Can Affirmative Action Policies be Inefficiently Persistent?*

Philippe Jehiel†
Paris School of Economics
& University College London

Mathieu V. Leduc‡
Paris School of Economics

7 April, 2024

Abstract

We develop a dynamic model where successive, decentralized policy makers must decide whether to implement affirmative action policies aimed at improving the performance of future generations of a targeted group. Employers do not perfectly observe if a worker benefited from affirmative action, but take that possibility into account, resulting in the devaluation of the worker's credentials and an associated feeling of injustice. We establish that, in equilibrium, affirmative action is implemented *perpetually* by benevolent policy makers, despite the feeling of injustice that eventually dominates the anticipated benefits. This contrasts with the first best, which requires affirmative action to be temporary.

Keywords: Affirmative Action, General Equilibrium, Loss Aversion, Prospect Theory, Moral Hazard, Game Theory, Tragedy of the Commons

JEL codes: D40; I28; I30; J15

*We thank seminar participants at the Paris School of Economics, the Institut Henri Poincaré, Cambridge University, the Stockholm School of Economics and meetings of the Econometric Society and the Royal Economic Society, namely Gabrielle Demange, Francis Bloch, Frédéric Koessler, Olivier Tercieux, Mikhail Safronov, Xin Gao. We also thank Jörgen Weibull, Andrea Galeotti, Matthew L. Elliott and Larry Samuelson for useful comments. We are also grateful to David Levine, two anonymous referees and an anonymous advisory editor for very constructive comments when revising the article. Jehiel thanks the ERC (grant no 742816) for funding.

†Email: philippe.jehiel@gmail.com

‡(Corresponding author) Email: mattvleduc@gmail.com. Mailing address: Paris Jourdan Sciences Economiques, 48 boulevard Jourdan, 75014 Paris, France. Tel.: (+33) 01 80 52 16 60

1 Introduction

The original rationale for affirmative action was to help underrepresented groups close achievement gaps and such policies were often anticipated to be temporary. Decades after their inception, affirmative action policies however often remain in place. Such policies have generated a deep interest and the literature spawned over the past decades is large. Some work argues that such policies can be used to attempt to close achievement gaps, while other work emphasizes certain inefficiencies mostly related to market distortions, such as mismatches between workers and jobs, and related consequences on productivity.

A form of inefficiency that has gathered less attention is the devaluing effect an affirmative action policy can have on the perception of a worker's curriculum vitae. Indeed, when an affirmative action policy is in place, the mere possibility that a worker may have benefited from it can have a devaluing effect on his diplomas. If the quality of a worker's curriculum vitae is perceived by employers to be lower than his actual skills level, the worker can then experience a stigma or a feeling of injustice.

The current article will thus attempt to provide a novel explanation for the apparent stickiness of affirmative action policies in a model that takes explicitly into account their effect on devaluing the perception of a worker's curriculum vitae.

In our approach, we will consider a decentralized setting in which successive policy makers in many different districts have to decide whether or not to implement affirmative action policies. Each district's population is composed of a group A (the main group) and a group B (the group targeted by the affirmative action policy). The policy makers can be thought of as local government representatives or as school or university managers in charge of choosing which pupils or students to admit. Moreoever, the policy maker of a given district anticipates that an affirmative action policy improves the talent distribution of group B in future periods in that district. This belief, held by the policy maker, is in line with popular role model theories (see, Chung (2000) or Beaman et al. (2012) for, respectively, a theoretical and empirical exploration of the role model effect), according to which witnessing certain members of an underrepresented group achieving success would lead other group members to achieve higher success in the future. After pupils/students have completed their schooling/university period, they enter the labor market, which is assumed to mix pupils/students from all districts.

The successive policy makers are assumed to be benevolent and we study their incentives to implement affirmative action policies during their tenure. We do so using a repeated game setting, with each successive localized policy maker seeking to maximize its local welfare.

In the main part of the paper, we will suppose that employers cannot condition wages on group identity. Although it is not necessary for our results to hold, it is in line with many anti-discrimination policies (for instance, direct discrimination that would allow conditioning wages on group identity, such as gender, is illegal in the U.S. since the passage of the Equal Pay Act in 1963).[1] In a perfectly competitive labor market, each employer pays a worker a wage equal to his expected performance given the district he comes from. The employer does not observe whether the worker benefited from

[1] We are aware that some studies have highlighted the presence of wage gaps across groups (see, for example, the OECD charts on the gender earning gap). However, for our purpose, the real question is whether there is a gap conditioning on the curriculum vitae (which is more debated). As employers may, in some cases, find ways to get around the legal constraints forbidding discrimination, reality most likely lies somewhere in between the no-discimination environment—studied in the main model—and the free discrimination environment—analyzed in the discussion section (Section 5.1.1). As we will see, our insights apply in both cases.

affirmative action or not and can only estimate this performance based on a curriculum vitae (which may be artificially improved by affirmative action), as well as some aggregate statistics describing the average level of affirmative action policy implemented over a range of districts. Paying workers a wage equal to their expected performance thus means that non-beneficiaries of affirmative action will get a wage below their true performance level. These non-beneficiaries can include members of both groups A and B since the affirmative action policy typically does not reach all members of the target group B. We postulate that in such a case, the worker suffers from a *feeling of injustice* that is proportional to the difference between his true performance (which the worker is assumed to know) and his wage.

Although our model is stylized, recall that this depressed wage can be understood, more broadly, as being associated with the devaluation of a worker's diplomas (or even career promotions), which results from the mere possibility that he may have benefited from affirmative action.[2] We believe such a feeling of injustice is very common. In the case of group B, this feeling can often be associated with the stigmatization felt by workers who did not benefit from affirmative action (or in more practical situations, even by those who did not need the policy in order to be accepted in a school or university), but are yet underrated due to the mere possibility that some members of their group may have benefited from the policy. In the case of group A, this feeling of injustice is also in line with not being favored by the policy.[3]

In a first-best scenario, this depressed wage given to non-beneficiaries means that affirmative action should not last permanently. The optimal duration would be determined by a number of parameters, namely by the weights in welfare assigned to members of the main and the targeted groups, the costs associated to the feeling of injustice, etc. However, affirmative action would necessarily be ended at some point, as long as non-beneficiaries suffer some (even very small) feeling of injustice in the long-term. This is so because after sufficently many implementations of affirmative action policies, the benefit of one more implementation (through the role model channel) would become vanishingly small whereas the cost of it (that results from the induced feeling of injustice) would remain significant, no matter how long such policies have been in place.

To the contrary, the unique equilibrium is such that decentralized policymakers choose to implement affirmative action policies on a *permanent* basis. The intuition is that, in our setting, affirmative action policies are not observed district by district by employers. Thus, if a policy maker were to not implement an affirmative action policy in some period and in some district, this policy maker could deviate without being observed, implement the policy, and this would have no effect on depressing wages. Since such a deviation would be believed to improve the future performance distribution of the targeted group (through a role model argument), the policy maker would do it, thereby showing that affirmative action policies are perpetually implemented in all districts. In other words, the non-transparency of the affirmative action policies creates a moral hazard environment, by which each policy maker necessarily chooses to implement an affirmative action policy and fails to internalize the effect that it has on devaluing diplomas (and thus on depressing wages).

[2] For example, Reardon (2021) reports that women publishing in elite medical journals are half as likely to be cited than their male counterparts. While there could be several explanations for this, one of them (based on the premise that those citing do not always have the full ability to assess the quality of the publications, relying instead on more accessible characteristics including the authors' gender) could be related to the devaluing effect of affirmative action policies.

[3] It is also in line with the policy being perceived as decreasing the quality of diplomas through other channels, such as lowering academic standards by relaxing entrance requirements.

We believe that our assumption—that the policies chosen by the policy makers are not observed precisely in the labor market—is justified when affirmative action decisions are implemented at a decentralized level, as considered in our model. Indeed, it is often very difficult in practice to determine whether a specific policy maker actually implemented an affirmative action policy or not. For example, in the United States, these policies are complex, they vary from state to state, even from school to school, and when they are not officially implemented, they may actually take place through private channels (e.g. non-governmental diversity enhancement programs, etc.). The large number of decentralized policymakers, along with their imperfectly observed actions, thus lead to a tragedy of the commons phenomenon, whereby they overuse the policy and disregard its negative reputational effects on workers.

The paper is organized as follows. In Section 2, we introduce the basic setting and define the workers' utilities and welfare. In Section 3, we study how employers set the wages they pay to workers and show that it leads to a feeling of injustice felt by non-beneficiaries of affirmative action (of both groups A and B). In Section 4, we analyze each policy maker's welfare maximization problem and present the two central results: (i) perpetual affirmative action as an equilibrium policy and (ii) the first-best policy in which affirmative action is ultimately ended. In Section 5, we discuss how our assumptions can be relaxed, as well as model extensions. We also compare our model with the existing literature. Proofs are relegated to Section 6. A supplementary appendix in Section 7 extends our model to a more general setting allowing for strategic behavior on the workers' side.

