why trans women are women.

According to Díaz León, we cannot say who counts as a member of the kind women without knowing what the term 'woman' means, i.e., what concept it expresses. So although we do have to determine the content of the concept 'woman' before determining who counts as a member, there is no alternative. How, she suggests, could there be a different order of explanation?

Díaz León suggests several times that she does not take her debate with Barnes to be a question of interpretation of my text. Rather, she maintains that the goal of the paper is to show that the deflationist position does not have the bad consequences that Barnes suggests: ontological deflationism about gender is a viable position and compatible with social constructionism about gender and race. In short, there is a meaningful position between ontological realism and ontological anti-realism.

I agree with Díaz León that ontological deflationism is a meaningful position, and to that extent, I agree with her arguments. However, I would like to suggest that there is another middle ground between an especially strong version of ontological realism and an anti-realism that does not take the form of Díaz León's (or Amie Thomasson's) deflationism.

To locate this additional position, it is important to draw a distinction between semantics, metasemantics, and ontology. Roughly, semantics considers what a term means. Metasemantics considers how a term gets the meaning it has. And ontology considers what there is in the world. Díaz León (and Thomasson (2015)), include in their semantics what some would locate in metasemantics. So there are really two disagreements between Barnes and Díaz León: one is about semantics, another about ontology.

Internalists about semantics will often employ and recommend a semantic strategy (Mallon 2006: 527) for doing ontology: start by figuring out what our terms mean (considered to be the application conditions) and through this figure out what there is by determining what satisfies the application conditions.[4] This methodology assumes a kind of neo-Fregean approach to meaning: words have senses that determine their referents. In order to say anything, we need a concept associated with a word that consists in a set of (application- or truth-) conditions; parts of the world become subject matter for our discourse by virtue of satisfying the conditions specified by the concept.

But this theory of meaning is only one contender for an adequate semantics. I reject internalism about meaning, and instead prefer to think of meanings in terms of informational content (see Haslanger 2010; Haslanger forthcoming). Following Stalnaker, we express, believe, suppose, (etc.) propositions, understood as "functions from possible circumstances to truth values, or equivalently, as sets of possible situations." (Stalnaker 1998: 343). Word meanings can be captured by sets of possible individuals or divisions in logical space; concepts are capacities we have to make distinctions (Yalcin 2016, Perez Carballo 2016). On this approach, it would not make sense to say that we have to sort out what the word means *before* determining its extension because the extension (across possible worlds) *is* its meaning. We use words without having a full grasp of what they mean, and it will often take empirical work to determine their meaning. To do metasemantics—to determine how the word has that meaning—we need to figure out the pattern of our distinction-drawing behavior, the purposes of that behavior, the bits of the world that motivate and make sense of that behavior, etc.

Because I embrace an externalist semantics, I disagree with Díaz León about the answers she offers to questions (2) and (3) above. There are several different sites for debates about gender. First, we might debate about: who is, or is not, a woman or man; whether some people are both women and men; whether some people are neither women nor men; whether one's being a woman or man depends on context, etc. On my view, these are not semantic or metasemantic

[4] In the context of philosophy of race, Glasgow (2009) adopts this strategy, Mallon (2006, e.g., 550) rejects it.

questions. In attempting to answer them we are theorizing about the world. We should, I believe, draw on biological, historical, anthropological, sociological, psychological, and normative inquiry (including feminist theory and queer theory) to answer the questions. Just as a biologist or medical researcher doesn't consider what the term 'pathogen' *means* to determine whether a microbe is a pathogen, gender theorists don't need to consult what 'woman' means to determine whether someone is a woman. (I use discussions of pathogens as my example because what counts as a pathogen is clearly interest-relative and involves background normative presuppositions.) The focus on just theorizing about the world, rather than worrying about "analytic" constraints on theorizing, is exactly what scientific essentialists emphasized in the 1970s (Kripke 1980, Putnam 1975).

