




Using Quotas as a Remedy for Structural Injustice

György Barabás^{1,2} · András Szigeti^{3,4} 

Received: 1 June 2021 / Accepted: 23 December 2021 / Published online: 13 February 2022
© The Author(s) 2022

Abstract

We analyze a frequent but undertheorized form of structural injustice, one that arises due to the difficulty of reaching numerically equitable representation of underrepresented subgroups within a larger group. This form of structural injustice is significant because it could occur even if it were possible to completely eliminate bias and overt discrimination from hiring and recruitment practices. The conceptual toolkit we develop can be used to analyze such situations and propose remedies. Specifically, based on a simple mathematical model, we offer a new argument in favour of quotas, explore implications for policy-making, and consider the wider philosophical significance of the problem. We show that in order to reach more equitable representations, quota-based recruitment may often be practically unavoidable. Assuming that members of groups in statistical minority are more likely to quit due to their marginalization, their proportions can stabilize at a low level, preventing a shift towards more equal representation and conserving the minority status of the subgroup. We show that this argument has important implications for addressing, preventing, and remediating the structural injustice of unfair representation.

1 Introduction

This paper deals with a frequent but hitherto undertheorized form of structural injustice, one that arises due to the difficulty of reaching numerically equitable representation of underrepresented groups within larger groups. This form of injustice is significant because, as we will try to explain, it can occur even when diversity,

✉ András Szigeti
andras.szigeti@liu.se

¹ Division of Theoretical Biology, Department IFM, Linköping University, 581 83 Linköping, Sweden

² Theoretical Biology and Evolutionary Ecology Research Group, Hungarian Academy of Sciences, Eötvös University, Pázmány Péter Sétány 1B, Budapest, Hungary

³ Department of Philosophy, Institute of Culture and Society (IKOS), Linköping University, 581 83 Linköping, Sweden

⁴ Department of Philosophy, Lund University, 221 00 Lund, Sweden

equity and inclusion are prioritized, relevant stakeholder bias and bad intentions are factored out, and institutional organizations are designed without preference for a specific group.

Combining mathematical modelling techniques and formal ethical arguments, we show that even a recruitment policy that is free of any explicit or implicit hiring bias can unwittingly perpetuate wider societal structures and maintain the minority status of disadvantaged groups within organizations. This finding has important implications for the question of how to address this form of structural injustice. It turns out that in order to achieve fair and equal representation of disadvantaged subgroups, the use of quota-based recruitment procedures may often be crucial. Moreover, the model presented here can help optimize quota-based recruitment in combination with other possible anti-discrimination measures, and can be used to make quantifiable policy recommendations.

Specifically, the model diagnoses intragroup dynamics as a potential source of structural injustice. Intragroup dynamics characterizes how the proportions of two or more subgroups constituting a larger group change over time. We show that non-trivial dynamic properties of the intragroup proportions can prevent achieving some desired level of representation of the subgroup. Consider any group composed of two or more subgroups, where the subgroups consist of representatives from specific classes of people. For example, the group may be employees at a workplace, and the subgroups women and men employees. Suppose that a certain representation of the subgroups is desirable for moral or other reasons¹—for instance, we may strive to have equal numbers of women and men employees at a workplace where women are in minority. One might expect that gender-neutral hiring would eventually equalize the proportion of women and men in the group. However, closer inspection of intragroup dynamics shows that this is not necessarily the case. When minority members (here, women) are more likely to quit due to their relative marginalization within the group (an assumption we will investigate in depth), their intragroup proportion can freeze at a certain (low) point, which we call a “point of recalcitrance”. Such points prevent a shift towards a fairer representation of subgroup members. As a result, even an ideal hiring policy without any explicit or implicit gender bias can end up conserving the status of women as a numerical minority.

Naturally, the model assumes a vastly simplified world and as such is unable to reflect either the complex everyday reality or the intricate ethical challenges of intergroup and intragroup relations, discrimination, and prejudice. It is also true that the degree to which the model can reveal something important about real life depends on whether its predictions can be confirmed empirically. In particular, in order to show that points of recalcitrance do in fact exist in real life, empirical support will be needed for our assumption that people’s behaviour can often be significantly affected by their being in a numerical minority. Admittedly, at this point this remains a speculative, albeit to our minds *prima facie* plausible, hypothesis. That said, we would like to highlight some of the advantages of using a precise mathematical model to

¹ We emphasize that we are not making any claims here about the intrinsic desirability or all-things-considered justifiability of certain intragroup proportions (such as equality). See also Sect. 3.1.

analyze intragroup dynamics. First, the model allows us to make quantifiable predictions concerning the relative effectiveness and combined impact of the pertaining anti-discrimination measures such as preferential hiring as against mitigating marginalization. Second, the existence of “points of recalcitrance” is by no means trivial. As will be seen, one interesting feature of those points is that close to them the intragroup proportions get stuck even if they are manifestly unfair, and despite the fact that hiring itself is completely gender-neutral (or even slightly preferential towards the minority). Third, the model shows that there can be several such points. Fourth, the model also shows that once we manage to push past the unstable equilibria, the fair proportions can become stable even without preferential hiring and even if the quitting rate of the underrepresented group is still higher. So, we can estimate how long one would have to rely on preferential hiring measures to counteract the kind of structural injustice to be examined in this paper.²

The general upshot is that injustice is not always attributable to individual agents or an organization’s leadership, and so it is rightly called “structural”. Inequity and injustice can be maintained or perpetuated even when all stakeholders and organizational actors prioritize diversity and inclusion, and strive to reduce the effects of bias. This, however, does not mean that we are powerless to transform or neutralize the effects of these unjust structures.³ To be successful in doing so, however, we also need to understand the broader systemic causes and the often unseen microforms of intragroup marginalization and exclusion.

2 A Simple Model of Intragroup Dynamics

2.1 Model Assumptions

We use a simple mathematical model to identify the nontrivial properties of intragroup dynamics mentioned above. We present its basic features and findings without any formal details in the main text (for a more technical and rigorous description, see the “[Appendix](#)”). Although the model is developed—by way of example—for two subgroups of women and men employees within the larger group of employees at a workplace, it is a proof-of-concept model that emphasizes relevant phenomena in the simplest possible context. We grant that in order to draw precise quantitative conclusions that might be used in real-world policy-making one would require a much more detailed model, parameterized using empirical data and state-of-the-art statistical techniques. Moreover, even the predictions of such a more complex model would have to be tested empirically to see what it can show about the real world. Here, we merely want to demonstrate the possibility that under a small set of reasonable assumptions, the intragroup proportions will be dominated by certain “points

² We are grateful to our reviewers for insisting that we clarify the advantages and limitations of our model-based approach.