2 Setting

There is a continuum J of districts (or jurisdictions), indexed by j, where j is uniformly distributed on $(0, \overline{J})$. At each time $t \in \mathbb{N}$, district policy makers must each decide whether to implement an affirmative action policy in their district for the duration of their tenure (one period). That is, the policy maker of district $j \in J$ chooses an action $\sigma_t^j \in \{0, 1\}$, where $\sigma_t^j = 0$ corresponds to no affirmative action and $\sigma_t^j = 1$ corresponds to affirmative action. One can think of policy makers as local government representatives or as private authorities such as school principals. In the following, we will be assuming that policy makers' interests are aligned with total welfare so that the inefficiencies we highlight cannot be attributed to conflicts of interests.

In each district, a population of workers consists of two groups: group A^j (the main group) and group B^j (the targeted group). A worker has a performance level $c \in [0, 1]$. This can be understood as his intrinsic productivity. A worker also has a curriculum vitae of quality $\overline{c} \in [0, 1]$, which is the potentially upward-biased signal about c that employers will observe.

At any time t, group A^j's performance density is $f_{A^j}(c)$ while group B^j's performance[4] density is $f_{B^j, n_t^j}(c)$, where $n_t^j = \sum_{s<t} \sigma_s^j$ is the number of times previous policy makers have implemented affirmative action policies in district j. $f_{A^j}(c)$ and $f_{B^j, n_t^j}(c)$ have support $[0, 1]$ and are non-degenerate. We will describe later how $f_{B^j, n_t^j}(c)$ varies with n_t^j but intuitively as n_t^j increases, $f_{B^j, n_t^j}(c)$ shifts lower values of c to higher values, resulting in first-order stochastic dominance. Each agent lives for only one period. At each time t, a mass $|A^j|$ and a mass $|B^j|$ of new workers from groups A^j and

[4]In the applications we will have in mind, it is reasonable to think that group A^j's performance distribution initially differs from that of group B^j, although this plays no role in our analysis. For instance, one could assume that initially $f_{B^j, n_0^j}(c) \prec f_{A^j}(c)$, where $\prec$ denotes first-order stochastic dominance, which could be used as a justification for allowing an affirmative action policy.

B^j respectively are born in district j to replace[5] the ones that have expired, with performance levels drawn according to $f_{A^j}(c)$ and $f_{B^j,n_t^j}(c)$, respectively.

2.1 Effect of affirmative action policy

An affirmative action policy has two effects. First, it gives an immediate artificial boost to the curriculum vitae of a worker benefiting from it. This models the fact that a beneficiary of affirmative action has expanded opportunities in terms of education (university admissions or other professional formations) compared to a non-beneficiary, thereby artificially enhancing the quality of his curriculum vitae. Second, it is also believed to have long-term, positive effects on the performance distribution of group B^j. This anticipated long-term effect is in line with popular role model theories (e.g. Chung (2000) and Beaman et al. (2012)). This second effect will be captured by the dependence of $f_{B^j,n_t^j}(c)$ on n_t^j.

It is important to note that an affirmative action policy can be interpreted[6] as anything that artificially increases the quality of a curriculum vitae (immediate effect) and is believed to improve the performance distribution of future generations (anticipated role model effect). To keep the exposition simple, we will often suppose that affirmative action takes place at the university level and that it grants certain people easier access to diplomas. Another interpretation, of course, is that affirmative action takes place at the workplace level and gives certain people easier access to promotions.

In a given period t where it is implemented, we will allow the affirmative action policy to only reach a fraction $\xi \in (0, 1]$ of the members of the targeted group B^j. Indeed, in practice, not all members of a targeted group may benefit from the policy[7] (furthermore, some members of the targeted group may not even be aware that such a policy is in place). While we believe this is a fairly plausible assumption, we note that our main results apply even when $\xi = 1$. The only case where we need the stronger requirement that $\xi < 1$ is when we allow for free discrimination in the labor market, i.e. allowing wages to also depend on the group to which the worker belongs.

2.1.1 Effect of affirmative action policy on curriculum vitae quality

When $\sigma_t^j = 1$, with probability $\xi \in (0, 1]$, a member of group B^j with performance level $c \in [0, 1]$ will have a curriculum vitae quality $\bar{c} = g(c)$, where g is an increasing function such that $g(c) > c$, $\forall c \in (0, 1)$, and $g(0) = 0$, $g(1) = 1$. The support of $\bar{c}$ is thus also $[0, 1]$. With probability $1 - \xi$, a member of group B^j with performance level c will have a curriculum vitae quality corresponding to his actual performance level: $\bar{c} = c$.

When $\sigma_t^j = 0$, a member of group B^j with performance level c will have a curriculum vitae quality corresponding to his actual performance level: $\bar{c} = c$.

Whether $\sigma_t^j = 0$ or 1, a member of group A^j with performance level c always has a curriculum vitae quality corresponding to his actual performance level: $\bar{c} = c$.

[5] The model can be extended to allow workers to live for more than one period and to have overlapping generations. On the other hand, as it will become clear, all the decisions in this model are made by policy makers and employers. Workers are essentially passive and thus their lifetime has no direct role in the analysis. For that reason, we have chosen the simpler setting in which workers live for only one period.

[6] See Section 5.1.4 for a discussion of how our model can accommodate even more general interpretations of affirmative action.

[7] See also Fryer Jr and Loury (2005) or Fershtman and Pavan (2021) for further discussion along this line.

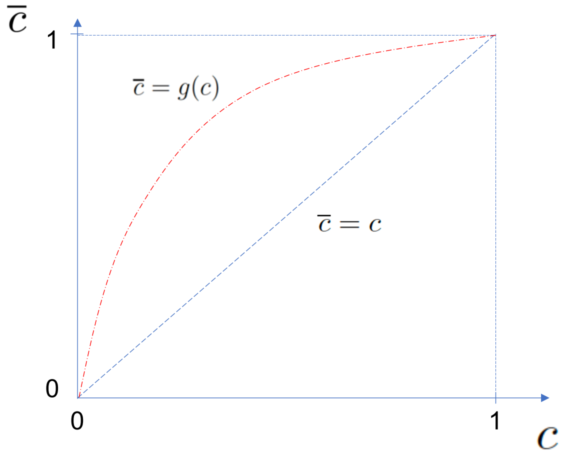

Figure 1: Effect of affirmative action on curriculum vitae quality $\overline{c}$. A beneficiary of affirmative action has curriculum vitae quality higher than his actual performance level: $\overline{c} = g(c) > c$ (red curve). A non-beneficiary has curriculum vitae quality corresponding to his actual performance level: $\overline{c} = c$ (blue line).

An affirmative action policy therefore increases the curriculum vitae quality of a beneficiary above his actual performance level, while it has no effect on the curriculum vitae quality of members of group A^j nor on those of members of group B^j who did not benefit from the affirmative action policy. That is, their curriculum vitae quality corresponds to their actual performance level. This is illustrated in Figure 1.

2.1.2 Anticipated effect of affirmative action policy on actual performance

Agents in this model believe that implementing an affirmative action policy has an improving impact on the performance distribution of future cohorts of B^j workers. Thus, we suppose that if $\sigma_t^j = 1$, then the next period's performance distribution of group B^j is shifted so that $f_{B^j,n_{t+1}^j}(c) \succ f_{B^j,n_t^j}(c)$, where $\succ$ indicates strong first-order stochastic dominance. Note that the effect of the shift is permanent, i.e. the improvement remains in all future periods.[8] This purported improvement in the performance of future cohorts of workers is consistent with the role model argument.

If $\sigma_t^j = 1$ for all t, then $f_{B^j,n_t^j}(c) \uparrow \overline{f}_{B^j}(c)$. Since $f_{B^j,n_t^j}(c)$ converges from below to a limiting distribution $\overline{f}_{B^j}(c)$, this implies that the distributional improvements become smaller and smaller as policy makers keep implementing affirmative action policies. Group A^j's performance distribution $f_{A^j}(c)$ does not vary with t. This is illustrated in Figure 2. Observe that the densities $f_{B^j,n_t^j}(c)$ could depend on ξ, as the larger ξ the more individuals in group B^j are likely to be exposed to the effect of the affirmative action policy in district j. Since our results do not rely on varying ξ, we omit an explicit reference to this dependence.

[8] While the empirical validity of this assumption may be debated, we make it in order to give the best chance to the affirmative action policy and then emphasize other side effects stemming from that policy. For empirical work consistent with this assumption, see for instance Antonovics and Backes (2014).

