Comments on Alex Madva's "A plea for anti-anti-individualism: how oversimple psychology mislead social policy"

Saray Ayala – *Sacramento State University*

Let me start by saying that I'm grateful and very happy for the opportunity to be part of this symposium. First, because I admire the The Brains Blog's initiatives, which facilitate interesting discussions and allow participation from a wide range of people without the need to travel. Second, because I admire Alex Madva's work and it's a pleasure to comment on it. In line with other papers by Madva, this is a rich and exciting piece of philosophy, which I highly enjoyed.

Madva's paper presents a well-woven and empirically informed argument against the idea that in order to alleviate inequality and discrimination, we need to start by changing structural and institutional dynamics, as opposed to interventions on individuals' minds. He labels this position "structural prioritizing". I agree with Madva's prescription that an integrative approach to social change, encompassing both structural and individual-level interventions, is more likely to succeed. I also agree with his call for a more fine-grained consideration of both types of interventions, and a detailed analysis of the reasons why specific interventions at the individual (or structural) level might or might not be successful. Instead of generalizing about the goodness or badness of structural or individual-level interventions, we need to explore specific interventions, for their effectiveness might have more to do with its specific character, rather than its being at this or that level. Failure or success of any one intervention at one or the other level is not indicative of the appropriateness of the level and thus does not justify any conclusion about structural or individual-level interventions in general. For example, if a particular intervention to reduce implicit bias against a particular social group does not work, it's unfair and misleading to conclude that any intervention at the individual level will meet the same fate, that the individual (as opposed to the social) is the wrong locus of change. It's unfair because as Madva notices, no intervention in isolation should be expected to solve all the problems. And it's misleading because implicit bias is not a unified, simple phenomenon, but heterogeneous (Holroyd & Sweetman 2016); we should expect complex interactions between biases, and biases with different functionalities (and perhaps different underneath mechanisms), therefore failure associated with addressing one sort of bias is not completely informative about how intervention on other biases will go. It's misleading also because attempts to reduce individuals' implicit bias is just one type of individual-level intervention.

In relation to this last point, Madva's work calls attention to an interesting question that is often either absent or not explicit in the contemporary debate on how to address discrimination and inequality: there is a variety of attitudes that are relevant for social justice, upon which intervention could have a bigger or a smaller (positive or negative) impact. Implicit bias and prejudices are just one type amongst them. Without making a taxonomy of social justice-related attitudes, a previous version of this paper

was more explicit about this variation, but the present paper is still clear enough about it. According to Madva's diagnosis, structural prioritizers are stuck in one kind of attitudes: prejudices and implicit biases against social groups. And they conclude that we should focus on structures after reasoning that intervention on prejudices and implicit biases is either not enough or not effective at all. But Madva warns: this is a non sequitur. Intervention on other kinds of attitudes could be (more) effective, and so the reputation of individual-level interventions rescued. One such candidate attitudes are those towards the malleability of social systems; as Madva points out, a study by Johnson and Fujita (2012) suggests that making malleability of a social system salient can increase individuals' motivation to change it. This intervention towards social change would count as an intervention at the individual-level, but not on implicit bias against any particular group. Thus, from the failure of intervention on implicit bias does not follow a general claim about the inefficiency of intervening on individual minds. Structural prioritizing, if based upon this reasoning, is misleading.

Instead of commenting on Madva's general conclusion on what successful interventions should include, I'll focus on how he makes his way towards such a conclusion. But before scrutinizing Madva's critique of structural prioritizers, let's first see how fair his characterization of them is. We could say that Madva's paper does not provide a contextualization of structural prioritizers' claims and this contributes to paint an unfair picture of them. A charitable interpretation of prioritizers should take into account that they are responding to a particular move, which we could call implicit-bias priorization: the priorization of intervention on implicit bias in order to address discrimination and inequality. And this motivation is well justified, for there is tremendous attention to implicit bias and a lot of literature (and it keeps increasing) on the topic (on measuring them, characterizing them, and alleviating them). If taken as a response to implicit-bias priorization, Madva's target reveals as more nuanced and less out-of-synch as it might seem. Reading structural prioritizers without that context is unfair, or at least only partially fair, for it takes out of the picture the strong motivation behind their concerns. Even if nobody had explicitly articulated a position such as implicit-bias priorization, the public attention (within different academic disciplines, and also outside of academia) that implicit bias is receiving justifies the warnings that structural prioritizers are putting forward: that *fixing* biased minds will not alone solve social injustice. That is, we could see structural prioritizers as reacting against the idea (tacit or insinuated, even if not explicitly articulated or defended) that individual-level change is sufficient for social change. If so, they could be taken to defend something like "structural necessarism": structural change is necessary (even if not sufficient). Although this reading adds nuance and complexity to the discussion, it is true that some of the claims Madva analyzes deserve the more radical reading he applies, for there is definitely a stronger position, something we could call structural sufficientism: that structural change is sufficient for social change. Madva's paper does a great job exploring and criticizing this more radical position.