³ And, of course, if those in charge fail to take action once they become aware of the existence of such unjust structures, they can be faulted for their omissions.

of recalcitrance” which prevent a shift towards other, potentially more desirable proportions. We then go on to argue for the philosophical significance of this finding.

The model’s assumptions are as follows. We only take two subgroups into account, and assume that the following three processes act to change their relative proportions: (1) new hires, happening at gender-specific annual rates; (2) retirements, assumed to be equal across genders; (3) finally, employees may also quit for reasons of marginalization caused by being a minority at work. This final rate depends on the subgroups’ current proportions at the workplace: the higher the proportion of one subgroup, the less likely it is for them to leave before retirement. Beyond a certain level of representation, the group is no longer a minority and there is no more incentive to quit for this reason anymore, so the quitting rate is effectively zero beyond that point (Fig. 1).⁴ Additionally, we assume compensatory hiring: new members are hired whenever an old one retires or quits, so the total number of employees is always the same.

Putting the three processes of hires, retirements, and premature quitting together, Fig. 2 gives a graphical overview of the model’s behavior, for two different scenarios (left and right panels): one with gender-neutral, and one with preferential minority hiring where 60% of hires are women. The graphs show, within the model, the rate of change of the proportion of women at the workplace as a function of their current proportion (see the “Appendix” for details of why the curves have the depicted shapes). A positive (negative) rate of change means that the proportion of women in the model is increasing (decreasing). If the rate of change is exactly zero, then the proportion does not change: the modelled gender ratio reaches an equilibrium. An equilibrium point is stable if small deviations away from it decay and the model system eventually returns to the same equilibrium, and unstable otherwise (indicated by solid/open circles in Fig. 2, respectively). Keeping these facts in mind, one can trace how the fraction of women changes in time within the model by following along the horizontal axis, moving in the direction indicated by the value of the rate of change: rightward for positive and leftward for negative rates.

In the left panel of Fig. 2 for instance, what we see is that unless the initial fraction of women is larger than ca. 31% (unstable equilibrium denoted by open circle), their modelled proportion is predicted to slide down to the leftmost stable equilibrium at around 18%. This is what we call an undesirable “point of recalcitrance”, preventing the group from reaching an equal gender ratio—despite the fact that hiring is completely gender-neutral. One way to get out of this trap and achieve gender equality in the modelled world (right panel) is to represent the preferential hiring of

⁴ It is worth emphasizing that we do not assume that being in a numerical minority *necessarily* leads to higher quitting rates. The model’s predictions are made merely on the assumption that being in a numerical minority *can* lead to higher quitting rates. While we will adduce some considerations below why this seems to be a plausible assumption, we will also discuss why this assumption may not obtain in certain situations, for instance, when some perceived or real benefit accrues to members of the minority group due to their minority status, or when the behaviour of members of the minority is simply unaffected by their minority status either way. In general, whether and when the assumption in fact obtains, and so too the prevalence of the predicted points of recalcitrance in real life, must be the object of careful empirical study. We were greatly helped by our anonymous referees in making these points clearer.

women by increasing their hiring rate to 60%. This eliminates the point of recalcitrance altogether, and thus enables the model system to be dragged over the critical threshold to reach more than 31% representation. Once this has been achieved, hiring in the model may be reset to being gender-neutral, recovering the scenario of the left panel, but with a crucial difference: the proportion of women would now be slightly above the unstable equilibrium at 31%. As its rate of change is now positive, the modelled proportion would approach the stable point at 50%, with equal representation. Unlike its counterpart down at 18%, this stable equilibrium is desirable: once the system is close to this point, deviations from it are counteracted and equality is eventually restored. Thus, the proposed model predicts that it may be possible to end up with a stable gender imbalance even with gender-neutral hiring, and that the use of hiring quotas could in principle alleviate this imbalance.

2.2 Discussion

The main predictions of the model can be summarized as follows. Given certain assumptions, if the proportion of the minority subgroup is low to begin with, this subgroup will not equal the proportion of the larger subgroup, even if hiring is in equal proportions. This will be the case so long as members of the smaller subgroup quit at a higher rate than members of the larger subgroup, which makes it impossible for their proportion to equalize. Rather, the proportion of the smaller subgroup stabilizes at some low level, which we call the “point of recalcitrance”. Under these assumptions, only if the proportion of the small subgroup reaches some threshold level can it subsequently increase to an even representation. Above this threshold then, the smaller subgroup is sufficiently represented so that its members no longer have a strong incentive to quit due to being in minority, so equal hiring will eventually equalize their proportion. Often, however, this threshold is well above the point of recalcitrance, and will therefore be simply unattainable through parity-based recruitment. So, we predict that in order to force the proportion of the smaller subgroup over the threshold, members of that group have to be given preference when new members are recruited.

Hopefully, the general advantages of using simple mathematical models such as ours have also become easier to discern even in view of the obvious simplifications and limitations the model makes relative to the complex reality of intragroup dynamics (some of which we will take up below). In general, models force one to make all assumptions explicit, and open up the toolbox of formal mathematics to aid the analysis. A model also allows one to make quantitative predictions: instead of simply concluding that preferential hiring is needed to equalize the intragroup proportions, one can determine how strong that preference should be, for how long it should be applied, and how much time it will take for these measures to actually lead to the eventual equal representation. Finally, an important advantage of a mathematical model is that it can lead to surprising conclusions that would have been much more difficult to imagine without its help.