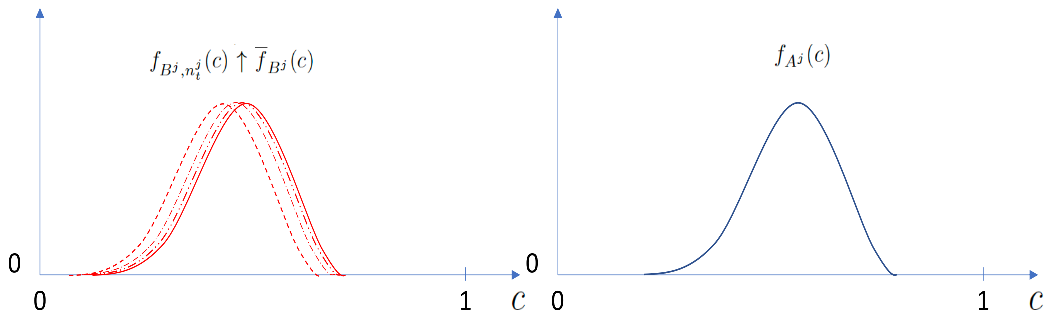

Figure 2: Anticipated effect of affirmative action policy on actual performance. If $\sigma_t^j = 1$, then the next period's performance distribution of group B^j is shifted so that $f_{B^j,n_{t+1}^j}(c) \succ f_{B^j,n_t^j}(c)$. If $\sigma_t^j = 1$ for all t, then $f_{B^j,n_t^j}(c)$ converges to a limiting distribution $f_{B^j,n_t^j}(c) \uparrow \overline{f}_{B^j}(c)$. Group A^j's performance distribution $f_{A^j}(c)$ is not affected.

2.2 Utilities and welfare

A worker is of type $\theta = (c, \bar{c}, G^j)$, where c is his true performance level, $\bar{c}$ is his curriculum vitae quality and $G^j \in \{A^j, B^j\}$ is the group G this worker belongs to and the district j he comes from. A time t worker knows his type and the wage function $\omega_t^j(\bar{c})$ set by employers, which is the wage the worker earns based on the information on his curriculum vitae (i.e. the curriculum vitae quality $\bar{c}$ and the district j the worker comes from).[9] This is formalized in the following definition.

Definition 1 *A wage function $\omega_t^j : [0,1] \to [0,1]$ determines, at time t, the wage a worker coming from district j earns when presenting a curriculum vitae of quality $\bar{c}$ to an employer.*

Note here that we chose not to allow employers to condition wages on the group A or B to which a worker belongs. This is motivated on grounds that such group-based discrimination is in general forbidden.[10] Our results are however robust to conditioning wages on group identity, i.e. giving a wage $\omega_t^{G^j}$ instead of ω_t^j. This is further discussed in section 5.1.1.

2.2.1 Utility

The utility of a type $(c, \bar{c}, G^j)$ worker at time t is

$$u_{G^j,t}(\bar{c}, c) = \omega_t^j(\bar{c}) - \gamma_{G^j} \max\{c - \omega_t^j(\bar{c}), 0\} \tag{1}$$

where $\gamma_{G^j} \max\{c - \omega_t^j(\bar{c}), 0\}$, for some $\gamma_{G^j} > 0$, captures the fact that a feeling of "injustice" is suffered when a worker gets a salary that is below his true performance level. Note that we allow $\gamma_{A^j} \neq \gamma_{B^j}$ so as to capture that the feeling of injustice may differently affect groups A and B in district j.

[9] If workers were to live several periods, we could envision a more elaborate model in which the wage earned in later periods would also depend on the true performance assumed to be partly observed then. Our qualitative insights would be unaffected.

[10] For instance, direct discrimination that would allow conditioning wages on group identity, such as gender, is illegal in the U.S. since the passage of the Equal Pay Act in 1963. Note, however, that even with a wage function ω_t^j that is common across groups, the average wage paid to group A and to group B may differ and hence a wage gap can remain between the groups. Our model is thus perfectly compatible with a wage gap existing between groups.

In particular, the utility of a type $(c, \overline{c}, G^j)$ worker who benefits from affirmative action has the form

$$u_{G^j,t}(\overline{c}, c) \quad = \quad \omega_t^j(g(c)) - \gamma_{G^j} \max\{c - \omega_t^j(g(c)), 0\}$$

since $\overline{c} = g(c)$, while the utility of a type $(c, \overline{c}, G^j)$ worker who does not benefit from affirmative action has the form

$$u_{G^j,t}(\overline{c}, c) \quad = \quad \omega_t^j(c) - \gamma_{G^j} \max\{c - \omega_t^j(c), 0\}$$

since $\overline{c} = c$. We will often denote by $u_{B^j,t}(g(c), c)$ (respectively, by $u_{B^j,t}(c, c)$) the utility of a group B^j worker benefiting (respectively, not benefiting) from affirmative action, while we will denote by $u_{A^j,t}(c, c)$ the utility of a group A^j worker.

In the above, we assume that workers have the correct perception[11] of their performance level c. We also note that there are no extra positive effects on utility of receiving a wage greater than the performance level. Such an asymmetry in the utility assessment of wages above or below the performance level is in line with well documented psychological studies (see in particular the prospect theory of Kahneman and Tversky (1979)), which suggest a different assessment for payoff realizations above or below the reference point (here naturally identified with the performance level).

2.2.2 Profit

Each employer produces a numeraire good of price equal to 1 with a constant return to scale technology and using labor as the input. The quantity of the numeraire good produced by a worker of performance level c is thus simply c. The profit generated by a district-j worker of curriculum vitae quality $\overline{c}$ and performance level c at time t, when he is paid a wage $\omega_t^j(\overline{c})$, is thus

$$\pi_t(\overline{c}, c) = c - \omega_t^j(\overline{c}).$$

2.2.3 Welfare

The welfare of each group in district j at time t is defined by taking the aggregate utility of that group. We thus have,

$$W_{A^j,t} = |A^j| \int_0^1 u_{A^j,t}(c, c) f_{A^j}(c) dc$$

$$W_{B^j,t} = |B^j| \int_0^1 \Big(\xi\sigma_t^j u_{B^j,t}(g(c), c) + (1 - \xi\sigma_t^j) u_{B^j,t}(c, c)\Big) f_{B^j, n_t^j}(c) dc$$

where σ_t^j is the actual policy decision made by the time t policy maker of district j.

[11] More realistically, we could assume that the worker only has a signal about his labor market performance c, which could be derived from his performance in school and maybe some other extracurricular experiences. The performance, as considered in our model, could then be thought of as the expected labor market performance conditional on the signal. Given the monotonic relationship between the signal and this expected performance, redefining the variable of interest to be the expected labor market performance would lead to the same model.

Likewise,

$$\Pi_{F^j,t} = |A^j| \int_0^1 \pi_t(c,c) f_{A^j}(c) dc + |B^j| \int_0^1 \Big(\xi \sigma_t^j \pi_t(g(c),c) + (1 - \xi \sigma_t^j) \pi_t(c,c)\Big) f_{B^j, n_t^j}(c) dc$$

is the total profit of the employers (i.e. firms) resulting from the labor provided by workers coming from district j.

Total welfare of district j at time t is then defined by

$$W_t^j = W_{A^j,t} + \lambda_{B^j} W_{B^j,t} + \lambda_{F^j} \Pi_{F^j,t}$$

where the weight on B^j's welfare, λ_{B^j}, is non-negative, while the weight placed on the employers' profits, λ_{F^j}, is also non-negative. The case where $\lambda_{B^j} = \lambda_{F^j} = 1$ corresponds to the standard total welfare criterion. $\lambda_{B^j} < 1$ reflects a preference for the main group A^j in the policy maker's objective, while $\lambda_{B^j} > 1$ reflects a preference for the targeted group B^j. We will generally concentrate on the case where $\lambda_{F^j} \leq \lambda_{B^j} \leq 1$.

Letting δ denote the common discount factor, total welfare of district j over all periods is then defined by

$$W^j = \sum_{t=1}^{\infty} \delta^t W_t^j$$

and total welfare in the economy is defined by

$$W = \int_{j \in J} W^j dj.$$

3 Effect of affirmative action policy on wage levels

We model a non-localized labor market, where workers educated in all districts match freely with employers and are paid wages by the latter.

3.1 Informational environment

While it is plausible to assume that employers observe some aggregate statistics about the decentralized affirmative action policy decisions, we believe that in many applications it is natural to assume that employers do not observe each σ_t^j separately. Indeed, in the face of a large number of districts, it would be very difficult to keep track of all decentralized policy decisions.