Putting now aside questions about what particular positions are actually being defended and how, let's get into the more philosophically interesting project of

exploring the general conceptual landscape. This exploration would serve to make a few comments on Madva's critique.

Causes and Interventions

Madva's argument against structural prioritizers focuses on one question: the effectiveness of interventions in alleviating discrimination and inequality; we'll call this question EFFECTIVENESS. Besides EFFECTIVENESS, there is at least one other question that appear here and there in Madva's argument: a question about the causes of inequality and discrimination (CAUSE) (another question, about explanations of inequality and discrimination, is also relevant to Madva's argument, but given space limitations I won't analyze it here). It's not clear throughout the paper, however, what the relationship between these two questions is. Madva's motivation is about EFFECTIVENESS, and so is his conclusion. But we also see CAUSE playing a role in his argument. I would like to follow Madva's call for more nuanced analyses and add to his contribution with an exploration of how these questions are related. In doing that, I hope to reveal that Madva's critique against structural prioritizers exploits a misleading relationship between CAUSE and EFFECTIVENESS.

Madva presents structural prioritizers as those who defend that the most effective interventions are going to be at the structural level, and argues against that. One particular reason why structural prioritizers get things wrong according to Madva is because they take structural factors to be *the cause* of certain attitudes and/or beliefs (e.g. prejudices against certain social groups). Madva calls this view MIRROR; according to MIRROR, we acquire our biases as a sort of infection out of living in an unjust society full of stereotypes and disparities between social groups. One possible reconstruction of structural prioritizers' reasoning is: even if we were to say that the proximal cause of social inequality and discrimination is (at least partly) people's prejudices, structural interventions should be prioritized because people get their prejudices from corrupted social dynamics and structures; therefore, if we remove or change those structures, this will affect attitudes and beliefs and pave the way for change. Madva uses MIRROR to explain the predicted failure of structural interventions to eliminate discrimination and inequality. His response is that MIRROR is wrong, and everyone knows it. So the whole reasoning behind prioritizing structural interventions falls apart. We can easily see how structural prioritizers' predictions are wrong: if we remove the social factors causing biased attitudes (e.g. with affirmative action procedures), this does not make those attitudes disappear. Even worse, Madva points out how it can actually reinforce those attitudes (e.g. diversity-promoting procedures can reinforce discriminatory attitudes among priviledged individuals. See Kaiser at al. 2013). So in a sort of reverse reasoning: because biased attitudes are still in place when we remove some social factors that have likely caused them, this means the latter did not actually cause the former. This implies that structural prioritizers are wrong in prioritizing intervention at the social level.

We see here how CAUSE appears in Madva's argument: he uses MIRROR, which is about the causal relationship between social-level factors and biased attitudes, for his

conclusion about EFFECTIVENESS. There are two points I would like to make about CAUSE that complicate Madva's critique:

A. First, we don't need to go as far as MIRROR to acknowledge that social-level factors are causally involved in producing biased attitudes. In Madva's characterization MIRROR postulates only one direction of influence, and it assumes total passivity of the mind. But we can postulate a two-way direction (between mind-environment) and give a more active role to the mind, while acknowledging that social factors have a causal influence on our attitudes. Compared to MIRROR, this more nuanced approach is not obviously wrong, and so not a weak point in the structural prioritizer's position. Moreover, it acknowledges that social factors can have a causal contribution on our attitudes, without necessarily being *the* cause or the sole causal contribution. This means that a change in those factors does not in principle make all causal contributions disappear, and therefore should not be expected to have an impact on all the effects.