Second, in theorizing about gender, we might debate metasemantics: what determines whether this or that carving of logical space constitutes the extension of the terms 'man' or 'woman' or 'gender queer.' There are many carvings of logical space that position different individuals on different sides. People use the terms in many different ways. Who counts as an authority on this? Should we endorse a social externalism (Burge 1979)? Method in metasemantics, is difficult and unclear. What makes it the case that our words and concepts have the content they do, and how do we have knowledge of their content? Past usage and explanatory purposes play a role (Schroeter and Schroeter 2015). But, at least in my view, no one has an adequate theory of this (and this is why the project of conceptual engineering/amelioration is so contested).

A possible advantage of the externalist approach, however, is that we can consider the question (3) as more direct, i.e., as a first-order or substantive question rather than a semantical one. Once theory yields its results, we can then embrace Barnes' claim that:

(12) 'Trans women are women' is true because trans women are women.

rather than:

(13) 'Trans women are women' is true because we have selected a set of application conditions for the concept 'woman' to include them and they satisfy the conditions.

I'm not altogether convinced that (12) is so much better than (13). But comparing the externalist and internalist semantics helps locate the disagreement. Díaz León claims, in responding to Barnes' explanatory priority complaint, that it is not intelligible to say trans women are women *just because* they are women, because *of course* the truth of 'Trans women are women' depends on the meaning of the term 'woman.' But because the internalist and externalist disagree about meaning, they disagree about what it is for truth to depend on meaning. For the internalist, the meaning of 'woman' is a set of conditions we associate with a term; for the externalist the meaning of 'woman' just is the set of (possible) women. So even if in both cases the truth depends on meaning, the site of our contribution differs. The externalist will allow that there is some sort of metasemantical dependence on us, but this does not displace the work of explaining the *truth* of the claim from the extension onto us.

I would suggest, however, that the debate between Barnes and Díaz León on explanatory priority would be resolved more fruitfully by returning to Garfinkel's suggestion that we need to look more carefully at what questions we ask: are Barnes and Díaz León answering the same question differently or answering different questions? As I read them, they are asking different questions:

(14) Is 'Trans women are women' true (v. false) because trans women are women?

(15) Does 'Trans women are women' have its propositional content (v. a different propositional content) because trans women are women?

Barnes' point is that in order for the internalist to answer (14), we have to consider our "concept of women" or "what 'woman' means" and then consider how or whether trans women satisfy the conditions. The externalist might require that we consult with theorists, but not semanticists. But in both cases, the truth of the claim depends—at least in part—on the fact that trans women are women. Their answers to (15) will also differ given their accounts of meaning, i.e., what counts as the propositional content of the sentence, and what determines the propositional content, as I've sketched.

Earlier I suggested that my own approach to the ontology of

gender and race would fall between a realist and anti-realist account. So far I've distanced myself from Díaz León's deflationism. So in what sense am I less than fully realist? Realism, of course, is difficult to characterize. In saying that I'm less than fully realist, I only mean to distance myself from a particular way of privileging properties or distinctions by reference to naturalness, or natural kinds, or fundamentality. I do not accept the idea that there are different degrees of reality that correspond to such distinctions. I believe that reality is binary: either you exist or you don't; either you are real or you aren't. There are epistemically and pragmatically interesting differences between real things: some are more explanatorily useful, more accessible, or more valuable. As indicated earlier, I believe that explanatory power depends on the question being asked. The distinction between hydrogen and non-hydrogen is important for explaining some chemical reactions, but it is useless when attempting to explain why Lisa (v. Larry) quit her job, or why women (v. men) tend to quit their jobs, or why I usually eat with a fork (rather than chopsticks). On my view, the natural world is neither explanatorily prior to nor ontologically more fundamental than the social world (even if the social globally supervenes on the natural). (See Epstein 2015.) But as I read Barnes, part of her overall point is to argue that neither naturalism nor fundamentality is the key concept for realism, or the project of metaphysics. I suspect that on this Barnes, Díaz León, and I can agree.