To formalize this idea most simply, we assume that employers at time t can observe an aggregate statistic $\overline{\sigma}_t$ of all policy decisions made by the different districts. Here we take $\overline{\sigma}_t$ to be simply the average policy across all districts, that is

$$\overline{\sigma}_t = \frac{1}{|J|} \int_{j \in J} \sigma_t^j dj. \tag{2}$$

Thus, they know the sequence $\{\overline{\sigma}_s\}_{s=1}^t$ of all average policy decisions made over time (up to time t).

Not knowing for sure whether affirmative action took place in a particular district, they may not be able to tell for certain whether a worker benefited from affirmative action or not. They can however compute the probability that a worker benefited from affirmative action, conditional upon observing his curriculum vitae quality, the district this worker comes from (e.g., where he graduated school or university), the aggregate policy statistics and considering a conjectured strategy played by policy makers.

Note that the form of the observed aggregate statistic in Eq. (2) can be generalized. In our model, the observation of the aggregate statistic—whatever its exact from—plays no role in the wage determination, nor in the equilibrium analysis describing policy makers' strategies. For example, an employer could observe a more localized average of the policies practiced around district j, such as $\overline{\sigma}_t^j = \frac{1}{2\epsilon}\int_{i=j-\epsilon}^{j+\epsilon} \sigma_t^i di$. What is key is that no inference can be made from the observed statistics about the value of a given σ_t^j, which sounds plausible when the number of districts is very large. We discuss this further in Section 5.1.2.

3.2 Setting wages

We consider a perfectly competitive labor market, where an employer pays a worker a wage equal to his expected performance level. In Section 7.1, this reduced-form approach is micro-founded based on a Bertrand-type model of competition between employers.

As mentioned earlier, we assume that employers are not allowed to take group information (A or B) into account when giving a wage to a particular worker. This is consistent with anti-discrimination laws enacted in many countries and occupational areas (although it is not necessary for our results to hold, as previously mentioned and as discussed in Section 5.1.1). Thus, they set a wage conditioned only on the curriculum vitae quality $\overline{c}$, the district j a worker is from, the observed sequence of aggregate policy statistics $\{\overline{\sigma}_s\}_{s=1}^t$ and a given conjectured policy sequence assumed by them $\sigma = \{\{\sigma_s^j\}_{j\in J}\}_{s=1}^\infty$. The wage $\omega_t^j(\overline{c})$ paid to a worker of type $(c, \overline{c}, A^j)$ or to a worker of type $(c, \overline{c}, B^j)$ is thus the conditional expectation $\mathbb{E}_t[c|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$ of the worker's true performance level c, expressed in the following lemma.

Lemma 1 (Wage function) *Given some conjectured policy sequence $\sigma = \{\{\sigma_s^j\}_{j\in J}\}_{s=1}^\infty$ and an observation of aggregate policy statistics $\{\overline{\sigma}_s\}_{s=1}^t$ consistent with σ, the wage paid at time t to a worker with curriculum vitae quality $\overline{c}$ coming from district j has the form*

$$\omega_t^j(\overline{c}) = \mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma) \cdot g^{-1}(\overline{c}) + \big(1 - \mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma)\big) \cdot \overline{c}$$

where

$$\mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma) = \frac{|B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c})}{|A^j|f_{A^j}(\overline{c}) + |B^j|(1-\xi\sigma_t^j)f_{B^j,n_t^j}(\overline{c}) + |B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c})}.$$

In the above, $\{aa\}$ is the event that a worker benefited from affirmative action, $(g^{-1})'(\overline{c})$ is the derivative of g^{-1} evaluated at $\overline{c}$ and n_t^j is required to be consistent with $\{\sigma_s^j\}_{s<t}$.

In words, $\omega_t^j(\overline{c})$ is a convex combination between $g^{-1}(\overline{c})$ and $\overline{c}$, where the weight assigned to $g^{-1}(\overline{c})$ is the probability that a worker with curriculum vitae $\overline{c}$ and coming from district j benefited from affirmative action at time t (taking into account the policy sequence σ believed to be followed by policy makers, hence the expression for $\mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma)$).

We now make the following assumption to ensure that a higher curriculum vitae quality is always associated with a (weakly) higher expected performance level and thus that the wage $\omega_t^j(\overline{c})$

is non-decreasing in $\overline{c}$. We do so for simplicity of exposition as it allows us to rule out strategic behavior by which a worker could present a curriculum vitae of lower[12] quality than $\overline{c}$. An extension where a worker is allowed to present a curriculum vitae of a different quality than $\overline{c}$ is presented in a supplementary appendix (Section 7.2), where the robustness of our results to such strategic behavior is established in a more general context, and which thus removes the need for Assumption 1.

Assumption 1 *(i) For any n_t^j, the likelihood ratio $\frac{f_{A^j}(\overline{c})}{f_{B^j,n_t^j}(g^{-1}(\overline{c}))}$ is non-decreasing in $\overline{c}$. (ii) The curriculum vitae enhancing function $g(c)$ is concave in c.*

Lemma 2 (Non-decreasing wage function) *There exists $\underline{\xi} \in (0,1)$ such that when $\xi > \underline{\xi}$ and when Assumption 1 holds, the wage function $\omega_t^j(\overline{c})$ is non-decreasing in $\overline{c}$.*

We will thus suppose, in the main part of the paper, that the conditions of Lemma 2 hold.

Relying on the expression of equilibrium wage derived in Lemma 1, we note that whether the earned wage lies above or below the performance level solely depends on whether or not the worker benefited from affirmative action:

Lemma 3 (Wage versus performance level)

(i) Suppose $\sigma_t^j = 1$. Then any district-j worker gets a wage lower than his curriculum vitae quality (i.e. $\overline{c} > \omega_t^j(\overline{c})$). Moreover, a worker benefiting from affirmative action gets a wage higher than his performance level (i.e. $c = g^{-1}(\overline{c}) < \omega_t^j(\overline{c})$), while a worker not benefiting from affirmative action gets a wage lower than his performance level (i.e. $c = \overline{c} > \omega_t^j(\overline{c})$).

(ii) Suppose $\sigma_t^j = 0$. Then any district-j worker gets a wage equal to his curriculum vitae quality and his performance level (i.e. $c = \overline{c} = \omega_t^j(\overline{c})$).

This lemma is illustrated in Figure 3.

3.3 Feeling of injustice and broader interpretation of the depressed wage

In our model, the wage is depressed due to the possibility that a worker benefited from affirmative action. This represents the fact that a certain curriculum vitae quality is, in expectation, no longer associated with the same performance level as if there were no affirmative action policy. Indeed, an affirmative action policy has the effect of devaluing the diplomas or promotions that figure on a worker's curriculum vitae, if there is only some chance that the worker may have benefited from such a policy.

Using Lemma 3, we now make the following observation.

Observation 1 (Feeling of injustice)

(i) Suppose $\sigma_t^j = 1$.

A worker not benefiting *from affirmative action gets a wage lower than his performance level and* suffers *a feeling of injustice. His utility is then*

$$u_{G^j,t}(c,c) \quad = \quad \omega_t^j(c) - \gamma_{G^j}(c - \omega_t^j(c)).$$

[12] Indeed, if the wage function $\omega_t^j(\overline{c}) = \mathbb{E}_t[c|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$ is decreasing on some parts of the support $[0,1]$, a worker could earn a higher wage by presenting a curriculum vitae of lower quality than $\overline{c}$.

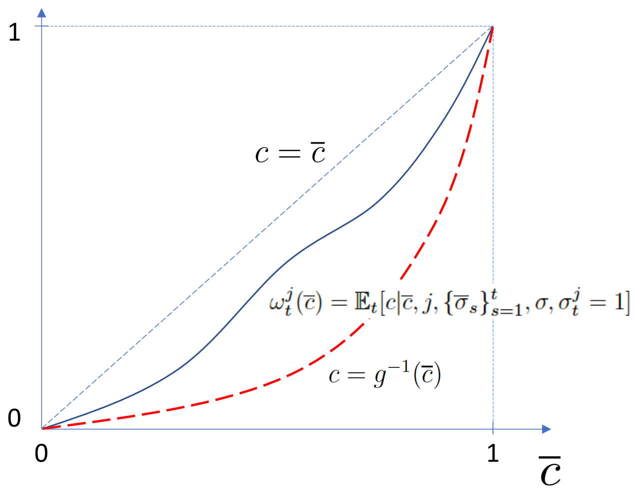

Figure 3: Illustration of Lemma 3. The curriculum vitae quality $\overline{c}$ is on the horizontal axis. (i) When affirmative action is implemented ($\sigma_t^j = 1$), we see on the vertical axis that the wage $\omega_t^j(\overline{c}) = \mathbb{E}_t[c|\overline{c}, j, \{\sigma_s\}_{s=1}^t, \sigma, \sigma_t^j = 1] < \overline{c}$ (full blue curve) is lower than the performance level of a non-beneficiary (thinly dotted blue line) and higher than the performance level of a beneficiary (thickly dotted red curve). (ii) When no affirmative action is implemented ($\sigma_t^j = 0$), then the wage $\omega_t^j(\overline{c}) = \mathbb{E}_t[c|\overline{c}, j, \{\sigma_s\}_{s=1}^t, \sigma, \sigma_t^j = 0] = \overline{c}$ corresponds to the performance level of any worker (i.e. thinly dotted blue line).