B. Second, and more important, we need to acknowledge the diachronic dimension of the mind-environment (two-way) causal relationship. Elsewhere I discuss a distinction that helps articulate this dimension and that is surprisingly absent in current discussions on social ontology and epistemology (Ayala, ms). I'm referring to the distinction between origin and maintenance.[1] Factors at the social level can be the causal origin of biased attitudes, and they can also be causally contributing to their maintenance, but these are two different stages in any causal relationship. Failure of a structural intervention to eliminate prejudices in people living in that society is not a sign that corrupted social dynamics are not at the causal origin of those prejudices. It could be the case that they are, but the intervention can only prevent new attitudes from being created; it cannot undo the ones that are already there. That X is the cause of Y doesn't mean that removing X will undo Y, only that after we remove X, no more X-caused Ys will happen (and this is assuming that X is *the* cause, as opposed to merely one causal factor amongst others).

Failure of a structural intervention to eliminate prejudices does not mean either that social factors are not causally contributing to the maintenance of those prejudices and biased attitudes. It could be the case that they are, but they are not the only factors maintaining those attitudes. Other attitudes and beliefs can have as strong a causal influence. That is, a structural intervention could perhaps prevent clean-mind people from getting corrupted, but we should not expect it to undo already biased minds.
If my reasoning is on track, Madva's critique is not tackling a weak point in structural prioritizers' position, and it doesn't have the impact he claims it has. MIRROR can be wrong, and we might see cases of structural interventions not eliminating biased attitudes, and still it can be the case that people's attitudes are the effect of corrupted social systems. So the structural prioritizers' reasoning that goes from CAUSE to EFFECTIVENESS could still be defended.

[1] This distinction mostly coincides with the notions of *morphogenesis* and *morphostasis* (Buckley 1967; Archer 1979).

Transition State vs. Goal State

Focusing now on EFFECTIVENESS, judging effectiveness of structural interventions deserves the same fine-grained analysis that Madva advocates for individual-level interventions. As he himself reasons about demands on individual-level interventions, demanding that one or several structural interventions will end all evils at once is not fair. Now we can see a good reason why. If we take into account the diachronic dimension and the different stages at which structural factors might have a causal influence (at both origin and during progress and maintenance), it's easy to see that one single intervention at one point in time will not do much to undo and/or reform. The debate over effectiveness seems at times wrong-headed, as if the goal is to identify the one single type of intervention that will bring social justice. But a more realistic approach finds transition states we need to reach first, before we get to the goal state, and that requires interventions that do not aim at the goal state. One such intervention could be the integration proposed by Elisabeth Anderson (2010). This intervention could take us to a transition state where some inequalities are alleviated (which is already a lot!), even if it does not eliminate people's biased attitudes. Failure of an intervention (be it at the structural or individual level) to bring about the goal state is not a sign that we are not approaching it. It might be taking us to a transition state from where the goal state is actually closer. Thus, one possible way to defend some of the structural interventions Madva criticizes is that they take us to a transition state. If so, this makes them successful, even if not final.

What are we fighting for?

Throughout the paper I wasn't sure about what, according to Madva, the aim of interventions is. One possible aim of interventions that seek social change is to reduce, and ideally eliminate, implicit bias and prejudices against certain social groups (BIASED ATTITUDES-ELIMINATION); another possible aim is to reduce and ideally eliminate inequalities (e.g. salary gaps, employment and education opportunities) (INEQUALITIES-ELIMINATION); and finally, they can aim at reducing social injustice altogether, and ideally attaining a just society (SOCIAL JUSTICE). These three aims are of course related. But they are also independent in important ways. Interventions within an INEQUALITIES-ELIMINATION project could reinforce biased attitudes, as Madva points out about the potential negative effects of diversity-promoting measures. Interventions focused on BIASED ATTITUDES-ELIMINATION could take us, even in an ideal case in which all biased attitudes are eliminated, to a society that is far away from a just one. As Sally Haslanger has pointed out (Haslanger 2014), a just society is not made up of good people only; good practices are also needed. In relation to this, I have commented elsewhere on the possibility of a society in which people hold no sexist, racist, and in general oppressive beliefs and/or attitudes, but their language usage is sexist, racist, and in general oppressive (Ayala & Vasilyeva 2016). In such society, people could, for example, introduce a sexist presupposition into a conversation, without actually believing it (because we can presuppose things without

necessarily believing they are true), and have patterns of (conversational) behavior that, without matching their beliefs and attitudes, make their interactions and social dynamics sexist.