In our case, the model has revealed that intragroup dynamics could be more nuanced than simple conceptual arguments may reveal. A claim one might have

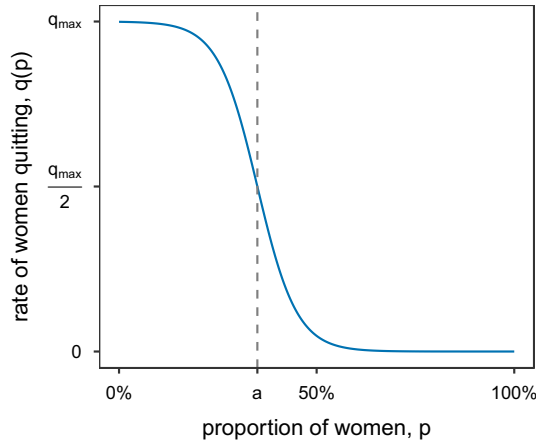


Fig. 1 An illustration of how the rate at which members of the minority subgroup (here, the subgroup of women) prematurely quitting their job can be modelled to change with their proportion at the workplace. This rate gradually decreases from its maximum as the proportion of minority subgroup increases. Beyond a certain proportion, there is no more incentive to quit due to being a minority, reflected by the curve approaching a quitting rate of zero as the proportion of women increases. The point of steepest descent of the function is at a (vertical dashed line); very roughly speaking, the minority group is quite likely to quit if their proportion is smaller than a , and no longer very likely to quit if their proportion exceeds a

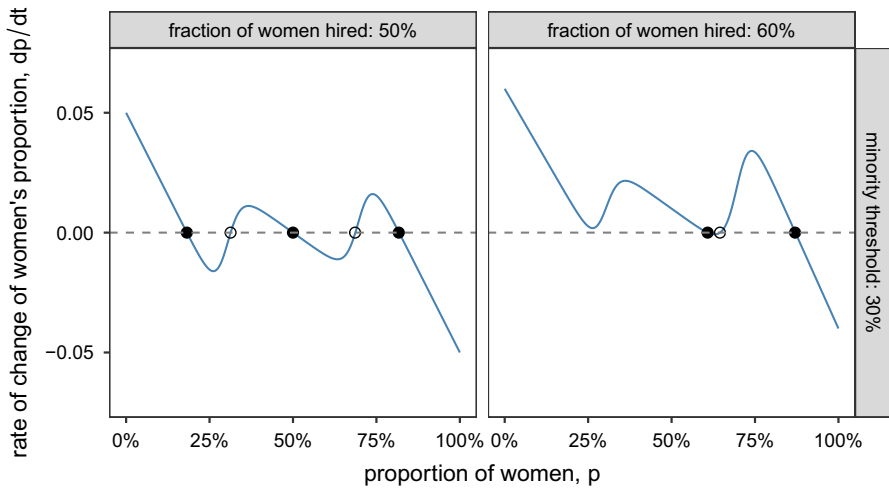


Fig. 2 The modelled rate of change of the proportion of women as a function of their actual proportion, when hiring is gender-neutral (left) versus preferential towards women (right). Solid/open circles show stable/unstable equilibria. Parameters: $r = 0.1$, $a = 0.3$, $q_{max} = 0.35$, $s = 0.04$, and f is either 0.5 (left) or 0.6 (right)

accepted without the model is that gender-neutral hiring will eventually equalize the workplace gender ratio. As we have seen, this may simply not turn out to be the case whenever members of the minority group are more likely to leave and their initial proportion is below the threshold of instability. Similarly, one might have thought that since hiring is equal but minority members leave at a higher rate, the resulting workplace proportions will trivially end up being unequal. Once again, the model shows that this does not have to be the case. If the minority subgroup's proportion starts above the unstable threshold (above ca. 31% in the left panel of Fig. 2), then it can be predicted to equalize eventually—despite the fact that, up until having reached equality, they are always in minority and therefore have a greater propensity to quit than members of the majority subgroup. Finally, the fact that there may exist multiple alternative stable states (where a subgroup is in strong minority, and one with equal representation) is a nontrivial consequence of our reasoning which would have been difficult, if not impossible, to intuit without a model.

It is also worth discussing in more detail here what we call the *marginalization assumption* in the model. This is the assumption that members of an underrepresented group are more likely to quit due to the very fact that they are a numerical minority in the organization. It plays an important part in our model, but naturally this assumption need not always be true. It is even possible that being in the minority will decrease the probability of quitting because there are ways in which that minority status could be seen as beneficial. Further, as noted, it may also be that the quitting rate of members of the minority group will not be significantly affected by their minority status—perhaps because members of the majority group work actively to create a welcoming and inclusive environment, or perhaps due to the presence of other, non-attitudinal factors. That said, it is plausible that the marginalization assumption applies in many concrete situations, for the following reasons. First, the expression of important aspects of one's identity (e.g., language, religion, gender, sexual orientation) and the representation of certain concerns and needs within a larger group depend on there being a sufficient number of other members sharing the same identity, concerns, and needs. Second, other things being equal, marginalization is also more likely to give rise to segregation, stigmatization, and discrimination.⁵ Third, there is an increased risk of an unfair distribution of resources due to the latent or explicit tyranny of the majority.

Future work can determine how widely applicable the marginalization assumption may be. It is worth noting though that empirical research has produced findings consistent with the marginalization assumption. Here is a sample of the available evidence.⁶ First, it has been found that the underrepresentation of women in philosophy departments is not just a “pipeline problem”.⁷ This means that the problem of underrepresentation cannot be resolved just by making sure that there are enough women PhDs and junior women philosophers in the profession. Second, studies of

⁵ Lippert-Rasmussen (2018).

⁶ We offer a systematic discussion of pertaining empirical material in a companion paper.

⁷ Dodds and Goddard (2013). Schiebinger (2000). Allen and Castleman (2001).

women in legislatures show that a critical mass of women is necessary to effectively promote a feminist political agenda broadly understood.⁸ And third, data show that African-American students suffer high attrition rates at law schools despite the fact that many of these schools use affirmative action to recruit students.⁹

There is one more issue regarding the generalizability of our model we would like to address here. As noted, we develop the model for the hiring of women and men but claim that it is applicable well beyond the area of gender relations. However, it could be argued that the case of women is special insofar as women constitute a group which is the object of discrimination but which actually constitutes a numerical *majority*. By contrast, many groups (e.g., ethnic or religious minorities) facing discrimination are often in a significant numerical minority in society at large. This situation might be seen to raise the problem that groups in a significant numerical minority (e.g., African-Americans at 12% of the population in the US) may never get past the point of recalcitrance.