By contrast, a worker benefiting *from affirmative action gets a wage higher than his performance level and* does not suffer *a feeling of injustice. His utility is then*

$$u_{G^j,t}(g(c), c) \quad = \quad \omega_t^j(g(c)).$$

(ii) Suppose $\sigma_t^j = 0$.

A worker gets a wage equal to his performance level and does not suffer *a feeling of injustice. His utility is then*

$$u_{G^j,t}(c, c) \quad = \quad c.$$

It is important to emphasize that workers from both groups (A and B) can experience a feeling of injustice. In the case of group B, this feeling can often be associated with the stigmatization felt by workers who did not benefit from affirmative action (or in more practical situations, even those who did not need it in order to be accepted in a school or university), but are yet underrated due to the mere possibility that some members of their group may have benefited from the policy. This is suggested by a good deal of empirical evidence (see, for example, Leslie et al. (2014), Heilman et al. (1997) or Heilman et al. (1992)) and some theoretical work (e.g. Kim and Loury (2018)).

3.4 Wage gaps across groups

It is important to note that even if employers do not condition the wage function $\omega_t^j(\overline{c})$ on group identity (A^j or B^j), a wage gap can be sustained across groups.

Indeed, even if employers must choose a common wage function $\omega_t^j(\bar{c})$, the actual *distributions* of wages that are given to workers of different groups can differ, as they are driven by the distributions $f_{A^j}(c)$ and $f_{B^j,n_t^j}(c)$. Indeed a group A^j worker receives a wage $\omega_t^j(c)$, whereas a group B^j worker receives a wage $\omega_t^j(c)$ if he did not benefit from the policy or a wage $\omega_t^j(g(c))$ if he did. We see how the wage distributions are driven by the c distributions and hence, the *actual* wages are allowed to differ across groups even if the wage *function* ω_t^j itself does not. Our model is thus compatible with a wage gap existing between groups.

Our model is agnostic as to the origins of the initial differences in $f_{A^j}(c)$ and $f_{B^j,n_t^j}(c)$. However, such differences could be the result of lower incentives to invest in skills acquisition in the past—which could have been driven by mechanisms inspired by the statistical discrimination literature (e.g. Phelps (1972), Arrow (1973), Lundberg and Startz (1983), Coate and Loury (1993a), Coate and Loury (1993b) or Echenique and Li (2022)) or even by taste-based mechanisms (e.g. Becker (1957))—and could have had a persistent impact on performance distributions.

Since our aim is to emphasize a mechanism for the devaluation of curricula vitae in the eyes of employers, we have chosen to leave such aspects out of our model, although they are not incompatible with it.

4 The policy makers' decision problem

4.1 Equilibrium policy

4.1.1 Main result

At any $t \geq 1$, a district-j policy maker wants to choose a policy σ_t^j in order to maximize the following objective function:

$$\max_{\sigma_t^j \in \{0,1\}} \quad \sum_{s=t}^{\infty} \delta^{s-t}(W_{A^j,s} + \lambda_{B^j} W_{B^j,s} + \lambda_{F^j} \Pi_{F^j,s}). \tag{3}$$

That is, we assume that the objective of the district-j policy maker coincides with the local welfare, defined in Section 2.2.3, aggregated over the remaining time periods.

Given some conjectured policy sequence $\sigma = \{\{\sigma_s^j\}_{j\in J}\}_{s=1}^{\infty}$ followed by policy makers across time, a district-j policy maker is able to compute $W_{A^j,s}$, $W_{B^j,s}$ and $\Pi_{F^j,s}$ for $s > t$, where $\omega_s^j(\bar{c})$ and $f_{B^j,n_s^j}(c)$ are taken to be consistent with σ.

Our first main result, Proposition 1, states that all district policy makers choosing to *perpetually* implement an affirmative action policy is the unique equilibrium.

Proposition 1 (Permanent affirmative action in equilibrium) *Let $\lambda_{B^j} > 0$ and $\lambda_{B^j} \geq \lambda_{F^j}$. Then there exists $\overline{\gamma}_{B^j}$ (or likewise $\underline{\xi}$) such that for any $\gamma_{B^j} < \overline{\gamma}_{B^j}$ (or likewise $\xi > \underline{\xi}$) the unique equilibrium is $\sigma_t^{j*} = 1$ for all t and j.*

The intuition behind Proposition 1 is that any district-j policy maker sees implementing an affirmative action policy as increasing the welfare of group B^j in two ways. To see this, denote the equilibrium wage function by ω_s^{j*}. First, implementing the policy at time t inflates the curricula vitae of group B^j members benefiting from the policy, which results in them receiving higher wages

$\omega_t^{j*}(g(c))$ as opposed to $\omega_t^{j*}(c)$. Second, the district-j policy maker believes that implementing an affirmative action policy improves the performance distribution $f_{B^j,n_s^j}(c)$ of future cohorts of B^j workers (at times $s > t$), which will thus also result in them receiving higher wages in the future by shifting the performance c towards higher values. Thus, the only reason the policy maker would choose not to implement such a policy would be to have an uplifting effect on the actual equilibrium wage function ω_t^{j*} chosen by employers (which, as we know from Lemmas 1 and 3, is depressed by the possibility that a worker has benefited from affirmative action). However, due to its small size (measure zero), a district-j policy maker cannot have any impact on the aggregate (average) statistic $\overline{\sigma}_t$, which is the only policy information actually "observed" by employers when setting wages. Therefore, there is no reason why a particular policy maker would deviate, by choosing $\sigma_t^j = 0$, from an equilibrium policy $\sigma_t^{j*} = 1$ in which it implements an affirmative action policy. Conversely, a deviation from a putative equilibrium in which $\sigma_t^{j*} = 0$ to $\sigma_t^j = 1$ would inflate the curricula vitae of B^j workers at time t and would be believed to improve the average performance of future cohorts of B^j workers at future times, without having a worsening impact on the wage function, since that deviation would not be reflected in the aggregate statistic $\overline{\sigma}_t$ observed by employers and thus on the wage function ω_t^{j*}. This establishes $\sigma_t^{j*} = 1$ as the unique equilibrium.

The intuition behind the sufficient (and not always necessary) conditions $\gamma_{B^j} < \overline{\gamma}_{B^j}$ (or likewise $\xi > \underline{\xi}$) is that although a deviation from a putative equilibrium in which $\sigma_t^{j*} = 0$ to $\sigma_t^j = 1$ would increase the average performance of future cohorts of B^j workers, it could also potentially increase the average feeling of injustice felt by B^j workers not benefiting from affirmative action in future periods. Indeed, the feeling of injustice could worsen following an increase in the performance level, if the latter increases faster than the wage received at a higher performance level. Recalling that the feeling of injustice is $\gamma_{B^j} \max\{c - \omega_s^{j*}(c), 0\}$, then if in some future periods $\frac{dc}{dc} = 1 > \frac{d\omega_s^{j*}(c)}{dc}$, increasing c could create more feeling of injustice among B^j workers since their performance would increase faster than their wage. As $\omega_s^{j*}(c)$ is endogenously determined, we can ascertain that this effect is of second order with sufficient conditions on exogenous parameters γ_{B^j} (or ξ). Namely, that the parameter γ_{B^j} be small enough[13] (or likewise that ξ be large enough, since then there would be a small enough fraction of B^j workers failing to benefit from affirmative action and thus experiencing the feeling of injustice).

We will now comment on when the purported improvement in the performance distribution $f_{B^j,n_s^j}(c)$ is crucial to drive this equilibrium result. Note that the wages are paid from employers to workers and are thus effectively a transfer of welfare from the employers to the workers. To make this explicit, consider the case when $\lambda_{B^j} = \lambda_{F^j}$. This case is interesting since it includes namely the utilitarian welfare objective (i.e. when $\lambda_{B^j} = \lambda_{F^j} = 1$). In this case, the sum of the welfare of a beneficiary of affirmative action (i.e. $\omega_s^{j*}(g(c))$) and the employer's profit coming from the work of this particular worker (i.e. $c - \omega_s^{j*}(g(c))$) is simply $\omega_s^{j*}(g(c)) + c - \omega_s^{j*}(g(c)) = c$. We then see that the beneficial effect of the inflated curricula vitae on group B^j workers benefiting from affirmative action is no longer the factor driving the district-j policy maker's decision, since in this case it no longer appears in the welfare function. In this case, the anticipated improvement in the performance distribution $f_{B^j,n_s^j}(c)$ (and thus the productivity c) of future cohorts of B^j workers is the only factor

[13] It is interesting to note that it is enough that such a parameter γ_{B^j}, capturing the feeling of injustice felt by members of group B^j not benefiting from affirmative action, corresponds to one chosen by the policy maker and it need not be the actual one felt in population B^j. Indeed, recall that the equilibrium wage ω_s^{j*} actually does not depend on γ_{B^j}. Only the welfare $W_{B^j,s}$ of group B^j does.

driving the policy maker's incentive to implement the affirmative action policy.