5 Conclusion

Again, I thank Esa Díaz León, Jennifer Saul, and Rachel Sterken, for their insightful and valuable commentaries. They raise many issues and arguments that I haven't addressed here. I have used the issue of explanation as a thread weaving through the discussions. As sketched above, I accept an erotetic account of explanation that takes explanations as answers to (certain kinds of) questions. I wholeheartedly agree with Garfinkel that what might appear to be theoretical or explanatory disagreements arise when parties to a debate are really trying to answer different questions, and much can be gained in understanding by clarifying the question at issue. As he suggests,

> What is needed is a critical philosophy of explanation. Its point would be to give us an understanding of what the objects of explanations are, what we want them to be, what forms of explanation are appropriate to those objects, and how various explanations fit together, excluding or requiring one another, supplanting one another historically, presupposing one another. (Garfinkel 1981: 14)

Garfinkel's work started on that task, and others have taken it up (e.g., Risjord 2000). I hope that my discussion here has mainly given us further motivation to continue it.

In response to Sterken, I have argued that although it might be that attention to local and flexible social structures are important to provide explanation of some phenomena, there are also context and questions that demand broader and deeper structures. Gender is a deep structure that explains many more superficial phenomena (Witt 2011). But as Sewell suggests, "Rather than staying at the deep structural level preferred by Levi-Strauss, I think we should, like most anthropologists, think of rules [structures, practices] as existing at various levels. [Those] nearer the surface may by definition be more "superficial," but they are not necessarily less important in their implications for social life" (Sewell 1992: 7). We should keep in mind also that both deep and superficial structures are sites for the circulation of power; deep ones, however, are often more unconscious and pervasive. I think Sterken and I can agree that what level of structure is apt as explanans depends on our explanatory purposes and the questions we ask.

Saul argued that both individual phenomena and structural phenomena are important for understanding persistent and resilient forms of injustice. I agree. I view systems of injustice as homeostatic systems that are held in place by many interdependent forces (Haslanger 2017); the forces include psychological attitudes (implicit and explicit), cultural schemas and social meanings, material conditions (including geography, biological constraints on human survival, technology), laws, etc. When we are trying to destabilize unjust systems, the interdependence of these forces matters a lot. I think Saul and I agree on this and that a purely individualistic approach (relying wholly on variations of "standard stories") is problematic. Attention to the parts, the whole, and the relations between them, will be required to make progress, and explanatory individualism doesn't have

the resources to address the structural and systematic dimensions of the problem. Different questions—or different tasks—drive us as we proceed, but, as I see it, Saul's and my projects are compatible. Where we might differ, however, is on the question of responsibility and accountability. I think we need more attention to collective responsibility and methods for crafting and motivating collective action. But I would grant that much of this work must start with individuals.

In considering Díaz León's commentary, I turned from social explanation to metaphysical explanation. I argued that debates over realism and deflationism about social kinds might be separated into questions about semantics and metasemantics, and questions about ontological structure. As I read Díaz León's response to Barnes, their disagreements are more semantic than ontological; but this is not to say that we don't need metaphysical explanation. In fact, I think we might all agree that the fact that trans women are women provides a kind of metaphysical explanation of why it would be both incorrect and morally/politically wrong of us not to count them as women.

But I think a deep issue remains: when doing ontology, is the project to limn the ultimate structure of reality (Quine 1960: 161), or is it to answer our questions as we struggle to figure out how best to go on? Quine thought that these two projects were not wholly distinct, and I agree with him (though I don't share his view of explanation). If this locates me as less than fully realist by current standards, then so be it. Philosophically I am committed to answering questions that arise in the struggle for social justice and I take all those engaged in the symposium to be so too. Our debates in this symposium (and others) demonstrate that sometimes rather specialized and esoteric tools are useful for understanding the social domain; and, more importantly, that tools we develop in the context of critical social theory can change the questions we ask, what forms of explanation are called for, and how we do philosophy.[5]

Sally Haslanger
Massachusetts Institute of Technology
shaslang@mit.edu

[5] Thanks very much to the participants in the 2016 NOMOS workshop, especially Teresa Marques for her insight, understanding, and patience. Thanks also to Shannon Doberneck, Kevin Dorst, Abigail Jaques, Elis Miller, Brad Skow, and Ege Yumusak for helpful discussion.