In response, we note that our model is limited to institutional contexts—e.g., education, employment—with regard to which it is typically quite reasonable to make the assumption that there will be enough sufficiently qualified people to be hired from the minority to push past the point of recalcitrance (especially when inclusivity measures to improve the group's retention are also used). All the more so as the distribution of any such group can be expected to be heterogeneous with respect to economic sector, geographical location, etc. (For example, even if African-Americans make up 12% of the US population this does not mean that this ratio is 12% in every relevant demographic, economic sector, place, and so on.) Further, we do not make any claims regarding what level of representation of the minority is desirable for moral or other reasons (more on this in Sect. 3.1 below). Thus, in some contexts it may be the case that if a certain group is in a significant numerical minority in society at large (say 3%), then there will be good reasons not to aim for a desired level of intragroup representation within the target hiring group well beyond that level (say at > 40%).¹⁰ In fact, one such reason for setting the desired level somewhat lower could be that the assumption just mentioned will not hold, namely there will simply not be enough sufficiently qualified candidates from the minority to maintain a higher level of intragroup representation.

But of course this last reason may not prevail all things considered. That is, it is indeed possible that while there are compelling reasons to set the desired level of representation within the target group at a certain level, there will simply not be enough sufficiently qualified candidates from the minority to push past the point of recalcitrance to reach that level. In such cases no quota-based increase in hiring will have any long-term effects because the equilibrium point corresponding to an

⁸ Beckwith and Cowell-Meyers (2007).

⁹ Sander (2004). See also (Fullinwider 2018). With respect to this last example we also note that switching from the context of gender to that of racial discrimination might very well raise specific problems as regards the applicability of our model (see the last three paragraphs of this section for the discussion of precisely such a problem). We thank an anonymous reviewer for pressing us on this point.

¹⁰ Needless to say, there can be very good reasons—e.g., remediation of historical injustice or importance of diversity—for setting the desired level of intragroup representation of any given minority group well above its share in the total population.

equitable representation is absent. In fact, one advantage of our model is that it can show under what conditions such cases are possible. However, we note, first, that for reasons mentioned above such cases are not likely to occur frequently in practice, and second, that when they do occur alternatives or complements to quota-based hiring—e.g., inclusivity measures to reduce the minority's quitting rate—will still be available.

3 The Moral Significance of Intragroup Dynamics

Here we discuss the implications of our model of intragroup dynamics for the debate regarding the permissibility and feasibility of affirmative action, on the one hand, and the understanding of structural injustice more broadly on the other. First, the conclusions reached from the model offer a new argument in favor of quota-based affirmative action. Second, the model also offers ways to evaluate the relative effectiveness and combined impact of diverse strategies for addressing structural injustice, insofar as it manifests itself in the unfair representation of subgroups within larger groups. Third, and more broadly, the model bears on the general debate about the causes of structural injustice.

3.1 The Debate About Affirmative Action

The model shows that if affirmative action is to succeed, then the use of quota-based recruitment procedures may often be unavoidable in practice. There are several novel features of this argument.

First, as just noted, it is independent from the aims and justification of affirmative action. The underrepresentation of a subgroup in a larger group could be unfair for a variety of reasons. A meritocratic reason would be that since members of the minority group are equally qualified, it is not fair that they are heavily underrepresented. However, there can be other, non-meritocratic reasons too why intragroup underrepresentation could be unfair. For instance, if women make up half of the population of a certain country, then it does not seem to do justice to the principle of representative democracy that only a small proportion of that country's members of parliament are women.¹¹ It is also possible that changing the intragroup proportions within a group is thought to be desirable for non-moral reasons. For example, urban planners may want to boost economic prosperity by raising the proportion of (say) young people among the inhabitants of a certain residential neighborhood. Whatever the justification, our model shows how to identify and overcome the obstacles posed by intragroup dynamics to realizing adequate representation.¹²

¹¹ Rosenblum (2007).

¹² That said, we are aware, as already noted, that the model's applicability might vary significantly depending on both which specific group characteristic—gender, race, age, etc.—is at issue and what the particular aim of affirmative action is.

Second, our argument implies that the reason for “administering justice” by quotas is not the cost or difficulty, but rather the impossibility of doing so at the individual level. As we noted, it may be unavoidable to employ some form of quota-based affirmative action policy in many cases. Once this is recognized, we can mount a strong response to certain criticisms of affirmative action such as the argument that affirmative action amounts to reverse discrimination.¹³ On this view, quota-based affirmative action is supposed to be problematic because “individuals [are] regarded merely as members of that group rather than in their individuality”.¹⁴

The standard reply to the reverse discrimination objection is that attempting to determine whether each individual member of a given target group is genuinely disadvantaged or not would be prohibitively difficult and too costly in practice.¹⁵ Our model offers a new angle on this debate showing that the reason for quota-based affirmative action is not the cost or difficulty of case-by-case adjudication. Rather, without using quota-based affirmative action, we may simply not be able to fulfill our obligations towards the relevant individuals at all. This is because the dynamics of intragroup proportions can be such (when quitting rates of the minority cannot be lowered) that without the use of quotas affirmative action could completely fail to improve representation.¹⁶

3.2 Redressing the Structural Injustice of Underrepresentation

The model identifies several possible measures to improve representation of the target subgroup. One of these measures is preferential hiring. Another is improving inclusiveness so that individuals from underrepresented subgroups are less inclined to leave the larger group. A considerable advantage of using a model-based approach with empirically-testable assumptions is that we can quantify the relative impact of these measures (e.g., how strongly should one skew the hiring policy in favor of the minority group if there is no possibility of reducing the rate of quitting?). While our proof-of-concept model will need future refinements to be used for making actual, effective real-world policy suggestions, the idea of how this could be done can be illustrated by considering the four scenarios below.

The first scenario (Fig. 3, top left), with equal hiring rates of women and men, is qualitatively identical to the one in the left panel of Fig. 2. The conclusions are

¹³ Cowan (1972), Nunn (1974), Goldman (1975), Sher (1975), Simon (1978), Goldman (2015), Fullinwider (2018), Lippert-Rasmussen (2018).

¹⁴ Cowan (1972).

¹⁵ Lippert-Rasmussen (2018), Nickel (1972), Thomson (1973), Nagel (1973), Nickel (1974).