In order to judge effectiveness of interventions, it is therefore important to know what the aim of interventions is. In the case of the unjust society made up of nice, unbiased people, only if we take the aim to be SOCIAL JUSTICE we would say that interventions failed. But if the aim is BIASED ATTITUDES-ELIMINATION, then the interventions didn't fail.

When Madva discusses Anderson's work, there could be a confusion about what Anderson's proposal aim is. He takes Anderson's structural proposal to aim at SOCIAL JUSTICE, and so her proposed interventions (e.g. integration) are expected to take us closer to a just society on all fronts, which would include reducing and ideally eliminating inequalities, and reducing and ideally eliminating biased attitudes. Then Madva notices that integration and imposed cooperative interactions will not reduce biases (they could actually backlash, as he warns). That is, in Madva's reading, Anderson's structural intervention fails, for it doesn't reduce bias and doesn't take us to a just society. But if we take Anderson's proposal to aim at INEQUALITIES-ELIMINATION, then the evaluation of its effectiveness could be different, and most important, her proposal wouldn't be open to Madva's criticism.[2]

One way to clarify the debate amongst structural prioritizers, individualist prioritizers, and critics of both, is to reframe it in terms of aims, and priorities over them. Is reducing inequalities and discriminatory practices the most important thing, or rather reducing discriminatory attitudes? Adding the "aims" layer in our analyses of interventions could prevent misleading evaluations of interventions.

On fixing minds

Finally, I would like to point out that Madva's argument is not sound if addressed against a general critical approach to individual-level interventions. Madva puts forward an invitation to explore the nuances and differences amongst interventions on people's attitudes and beliefs, and gives us good reasons on why lack of success of one intervention should not make us conclude that the individual level is the wrong locus of intervention. Even if we acknowledge the need for this finer-grained approach and for a detailed analysis of the reasons why this or that intervention fails, his argument does not support individual-level interventions in general. As I argue elsewhere (Ayala, 2016), we might have legitimate reservations about intervening on people's minds in order to build a better society, independently of concerns about effectiveness. One possible problem with e.g. fixing people's biased attitudes, is that those biases might not be bad across all contexts. What if at least some biases respond not to social identities, but to the positions individuals occupy? That is, given the

[2] I would like to note that in page 14, Madva doesn't opt for a charitable interpretation of Anderson in general. At times he seems to imply that Anderson's target is not just methodological individualism, but also ontological individualism.

overlap between certain social identities and certain social positions, it could be the case that at least some biases respond not to identity but to social position. In this society, where certain social identities systematically overlap with certain social positions, being biased against those social positions is morally problematic, because you are as a result biased against those social identities. But in a different society, where social identities and social positions do not overlap in systematic ways, these biases might not be something to avoid. This is a reason to be critical of individual-level interventions. And this is not concerned with how (in)effective they are. We could have this critical approach to interventions on individuals' minds, and still agree with Madva.

References

Archer, M. 1979. *Social Origins of Educational Systems*. London: Sage.

Anderson, E. 2010. *The Imperative of Integration*. Princeton University Press.

Ayala, S. 2016. Structural Explanations and Agency. Presented at the *Bias In Context* Conference. University of Sheffield 5-6, 2016.

-------. ms. Causes & Causal Contributions in Social Dynamics.

Ayala, S. & N. Vasilyeva. 2016. Responsibility for Silence. *Journal of Social Philosophy* 47(3):256-272.

Buckley, W. 1967. *Sociology and Modern Systems Theory*. Englewood Clifs, NJ: Prentice Hall.

Haslaner, S. 2014. What is the Domain of Social (not Political) Justice? *Political Philosopher Blog*, published on January 31, http://goo.gl/orW8NQ

Holroyd, J. & J. Sweetman. Holroyd, J. & Sweetman, J. 2016. The Heterogeneity of Implicit Bias, in J. Saul & M. Brownstein, eds. *Philosophy and Implicit Bias*, Oxford University Press.

Johnson, I. R., & Fujita, K. 2012. Change We Can Believe In: Using Perceptions of Changeability to Promote System-Change Motives Over System-Justification Motives in Information Search. *Psychological science*,*23*(2), 133-140.

Kaiser, C. R., Major, B., Jurcevic, I., Dover, T. L., Brady, L. M., & Shapiro, J. R. 2013. Presumed fair: Ironic effects of organizational diversity structures. *Journal of Personality and Social Psychology*, 104(4): 1063-1118.