4.1.2 Information available to policy makers

Finally, note that we did not explicitly assume anything about what district policy makers actually observe (the actual decision σ_s^j's of other policy makers or the aggregate statistics $\overline{\sigma}_s$). We wish to stress that our analysis remains identical whether or not we assume that the actual σ_s^j are mutually observed by the various district policy makers, as long as we retain our crucial assumption that *employers* do not observe them (i.e. they only have access to aggregate statistics about these policy decisions, as already discussed). This is so because a district policy maker cares about what other policy makers might do only indirectly, through the effect on the wages. Given that employers do not observe the policy makers' decisions, this implies that in equilibrium the best-response of a district policy maker is unaffected by what other policy makers may decide (now or in the future), irrespectively of the history of play.

4.2 First-best policy

4.2.1 Main result

Our second main result, Proposition 2, states that in the first-best scenario, affirmative action policies *always* end after a finite number of periods.[14]

Proposition 2 (Temporary affirmative action as first-best policy) *Suppose that at time* $t = 0$*, a single centralized policy maker announces (and commits to) the policy plan* $\hat{\sigma} = \{\{\hat{\sigma}_t^j\}_{j \in J}\}_{t=1}^{\infty}$ *that maximizes the welfare function* $\sum_{t=1}^{\infty} \int_{j \in J} \delta^t (W_{A^j,t} + \lambda_{B^j} W_{B^j,t} + \lambda_{F^j} \Pi_{F^j,t}) dj$*, and assume* $\gamma_{A^j} \neq 0$ *(or likewise* $\gamma_{B^j} \neq 0$ *when* $\xi < 1$*). Then for any* $\lambda_{B^j} \in [0,1]$*, there exists* $\overline{\delta} \in (0,1)$ *such that for all* $\delta \in (\overline{\delta}, 1)$*,* $\hat{\sigma}$ *has a threshold form:* $\hat{\sigma}_t^j = 1$ *for* $t < \overline{T}^j$ *and* $\hat{\sigma}_t^j = 0$ *for* $t \geq \overline{T}^j$*, for some (finite)* $\overline{T}^j \in \mathbb{N}$.

Proposition 2 essentially means that if different policy makers were able to coordinate their actions over time periods so as to maximize global welfare, they would never choose to make affirmative action permanent. The intuition is quite simple: After a certain number of periods the improvement in the performance distribution becomes marginal, while the depressing effect on wages (corresponding to curricula vitae being devalued) is not. As a matter of fact, $f_{B^j,n_t^j}(c)$ converges from below to a limiting distribution $\overline{f}_{B^j}(c)$, implying that the distributional improvements become smaller and smaller as affirmative action policies are implemented over time. By contrast, the welfare costs induced by the feeling of injustice remain bounded away from 0, no matter what the history of affirmative action policies is.

The optimal threshold $\overline{T}^j$, while always finite, depends on the relative weight placed by policy makers on the welfare of the targeted group B^j relative to the main group A^j, i.e. on λ_{B^j}, as well as on the intensity of the feeling of injustice felt by non-beneficiaries of both groups, i.e. on γ_{A^j} and γ_{B^j}, and on the fraction ξ of group B^j reached by the affirmative action policy.

[14] In some sense, the difference between the commitment case, leading to the first-best policy, and the equilibrium case, as described in Proposition 1, can be related to the inefficiencies caused by governmental time inconsistencies, as in the pioneering work of Kydland and Prescott (1977). The informational channel through which this occurs in the present article is however different from the one considered in the literature following Kydland and Prescott (1977). See also Bisin, Lizzeri, and Yariv (2015) for more recent work relating the inefficiencies of the governmental actions to the time inconsistency of economic agents (rather than that of governments).

Note that when $\lambda_{B^j} < 1$ (i.e. when the policy maker cares relatively more about group A^j than group B^j), the parameter γ_{A^j} governing the feeling of injustice of group A^j can be 0 and the first-best policy will still prescribe stopping affirmative action after a finite number of periods, because the depressed wage penalizes group A^j sufficiently while the performance distribution of group B^j is only marginally improved.

When $\lambda_{B^j} = 1$ (i.e. when the policy maker cares equally about group A^j and group B^j), then since the average wage is equal to the average performance level across the district (i.e. $\mathbb{E}_t[\omega_t^j(\bar{c})] = \mathbb{E}_t[c|j]$), an affirmative action policy effectively represents just a transfer of welfare from the non-beneficiaries to the beneficiaries. Indeed, this transfer of welfare takes place through non-beneficiaries of both groups A^j and B^j receiving wages lower than their performance levels while beneficiaries receive wages higher than their performance levels. In this case, as long as the parameter γ_{A^j} (or likewise $\gamma_{B^j} \neq 0$ when $\xi < 1$) is *strictly greater* than 0 (no matter how small it is), a first-best policy will prescribe stopping affirmative action after a finite number of periods because otherwise the feeling of injustice felt by non-beneficiaries would become worse than the improvement in the performance distribution of group B^j after sufficiently many implementations of the affirmative action policy.

Finally, if λ_{B^j} were to be strictly greater than 1 (i.e. when the policy maker cares relatively more about group B^j than group A^j), then we might need γ_{A^j} to be sufficiently positive in order to justify stopping affirmative action in the case when $\xi = 1$ (i.e. when all group B^j members benefit from affirmative action). Otherwise, when $\xi < 1$, any feeling of injustice felt by non-beneficiaries of group B^j (i.e. $\gamma_{B^j} \neq 0$) will justify stopping affirmative action at some point.

4.2.2 Potential ways to remedy the informational inefficiency

Now that we have seen that the equilibrium behavior described in Proposition 1 is inefficient, we may ask how this inefficiency could be alleviated. As our model emphasizes an inefficiency that is purely informational, a potential solution may be to design an informational structure that is more transparent. Namely, making the affirmative action policies more transparent in terms of which worker benefited from them—and allowing employers to condition wages on that information—could help remedy the inefficiency.

To illustrate this, call $\{aa\}_i$ the event that worker i benefited from the policy. If $\{aa\}_i$ is observed in the labor market, then the wage given to a worker i at time t is $\omega_t(\bar{c}, \{aa\}_i) = \mathbb{E}_t[c|\bar{c}, \{aa\}_i]$. Thus, $\omega_t(\bar{c}, 1) = g^{-1}(\bar{c}) = c$ is the salary paid to a beneficiary of affirmative action with performance level c and $\omega_t(\bar{c}, 0) = \bar{c} = c$ is the salary paid to a non-beneficiary of affirmative action with performance level c. Hence, this allows each worker to be paid a wage equal to his performance level, thereby eliminating the informational inefficiencies and any feeling of injustice felt by non-beneficiaries (of both groups A^j and B^j). The beneficiaries are then naturally paid less than before, as $\omega_t(\bar{c}, 1) = c < \omega_t^j(g(c))$. Namely, they no longer get the artificial enhancement $g(c)$ to their curriculum vitae, but they still get the long-term distributional improvements to $f_{B^j, n_t^j}(c)$ (and the improved future salaries associated with higher performance levels). As for the non-beneficiaries (of either group A^j or B^j), they are paid more than before since $\omega_t(\bar{c}, 0) = c > \omega_t^j(c)$.

While it may be impractical to fully observe who benefited from an affirmative action policy, improving the information available to employers nevertheless helps. Indeed, in Section 5.1.1, we examine the effect of allowing the employers to condition wages on group identity, i.e. giving a wage

$\omega_t^{G^j}(\overline{c}) = \mathbb{E}_t[c|\overline{c}, G^j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$ to a worker from group G^j. As we further explain in that section, while this eliminates the feeling of injustice felt by group A^j, it does not eliminate the feeling of injustice felt by non-beneficiaries of group B^j. This can only happen if the affirmative action policy reaches *all* members of group B^j (i.e. $\xi = 1$) and hence no member of this group suffers a feeling of injustice. Indeed, then all workers from group B^j receive wages equal to their performance levels, as there is no uncertainty as to whether they benefited from the policy or not. As a result, the equilibrium coincides with the first-best.