¹⁶ This should not be taken to imply that there is an all-things-considered obligation to use preferential hiring in the relevant cases. We hasten to make this clarification anticipating another objection to affirmative action, the so-called “miner’s son objection”, also known as the “mismatch objection”; see Lippert-Rasmussen (2018). The objection is that preferential hiring can lead to the unfair treatment of individuals who are equally or more disadvantaged than members of the minority group, such as the gifted and hard-working son of an unemployed miner from a poverty-stricken region (Nunn 1974; Goldman 1975, 2015). We suspect that no argument can show that underprivileged but deserving individuals who are passed over due to a policy of preferential hiring cannot even have so much as a *pro tanto* complaint.

therefore also the same: if the initial proportion of women is below the unstable equilibrium point at around 33% (open circle to the left), then their modelled proportion will move towards the point of recalcitrance, the stable equilibrium near 13% (leftmost solid circle). Only if the initial proportion exceeds 33% would it eventually stabilize at 50% (solid circle in the middle). Finally, if the initial proportion exceeds 67% (open circle to the right), then, since the rate of change is now positive, the proportion would increase until 87% (rightmost solid circle). This would lead to a reversal of gender roles: now men are the minority with a high propensity to quit prematurely.

The second scenario (Fig. 3, bottom left) has the same equal hiring rate, but quitting rates from being in minority are reduced by placing the point of transition from high to low quitting rates earlier. In terms of our model, this means that the parameter a is lowered from 30 to 20%.¹⁷ That is to say, in this model scenario minority members can be less numerous than before to no longer feel marginalized. What the model shows for this parameterization is that the proportion of women needs to exceed 20% (open circle to the left) to converge to equal representation at 50%. Significantly, however, for initial proportions less than 20%, the representation of women would still stabilize at a low 13%, despite the fact that women are less affected by marginalization in this scenario.

In the third scenario (Fig. 3, top right), quitting rates are as high as in the first one, but the hiring rate is skewed to favor women ($f = 0.6$). It is interesting to observe that, by itself, a *moderate* affirmative action regime such as this one does not improve matters significantly in the model. The point of recalcitrance will still only move to around 18% (solid black circle to the left), and the proportion of women must exceed 31% (open circle) to increase above 18% in the long run.

In fact, one can calculate that in this model example, the hiring rate of women would have to be increased to at least 69% to eliminate the lower point of recalcitrance—that is, to reach a high representation of women even if their initial proportion is low. In reality, since enforcing such a strict quota in which approximately 7 out of every 10 hires are women may be difficult for administrative or other reasons, it is worth taking a look at the potential impact of a combination of measures. For example, one could skew hiring rates to favor women as in the third scenario *and* assume a reduced effect of marginalization as in the second.

This possibility is shown in our fourth and final scenario (Fig. 3, bottom right). Thanks to a combination of measures, it is suddenly possible for women to achieve high representation in the model, even if their initial proportion is low. As can be seen, the lower point of recalcitrance is eliminated, and instead the system approaches the stable equilibrium at 60%. While this scenario does not yield gender equality, balance can be achieved by resetting the hiring rate to $f = 0.5$ and stopping affirmative action after the proportion of women has reached 50%. Once this is done, the dynamics of the second scenario take over, with gender parity becoming stable in the long run.

¹⁷ See Fig. 1 and the “Appendix”.

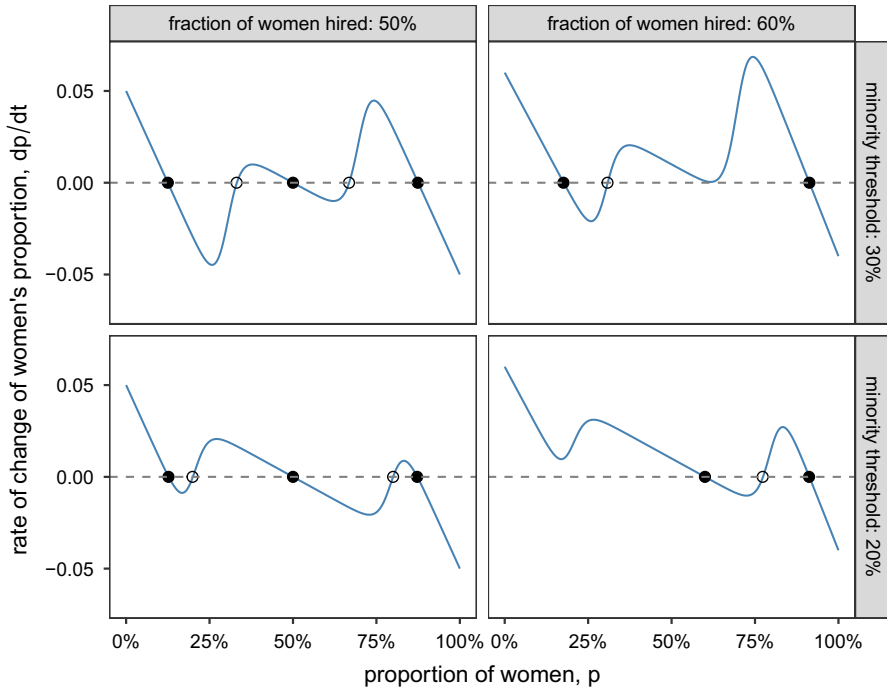


Fig. 3 The rate of change of the proportion of women in the model, dp/dt , as a function of the actual proportion p , for four different model parameterizations. Solid/open circles show stable/unstable equilibria. Columns show scenarios with the same relative hiring rate ($f = 0.5$ for the left and 0.6 for the right column); rows show scenarios with the same inflection point a of the quitting function ($a = 0.3$ in the top and 0.2 in the bottom row; see Fig. 1 and the “Appendix” for details). All other parameters are fixed and equal across scenarios: $r = 0.1$, $q_{\max} = 0.6$, $s = 0.04$

These four scenarios are instructive because they allow us to quantify the impact of quota-based affirmative action. They illustrate how to eliminate the stable equilibrium—the point of recalcitrance—where the proportion of women is low. Sometimes, we predict, this will only be possible by means of a strongly preferential hiring regime. If other measures can also be used to reduce the probability that minority members will leave (represented by adjustments in other model parameters), then affirmative action may be combined with those measures to yield a more effective way of achieving equality. So, while our model helps us see how to fight the adverse effects of certain structural obstacles by using quotas, it also throws into sharper relief the importance of measures aimed at improving inclusivity and reducing social hostility.¹⁸

¹⁸ We have created a web application at <https://dysordys.shinyapps.io/shinyapp/> where every parameter can be individually manipulated to examine their effect on the model.