References

Anderson, Elizabeth S. 1995. Knowledge, human interests, and objectivity in feminist epistemology. *Philosophical Topics* 23(2): 27–58.
Cudd, Ann. 2006. *Analyzing Oppression*. Oxford University Press.
Barnes, Elizabeth. 2014. Going beyond the fundamental: feminism in contemporary metaphysics. *Proceedings of the Aristotelian Society* 104(3): 335–51.
Barnes, Elizabeth. 2017. Realism and social structure. *Philosophical Studies* 174(10): 2417–33. DOI: 10.1007/s11098-016-0743-y
Burge, Tyler. 1979. Individualism and the mental. *Midwest Studies in Philosophy* 4 (1): 73–122.
Epstein, Brian. 2009. Ontological individualism reconsidered. *Synthese* 166(1): 187–213.
Epstein, Brian. 2015. *The Ant Trap*. Oxford: Oxford University Press.
Garfinkel, Alan. 1981. *Forms of Explanation: Rethinking the Questions in Social Theory.* New Haven: Yale University Press.
Glasgow, Joshua. 2009. *A Theory of Race.* New York: Routledge.
Haslanger, Sally. 2000 Gender, race: (what) are they? (What) do we want them to be? *Noûs* 34(1): 31–55.
Haslanger, Sally. 2006. What good are our intuitions: philosophical analysis and social kinds. *Proceedings of the Aristotelian Society Supplementary Volume* 80(1): 89–118.
Haslanger, Sally. 2010. Language, politics and 'the folk': looking for the Meaning of 'Race'. *The Monist* 93(2): 169–87.
Haslanger, Sally. 2016. What is a (social) structural explanation? *Philosophical Studies* 173: 113–30.
Haslanger, Sally. Forthcoming. Going on, not in the same way. In *Conceptual Ethics and Conceptual Engineering*, ed. by Herman Cappelen and David Plunkett. Oxford: Oxford University Press.
Haslanger, Sally. 2017. Racism, ideology, and social movements. *Res Philosophica* 94(1): 1–22.
Jackson, Frank; and Philip Pettit. 1992. Structural explanation in social theory. In *Reduction, Explanation and Realism*, ed. by David Charles and Kathleen Lennon. Oxford: Oxford University Press.
Kripke, Saul. 1980. *Naming and Necessity*. Cambridge, MA: Harvard University Press.
Mallon, Ron. 2006. 'Race': normative, not metaphysical or semantic. *Ethics* 116(3): 525–51.
Okin, Susan. 1989. *Justice, Gender and the Family*. NY: Basic Books.
Pérez Carballo, Alejandro. 2016. Structuring logical space. *Philosophy and Phenomenological Research* 92(2): 460–91.
Quine, W. V. O. 1960. *Word and Object*. Cambridge, MA: Harvard University Press.
Putnam, Hilary. 1975. The meaning of 'meaning'. In *Mind, Language and Reality. Vol. 2 of Philosophical Papers*. Cambridge: Cambridge University Press.

Risjord, Mark W. 2000. *Woodcutters and Witchcraft*. Albany: SUNY Press.
Schroeter, Laura; and Francois Schroeter. 2015. Rationalizing self-interpretation. In *The Palgrave Handbook of Philosophical Methods*, ed. by Chris Daly. Basingstoke: Palgrave-Macmillan, pp. 419–47.
Stalnaker, Robert. 1998. What might nonconceptual content be? *Philosophical Issues* 9: 339–52.
Sewell, William. 1992. A theory of structure: duality, agency, and transformation. *American Journal of Sociology* 98(1): 1–29.
Skow, Bradford. 2018. *Causation, Explanation, and the Metaphysics of Aspect*. Oxford: Oxford University Press.
Thomasson, A. 2015. *Ontology Made Easy*. Oxford University Press.
Tilly, Charles. 2002. The trouble with stories. In *Stories, Identities and Political Change*. Lanham, MD: Rowman and Littlefield, pp. 25–42.
Witt, Charlotte. 2011. *The Metaphysics of Gender.* Oxford: Oxford University Press.
Yalcin, Seth. 2016. Belief as question sensitive. *Philosophy and Phenomenological Research* 97(1): 23–47.
Zheng, Robin. 2018. Bias, structure, and injustice: a reply to Haslanger. *Feminist Philosophy Quarterly* 4(1): Article 4. doi:10.5206/fpq/2018.1.4.