4.2.3 Other interpretations of the parameter γ_G and alternative formulations

As discussed in Section 4.2.1, the feeling of injustice γ_G helps establish the inefficiency of permanent affirmative action policies. Indeed, the friction it introduces (the asymmetry in how the policy affects beneficiaries and non-beneficiaries) can be necessary to avoid the possibility that such policies are mere transfers across groups.

Alternative formulations are however possible. Namely, it should be highlighted that any depreciation of the workers' utility, when receiving a payoff lower than their productivity, would lead to the same conclusions. For example, one could think of workers who would be less motivated in such a case, thereby enjoying less utility from their time at work. Likewise, we could think of workers who would not optimally exploit a previous investment made in skill acquisition, given that this investment was made on the premise that they would earn a wage at least equal to their productivity. Another possible formulation could be to have a cost resulting from a labour market congestion effect, as discussed in detail in Section 5.1.8.

5 Some extensions and further discussion

5.1 Discussion of assumptions

In the next subsections, we show that our main results are quite robust and most often hold, even if we relax the assumptions made in the main part of the paper. We explain that all we really need for our main results to hold is that an employer cannot be certain that a worker from group B^j has not benefited from affirmative action.

5.1.1 Allowing employers to condition wages on group identity

In our main model, we have not allowed employers to condition wages on the group A or B to which a worker belongs. This was motivated on grounds that such discrimination is in general forbidden. If conditioning wages on group identity were allowed, our results would actually hold as long as some members of the targeted group B do not benefit from affirmative action (a fairly weak assumption). In such a case, the feeling of injustice is suffered entirely by them (and not also by members of group A) and this is enough for our results to hold.

More specifically, a group G^j worker would receive a wage $\omega_t^{G^j}(\overline{c}) = \mathbb{E}_t[c|\overline{c}, G^j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$. In this case, a group A^j worker would receive a wage $\omega_t^{A^j}(\overline{c}) = \overline{c} = c$ equal to his performance level, since group A workers do not benefit from affirmative action. Group A workers would then suffer no feeling of injustice and be unaffected by the affirmative action policy. When affirmative action only reaches a fraction $\xi \in (0, 1)$ of group B^j, then the share $1 - \xi$ of group B^j workers not benefiting from affirmative action would still get a wage $\omega_t^{B^j}(\bar{c}) < \bar{c} = c$ lower than their performance level and thus suffer a feeling of injustice. The share ξ of group B^j workers benefiting from the policy would still get a wage $\omega_t^{B^j}(\bar{c}) > g^{-1}(\bar{c}) = c$ higher than their performance level and thus would not suffer a feeling of injustice.

We see that allowing employers to condition wages on the group A or B to which a worker belongs leads to similar insights as in the main model, although the key tensions now take place entirely within group B.

The particular case of $\xi = 1$, that is when affirmative action reaches *all* members of group B^j, is worth commenting on. In this particular case, $\omega_t^{B^j}(\bar{c}) = g^{-1}(\bar{c}) = c$, and workers from both groups A and B would receive wages equal to their performance levels, as there would be no uncertainty as to whether they benefited from the policy or not. As a result, the equilibrium would coincide with first-best.

5.1.2 Variations on the informational assumptions

We could consider several variations in what we assume employers can condition their wages on.

As already said, in our environment with a continuum of districts, assuming that employers observe the average policy over various ranges of districts would not affect the analysis, as long as the averages are taken over positive Lebesgue measures of districts. This is so because a deviation by a single district policy maker would not affect such statistics, even if finer than the one considered in the main model.

From another perspective, one might request that employers do not condition their wage on the district j the worker comes from. Such constraints may be the result of anti-discrimination considerations, this time based on location rather than group membership, as it may reasonably be argued that employers would naturally have access to where the worker completed his studies. Our equilibrium analysis would be unchanged in this setting. One might have thought that having the same wage across districts would create an additional externality between the policy makers of the various districts, to the extent that a choice of policy in some districts could now adversely affect the wages received by workers in other districts. However, in our setting where only aggregates are observed, the equilibrium already involves perpeptual affirmative action even without this externality. It then follows that the same property of perpetual affirmative actions would a fortiori hold when such an externality is present[15].

The nature of the first-best policy (a threshold form) would be unaffected by such modifications. This establishes the robustness of our main insights with respect to a broad class of observational environments.

5.1.3 Making the number of districts finite

We now consider an analogous variant of our model where the number of districts is finite, i.e. $|J| \in \mathbb{N}_+$. Then, Proposition 1 still holds as long as there are sufficiently many districts and $\lambda_{B^j} > \lambda_{F^j}$.

[15] By contrast, such an externality would have an effect (resulting in shifting towards more affirmative action policies being implemented) in contexts where the affirmative action policies would be observed, district by district, by employers.

To see this, recall that the only reason the policy maker would choose not to implement an affirmative action policy is to have an uplifting effect on the actual equilibrium wage function ω_t^{j*} chosen by employers, which is depressed by the possibility that a worker has benefited from affirmative action. However, as only an anonymous aggregate statistic $\overline{\sigma}_t = \frac{1}{|J|}\sum_{j=1}^{|J|}\sigma_t^j$ is observed by employers when setting wages, when $|J|$ is high enough the effect of some district j's policy σ_t^j on the $\overline{\sigma}_t$ (and thus on the wage) can be made arbitrarily small.

Therefore, there is no reason why a particular policy maker would deviate, by choosing $\sigma_t^j = 0$, from an equilibrium policy $\sigma_t^{j*} = 1$ in which it implements an affirmative action policy.

Conversely, a deviation[16] from a putative equilibrium in which $\sigma_t^{j*} = 0$ to $\sigma_t^j = 1$ would inflate the curricula vitae of B^j workers at time t, while having only an arbitrarily small impact on the statistic $\overline{\sigma}_t$ observed by employers and thus an arbitrarily small worsening effect on the wage function ω_t^{j*}. This again establishes $\sigma_t^{j*} = 1$ as the unique equilibrium when the number of districts is finite, but large enough.

In the special case when $\lambda_{B^j} = \lambda_{F^j}$, then the policy maker in district j will ultimately stop implementing the policy, but later than in the first-best scenario. Indeed, in this particular case, we recall that the wages paid from employers to workers are effectively a transfer of welfare from the employers to the workers: The sum of the welfare of a beneficiary of affirmative action (i.e. $\omega_s^{j*}(g(c))$) and the employer's profit coming from the work of this particular worker (i.e. $c-\omega_s^{j*}(g(c))$) is simply $\omega_s^{j*}(g(c)) + c - \omega_s^{j*}(g(c)) = c$. As previously discussed in Section 4.1, the beneficial effect of the inflated curricula vitae on group B^j workers benefiting from affirmative action is no longer the factor driving the district-j policy maker's decision, since in this case it no longer appears in the welfare function. The anticipated improvement in the performance distribution $f_{B^j,n_s^j}(c)$ (and thus the productivity c) of future cohorts of B^j workers was the only factor driving the policy maker's incentive to implement the affirmative action policy. Since this distributional improvement becomes marginal over time as $f_{B^j,n_s^j}(c)$ converges to its limit, then the worsening effect on the wage, no matter how small, will ultimately be the dominating factor and thus the policy maker will choose to stop implementing the policy.

Consider now a different variant of our model where instead of (accurately) observing an aggregate statistic $\overline{\sigma}_t$, as in Eq. (2), employers were to observe a noisy signal $\tilde{\sigma}_t^j$ about σ_t^j (with a support of signal realizations that would be the same whether $\sigma_t^j = 0$ or 1, for instance $\tilde{\sigma}_t^j = (1-\chi_t^j)\sigma_t^j + \chi_t^j(1-\sigma_t^j)$ with χ_t^j being any Bernouilli random variable). Then $\sigma_t^{j*} = 1$ for all t would remain an equilibrium. This holds even if the signal is only slightly noisy. Indeed, in such an equilibrium, the signal $\tilde{\sigma}_t^j$ observed by employers would be perceived to be uninformative and thus would have no effect on the chosen wage, thereby providing the incentive to policy makers to choose $\sigma_t^j = 1$ in all periods[17]. Moreover, note that this works for *any* finite number of districts $|J| > 0$ (including a single district, when $|J| = 1$). All that is needed is that the implemented policy σ_t^j is not perfectly

[16]Note that there is no way that an employer, upon observing $\overline{\sigma}_t$, can guess which district policy maker is responsible for that deviation since the statistic is anonymous. While, in the eyes of an employer, there are multiple strategy profiles $\sigma_t = \{\sigma_t^j\}_{j=1}^{|J|}$ consistent with $\overline{\sigma}_t$ after such a deviation, we can assume that he forms a uniform belief about them. That is, letting $\Gamma = \{\sigma_t \mid \frac{1}{|J|}\sum_{j=1}^{|J|}\sigma_t^j = \overline{\sigma}_t\}$ be the set of policy decisions profiles compatible with the observed statistic $\overline{\sigma}_t$, then $\mathbb{P}\{\sigma_t\} = \frac{1}{|\Gamma|}$ for any $\sigma_t \in \Gamma$.