3.3 Structural Injustice and Agency

The conclusions drawn from the model suggest a more general lesson as well. In practice, the complete elimination of explicit and implicit bias may not be achievable in the treatment of various groups. In fact, through the marginalization assumption our own model also leaves room for the impact of bias in the explanation of the persistence of underrepresentation, since one of the reasons for the higher quitting rate of the numerical minority could well be their exposure to various forms of bias. What our model shows, however, is that even a hiring policy that is free of a specific form of bias, namely hiring-bias, can fail to ensure adequate intragroup representation. This in turn, we submit, provides a vivid illustration of the claim that, while efforts to expose and correct explicit and implicit bias are important, “the most profound injustices will remain entrenched” without accompanying structural change and the redistribution of resources.¹⁹ When underrepresentation is unjust, the non-trivial characteristics of intragroup dynamics constitute a kind of mechanism—and quite a robust one at that—by means of which structural injustice is perpetuated. Our model highlights how this can happen even when all the stakeholders act to eliminate explicit and implicit bias from their hiring practices and are genuinely committed to rectifying the problem.²⁰

As such, the model bears on the ongoing debate between individualist and structuralist explanations of various forms of inequality.²¹ Structuralists argue that the most important forms of discrimination and underrepresentation often do not derive from the morally objectionable behaviour of individual agents. Rather, in many cases they are to be traced back to complex systems of disadvantage maintained by political, social, and cultural institutions and norms.²² When talking about structural injustice in this paper, we too understand the adjective “structural” to indicate that the form of injustice in question cannot be attributed to the culpable agency of individuals (or groups). Structuralists do of course accept that individual attitudes—e.g., explicit and implicit bias—and behaviours can play a part. However, they argue that structural constraints and mechanisms are often necessary and sometimes even in themselves sufficient to create and maintain unjust practices and arrangements.

¹⁹ Haslanger (2015). We do not mean to imply of course that Haslanger would necessarily regard the model-based policies we recommend as the kind of structural change she advocates.

²⁰ We also note that while in practice the marginalization and therefore the higher quitting rate of minorities is likely to be due at least in part to their experience of various forms of bias, there can be many reasons behind the higher quitting rates of the minority which are not due to bias. For example, it may not be expedient to use more than one language in the group, and for various reasons it may be prohibitively costly to adopt the numerical minority’s language as the common one. That fact, in itself unrelated to bias, could under some circumstances significantly contribute to the minority’s higher quitting rate. We are grateful to an anonymous reviewer for urging us to circumscribe more precisely the role of bias in our model.

²¹ Zheng (2018a, 2021).

²² Lavin (2008), Young (2010), Haslanger (2015), Kaufman (2020).

Individualists, by contrast, point to the many cases in which individual attitudes and behaviours clearly engender and maintain unequal and unfair practices.²³ Moreover, they also argue that the appeal to structures lets individual agents off the hook all too easily thereby conveniently sidestepping the question of the moral responsibility of individual agents.²⁴ This potential implication of structuralism is seen by individualists as hugely problematic because in the absence of individual moral responsibility it becomes unclear who should be in charge of addressing structural injustice and remediating the harms it causes. In addition, individualists have also made use of ontological and action-theoretic arguments to the effect that, ultimately, no structures would exist if individuals did not sustain them through their actions and omissions.²⁵

Our findings provide support for some of the insights of both individualists and structuralists. On the one hand, our analysis shows that the dynamics of intragroup proportions can result in unfair underrepresentation even when every effort is made by the stakeholders to adopt hiring policies free of explicit or implicit bias. Moreover, even if the impact of bias cannot be eliminated in practice, our model points to an important factor in addition to individual bias and culpable behaviour, one that plays a significant role in explaining the difficulty of reaching the desired numerical level of representation of the minority group. To put it differently, since higher quitting rates in the minority group are not exclusively due to pernicious attitudes and actions (of individuals or groups), even in a world of perfect angels the problem of points of recalcitrance will have to be dealt with if desirable representations are to be achieved. These predictions, we claim, sit comfortably with the views of structuralists.

On the other hand, our model also shows that individual attitudes and behaviour can also play a part in the persistence of the kind of structural injustice we are interested in. As the model shows, the higher quitting rate of the underrepresented group can be crucial in explaining why more equitable intragroup proportions are difficult to reach. We noted in this connection that the higher quitting rate can be due to the fact that members of the underrepresented group feel (with or without reason) marginalized, disadvantaged, and exposed to various forms of bias, discrimination and stigmatization. Quite obviously, individual attitudes and behaviour exhibited by members of the larger group will have a great impact on whether members of the underrepresented group will feel that way or not. As noted, adopting a more welcoming attitude and undertaking other measures aimed at improving inclusivity are viable ways in which quitting rates can be reduced even when a minority group does not reach numerical parity. Conversely too, in certain cases even being in a numerical parity or majority may not reduce quitting rates if the effects of social animosity or prejudice continue to be intensely experienced.²⁶ So,

²³ For insightful discussions of how individuals contribute to unjust practices, see esp. Saul (2013), Brownstein and Jennifer (2016), Zheng (2018a).

²⁴ Zheng (2018b, 2021).

²⁵ Ludwig (2016, 2017).

²⁶ A further potential complication is that in certain groups the use of quotas might lead to the increase of bias and discrimination by (members of) the majority against (members of) the minority. Counteracting this might require even more strongly preferential hiring, at least temporarily. In order to be applicable to such groups, a more complex version of the model is needed that can represent the impact of such vicious loops.

again, in light of their complexity the analysis (and rectification) of real-life situations involving discrimination and injustice will always require careful empirical scrutiny, but our model does point at possible ways to enhance our understanding of how structural factors can interact with attitudinal ones.

Relatedly, our approach can steer clear of controversies regarding who or what can be a morally responsible agent. Collectivists argue that groups can be agents and be morally responsible.²⁷ Others disagree. It is said that groups are not the kind of things that can be agents, or even if they can be called agents, then only in an attenuated or metaphorical sense so that attributions of non-distributive moral responsibility to them would not be warranted.²⁸ Avoiding these theoretical controversies, our model shows that unfair representation of minority groups cannot always be traced back to the influence of specific agents—regardless of whether they are individual or collective agents.

This in turn has important implications regarding the question of who should address and remediate structural injustice. It is often assumed that duties of remediation and redress befall primarily on those blameworthy for committing structural wrongdoing in the first place.²⁹ However, our analysis shows that in some typical cases of structural injustice it is not possible to pinpoint salient agents—neither individuals nor groups—as responsible for structural injustice.

It is worth stressing that this does not mean that no individual or collective has the duty to address structural injustice and remediate its harms.³⁰ While we do not directly address here the question of allocating remedial duties, we have identified and evaluated available measures for remediation and prevention of future re-occurrence (see Sect. 3.2 above). This should make it easier to establish to whom (which groups and individuals) we should assign such remedial duties.

4 Conclusion

Based on a simple model of intragroup dynamics, this paper has argued that if a given ratio between two subgroups of a larger group is desirable for some reason, then preferential hiring in favor of the smaller group may in many cases be unavoidable. This can happen whenever members of the minority subgroup are more likely to leave due to their marginalization. As we have shown, higher quitting rates can lead to undesirable intragroup ratios becoming stable equilibria—what we have dubbed “points of recalcitrance”. In this situation, the minority subgroup will continue to be underrepresented.

We argued that our model of intragroup dynamics has important implications for the affirmative action debate as well as the ethics of structural injustice. We not only offered a new argument in favor of quota-based affirmative action, but we

²⁷ List and Pettit (2011).

²⁸ Miller and Makela (2005), Haji (2006), Ludwig (2016, 2017).

²⁹ Holroyd (2012, 2015), Brownstein (2016).

³⁰ Singer (1972), Miller (2001), Kutz (2007).

also used the model to evaluate the relative effectiveness and combined impact of diverse strategies for addressing the structural injustice of underrepresentation. In particular, we predicted that while affirmative-action quotas may often be crucial to achieve a fair, inclusive, and representative distribution of diverse groups, it is also important to pay close attention to the retention of members of underrepresented groups, thus making the success of those policies sustainable and enduring.

Finally, we highlighted the fact that this form of structural injustice can be caused by intragroup dynamics even in the absence of explicitly or implicitly biased or prejudiced hiring practices. At the same time, we insisted that even if no agents are culpably involved in causing or maintaining underrepresentation this should not be taken to imply that nobody can or should do anything about it, once the existence of this form of injustice becomes common knowledge. We hope that the model and its interpretation presented in this work can prove useful in identifying and fine-tuning measures to remediate and prevent the structural injustice of underrepresentation.

Appendix

Here we present the derivation of our model. Consider an institution hiring employees, at gender-specific rates h_w (women) and h_m (men). Employees retire at a rate r . Also, employees may quit prematurely, for reasons of marginalization caused by being a minority member at work. These rates, denoted q_w for women and q_m for men, depend on their current proportions at the workplace. Generally speaking, the higher the proportion of one gender, the less likely it is for them to leave before retirement. Assuming that the above three processes (hires, retirements, and premature quitting) are the only factors changing the number of employees, the model may be cast in differential equation form:

$$\frac{dw}{dt} = h_w - wr - wq_w, \quad (1)$$

$$\frac{dm}{dt} = h_m - mr - mq_m, \quad (2)$$

where w and m denote the number of women and men employees, t is time, and d/dt denotes the derivative (instantaneous rate of change) with respect to time. Since retirements and premature quittings happen to individuals, with each individual having a certain probability of leaving within a given time period, r , q_w , and q_m are interpreted as per capita rates. That is why they are multiplied in Eqs. (1) and (2) by the number of women and men, respectively, to obtain group-level rates.

We assume that the total number of employees at the workplace is kept constant: hiring rates are always adjusted to fill in the positions of leaving employees. Denoting the total number of employees $w + m$ by N , this means that

$$\frac{dN}{dt} = 0. \quad (3)$$

But since $N = w + m$, this can also be written as

$$\begin{aligned} \frac{dN}{dt} &= \frac{d(w + m)}{dt} \\ &= h_w - wr - wq_w + h_m - mr - mq_m, \end{aligned} \quad (4)$$

where we used Eqs. (1)–(2) and the sum rule of differentiation (the derivative of a sum is the sum of the derivatives). Equations (3) and (4) together imply

$$h_w - wr - wq_w + h_m - mr - mq_m = 0, \quad (5)$$

which means that the hiring rates must sum to

$$h_w + h_m = wr + wq_w + mr + mq_m. \quad (6)$$

Denoting by f the fraction of hires who are women (then the fraction of men hires is $1 - f$), the individual gender-specific hiring rates are written

$$h_w = (wr + wq_w + mr + mq_m)f, \quad (7)$$

$$h_m = (wr + wq_w + mr + mq_m)(1 - f). \quad (8)$$

Our focus is the fraction of women and men at the workplace. Let us denote the fraction of women by $p = w/(w + m) = w/N$; we then have that the fraction of men is $1 - p = m/N$. The dynamics of p can be obtained by substituting Eq. (7) into Eq. (1) and dividing by N :

$$\begin{aligned} \frac{dp}{dt} &= \frac{1}{N} \frac{dw}{dt} \\ &= \left(\frac{w}{N}r + \frac{w}{N}q_w + \frac{m}{N}r + \frac{m}{N}q_m \right) f - \frac{w}{N}r - \frac{w}{N}q_w. \end{aligned} \quad (9)$$

Using $p = w/N$ and $1 - p = m/N$:

$$\frac{dp}{dt} = [pr + pq_w + (1 - p)r + (1 - p)q_m]f - pr - pq_w, \quad (10)$$

which can be rearranged to read

$$\frac{dp}{dt} = (r + q_m)f(1 - p) - (r + q_w)(1 - f)p. \quad (11)$$

To move forward, we assume that premature quitting is symmetric with respect to gender: the probability of a woman quitting a workplace where a fraction x of workers are women is the same as the probability of a man quitting a workplace where the fraction of men is x . Mathematically, since the fraction of women is p and the fraction of men $1 - p$, this means that we can express both q_w and q_m through the same function: $q_w = q(p)$ and $q_m = q(1 - p)$, where $q(p)$ is the rate of quitting by

women given that they make up a fraction p of the workplace. With this assumption, Eq. (11) is written

$$\frac{dp}{dt} = [r + q(1 - p)]f(1 - p) - [r + q(p)](1 - f)p, \quad (12)$$

which is the final form of our model.

An illustration of the general shape of the function $q(p)$, and the one we adopt here, is given by

$$q(p) = \frac{q_{\max}}{2} \left[1 - \tanh \left(\frac{p - a}{s} \right) \right], \quad (13)$$

where $\tanh(\cdot)$ is the hyperbolic tangent function (Fig. 1). Substituting this into Eq. (12), we obtain the functions on the right hand side that are plotted in Figs. 2 and 3. See the captions for the specific parameter values used in each figure.

To be able to explore the model's behavior for alternative parameter choices and to facilitate a better understanding of how it works, we have created a web application where every parameter can be individually manipulated and their effects on the model's behavior examined (<https://dysordys.shinyapps.io/shinyapp/>).

Acknowledgements The authors would like to thank two anonymous reviewers as well as Robin Abbey-Lee, Kasper Lippert-Rasmussen, Geneviève Metson, the Division of Theoretical Biology at Linköping University, and other audiences in Aarhus, Lund and elsewhere for helpful discussions. GB was funded by the Swedish Research Council (Vetenskapsrådet), grant VR 2017-05425. ASz gratefully acknowledges financial support by the Lund Gothenburg Responsibility Project (LGRP) funded by the Swedish Research Council.

Funding Open access funding provided by Lund University.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Allen, M. & Castleman, T. (2001). Fighting the pipeline fallacy. (In A. Brooks & A. Mackinnon (Eds.), *Gender and the restructured university* (pp. 151–165). Buckingham: Open University Press.)
- Beckwith, K. & Cowell-Meyers, K. (2007). Sheer numbers: Critical representation thresholds and women's political representation. *Perspectives on Politics*, 5, 553–565.
- Brownstein, M. (2016). Attributionism and moral responsibility for implicit bias. *Review of Philosophy and Psychology*, 7(4), 765–786.
- Brownstein, M. & Jennifer, S. (2016). *Implicit bias and philosophy: Moral responsibility, structural injustice, and ethics* (vol. 2). Oxford: Oxford University Press.
- Cowan, J. L. (1972). Inverse discrimination. *Analysis*, 33, 10–12.

- Dodds, S. & Goddard, E. (2013). Not just a pipeline problem. In K. Hutchison & F. Jenkins (Eds.), *Women in philosophy* (pp. 143–164). Oxford: Oxford University Press.
- Fullinwider, R. (2018). Affirmative action. In Edward N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/entries/affirmative-action/>
- Goldman, A. H. (1975). Reparations to individuals or groups. *Analysis*, 35, 168–170.
- Goldman, A. H. (2015). *Justice and reverse discrimination*. Princeton: Princeton University Press.
- Haji, I. (2006). On the ultimate responsibility of collectives. *Midwest Studies in Philosophy*, 30(1), 292–308.
- Haslanger, S. (2015). Social structure, narrative and explanation. *Canadian Journal of Philosophy*, 45(1), 1–15.
- Holroyd, J. (2012). Responsibility for implicit bias. *Journal of Social Philosophy*, 43(3), 274–306.
- Holroyd, J. (2015). Implicit bias, awareness and imperfect cognitions. *Consciousness and Cognition*, 33, 511–523.
- Kaufman, C. (2020). *Challenging power: Democracy and accountability in a fractured world*. Bloomsbury: Bloomsbury Publishing.
- Kutz, C. (2007). *Complicity: Ethics and law for a collective age*. Cambridge: Cambridge University Press.
- Lavin, C. (2008). *The politics of responsibility*. Illinois: University of Illinois Press.
- Lippert-Rasmussen, K. (2018). The ethics of anti-discrimination policies. (In A. Lever & A. Poama (Eds.), *The Routledge handbook of ethics and public policy* (pp. 267–280). New York: Routledge.)
- List, C. & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford: Oxford University Press.
- Ludwig, K. (2016). *From individual to plural agency: Collective action* (vol. 1). Oxford: Oxford University Press.
- Ludwig, K. (2017). *From plural to institutional agency: Collective action* (vol. 2). Oxford: Oxford University Press.
- Miller, D. (2001). Distributing responsibilities. *Journal of Political Philosophy*, 9(4), 453–471.
- Miller, S., & Makela, P. (2005). The collectivist approach to collective moral responsibility. *Metaphilosophy*, 36(5), 634–651.
- Nagel, T. (1973). Equal treatment and compensatory discrimination. *Philosophy & Public Affairs*, 2, 348–363.
- Nickel, J. W. (1972). Discrimination and morally relevant characteristics. *Analysis*, 32, 113–114.
- Nickel, J. W. (1974). Should reparations be to individuals or to groups? *Analysis*, 34, 154–160.
- Nunn, W. A. (1974). Reverse discrimination. *Analysis*, 34, 151–154.
- Rosenblum, D. (2007). Loving gender balance: Reframing identity-based inequality remedies. *Fordham Law Review*, 76, 2873.
- Sander, R. H. (2004). A systemic analysis of affirmative action in American law schools. *Stanford Law Review*, 57, 367.
- Saul, J. (2013). Implicit bias, stereotype threat, and women in philosophy. In K. Hutchison & F. Jenkins (Eds.), *Women in philosophy: What needs to change* (pp. 39–60). Oxford: Oxford University Press.
- Schiebinger, L. (2000). Has feminism changed science? *Signs: Journal of Women in Culture and Society*, 25, 1171–1175.
- Sher, G. (1975). Justifying reverse discrimination in employment. *Philosophy & Public Affairs*, 4, 159–170.
- Simon, R. L. (1978). Statistical justification of discrimination. *Analysis*, 38, 37–42.
- Singer, P. (1972). Famine, affluence, and morality. *Philosophy & Public Affairs*, 1, 229–243.
- Thomson, J. J. (1973). Preferential hiring. *Philosophy & Public Affairs*, 2, 364–384.
- Young, I. M. (2010). *Responsibility for justice*. Oxford: Oxford University Press.
- Zheng, R. (2018a). Bias, structure, and injustice: A reply to Haslanger. *Feminist Philosophy Quarterly*, 4(1), 1–29.
- Zheng, R. (2018b). What is my role in changing the system? A new model of responsibility for structural injustice. *Ethical Theory and Moral Practice*, 21(4), 869–885.
- Zheng, R. (2021). Moral criticism and structural injustice, *Mind*, 130(518), 503–535.