[17]This observation is related to one made by Bagwell (1995) in the context of Stackelberg interactions in which the action of the first-mover would be observed with noise. He notes that the Nash equilibrium of the normal form game is then a Perfect Bayesian Equilibrium of the two-stage interaction. Van Damme and Hurkens (1997) later note that there are other equilibria involving mixed strategies, but such equilibria can be regarded as being less robust to the extent that they rely on indifferences of the players.

observable.

The first-best policy would be unaffected by such modifications and this again establishes the robustness of our main insights with respect to a broad class of observational environments.

5.1.4 Accommodating other forms of affirmative action

The leading interpretation of our model so far is that affirmative action takes the form of favoring, in their school/university studies, (some share ξ of the) members of group B^j as opposed to members of group A^j. Moreover, our model has emphasized the decentralized nature of affirmative action decisions so as to motivate our key informational assumption that affirmative action policies are not observable, district by district, by employers.

We however made clear in Section 2.1 that an affirmative action policy can be interpreted as anything that increases the quality of a curriculum vitae above the actual performance level (direct effect) and that is believed to improve the performance distribution of future generations (anticipated role model effect). An alternative form in which affirmative action can take place is instead through biased promotions, rather than through biased school admissions. Thus, think now of the life of a worker as having two phases: The *early* phase and the *mature* phase. In the early phase, we assume employment takes place in the worker's own district j, while in the mature phase employment takes place in a non-localized labor market. That is, we have in mind that workers in their early phase go to the local labor market and then get rematched to new firms in the mature phase, and that this rematching is not localized. It is not difficult to see how the model developed so far transposes to such a variant. A district-j firm, when considering workers in the early phase, may decide to promote more easily workers from group B^j in an attempt to increase the ability distribution of group B^j workers (through increased motivation, say) in the firm. How a given company favors the promotion of group B^j workers would hardly be known to outsiders, which is in line with our view that, at the rematching stage, it would be difficult to determine whether a group B^j worker benefited from an artificial boost in his early career. On the other hand, such biased promotions in the early careers of (some share ξ of) group B^j workers will depreciate the assessment of early-career promotions in the mature phase, leading to a feeling of injustice among the non-beneficiaries of such biased early-career promotions.

Other forms of affirmative action decisions could also be studied. For instance, since an affirmative action policy only reaches a fraction ξ of group B^j, a policy maker could be tempted to select *which* workers are to be enrolled in an affirmative action program. Indeed, workers of middle performance get a superior curriculum vitae improvement compared to low and high-performance ones (cf. Fig. 1). On the other hand, the anticipated role model effect—affecting future generations—may be more effective when higher-performance workers get enrolled in the program. Thus the optimal decision as to which workers to enroll may reflect a tradeoff. We leave such an extension for future research.

5.1.5 A model of grade inflation

While we developed our model in the context of affirmative action policies, it readily extends to an application to grade inflation. Grade inflation has generated considerable interest in the last decades and increasing trends in grades have often been observed at the school and university levels. See for instance, Sabot and Wakeman-Linn (1991), Bar et al. (2009) or Tyner and Gershenson (2020).

Grade inflation devalues the information content of grades in general and can cause a feeling of injustice among those who do not benefit from it (or even those whose grade advantage is reduced by this practice).

In our model, we could easily reinterpret a policy decision as whether or not to inflate grades in a particular class or school. A teacher or school administrator (i.e. the equivalent of a policy maker) may want to inflate the grades of his students as it improves their transcripts (the analog of their curriculum vitae), which translates into expanded future opportunities. He may also perceive that such increased opportunities may lead to long-term improvements in the performance distribution through a role model effect. Naturally, grades may not be inflated uniformly and a certain fraction of students may benefit from it, while the remaining fraction may not. A version of the model that directly applies to this situation is the one with a single group (group B) where a random fraction ξ may benefit from the policy, while the remaining fraction $1-\xi$ may not. In this context, our model predicts that a teacher would always choose to inflate grades for the same reasons as described before (see the discussion in Section 5.1.1, which applies directly here), in spite of the feeling of injustice felt by the non-beneficiaries, which makes this equilibrium inefficient.

5.1.6 Generalization to more than two groups

Suppose that instead of a group A^j and a group B^j, there is a set $\mathcal{G}^j = \{G_1^j, G_2^j, ..., G_N^j\}$ of up to N groups in each district, with the non-targeted group being group $G_1^j = A^j$ and the other groups (each with a potentially different $\xi_{G_n^j} \in (0,1)$) being targeted by the affirmative action policy $\sigma_t^j \in \{0,1\}$. Then, all the main results of the paper still hold.

Indeed, implementing the policy benefits the beneficiaries of each target group and, the actual policy decision being unobservable, a policy maker has no incentive to refrain from implementing it since he cannot have any effect on the wage function. Thus, Proposition 1 (permanent affirmative action) holds without changes, with the corresponding conditions on all the $\lambda_{G_n^j}$ and $\gamma_{G_n^j}$ (or $\xi_{G_n^j}$).

Likewise, Proposition 2 (temporary affirmative action as first-best policy) holds without changes—with the corresponding conditions on all the $\lambda_{G_n^j}$, $\gamma_{G_n^j}$ and $\xi_{G_n^j}$—for the same reasons as the ones discussed in Section 4.2.1. Namely, the feeling of injustice felt by non-beneficiaries ultimately dominates the improvements in the performance distributions.

In another variant of this multi-group model, we could also consider a more complex case in which a district-j policy maker chooses a separate action $\sigma_t^{G_n^j} \in \{0,1\}$ for each targeted group G_n^j in his district. Since those decisions $\sigma_t^{G_n^j}$ are separable across groups, then the same results would hold insofar as the actual policy decisions are not directly observable by the labor market. This establishes the robustness of our main results to a broad class of environments.

5.1.7 Allowing affirmative action to improve the performance in the current period

In the model presented so far, agents anticipated that the affirmative action policy had an improving effect on the actual performance distribution of group B^j in *future* periods, through a role model argument.

We could extend the model to allow the policy to also have an improving effect on the actual performance of beneficiaries in the *current* period. For instance, being enrolled in a better university does not only improve one's curriculum vitae, but can also improve actual skills. Then, when $\sigma_t^j = 1$, the performance distribution of the beneficiaries of affirmative action in period t would be $f^{aa}_{B^j,n^j_t}(c) \succ f_{B^j,n^j_t}(c)$. The performance distribution of the non-beneficiaries of group B^j would remain $f_{B^j,n^j_t}(c)$. In such a case, the probability that a worker benefited from affirmative action would simply be written as

$$\mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}^t_{s=1}, \sigma) = \frac{|B^j|\xi\sigma^j_t f^{aa}_{B^j,n^j_t}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c})}{|A^j|f_{A^j}(\overline{c}) + |B^j|(1-\xi\sigma^j_t)f_{B^j,n^j_t}(\overline{c}) + |B^j|\xi\sigma^j_t f^{aa}_{B^j,n^j_t}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c})},$$

which has a similar form as in Lemma 1. The wage is then still formed as

$$\omega^j_t(\overline{c}) = \mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}^t_{s=1}, \sigma) \cdot g^{-1}(\overline{c}) + \big(1 - \mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}^t_{s=1}, \sigma)\big) \cdot \overline{c}.$$

The curriculum vitae boost function g still overvalues the workers benefiting from affirmative action (even if their skill level now tends to be higher than in the main model) and thus the mechanisms generating salary distortions and the feelings of injustice remain.

5.1.8 Accounting for labor market congestion

Here we introduce a labor market congestion externality caused by affirmative action, which allows us to capture at least to some extent the fact that jobs obtained by beneficiaries of affirmative action are no longer available to non-beneficiaries. We show that our main insights go through even in the presence of such elaborations.

In our model, non-beneficiaries of affirmative action suffer from receiving a wage that is lower than their actual performance level, while beneficiaries of affirmative action receive a wage that is higher than their actual performance level. Therefore, a transfer of utility between beneficiaries and non-beneficiaries arises through the wage channel.

From another perspective, affirmative action is often thought of as an allocation problem, e.g. allocating a finite number of jobs between two groups, which would result in extra transfers between beneficiaries and non-beneficiaries of affirmative action in addition to the wage effect considered in our main model. While modeling a full-scale matching process is beyond the scope of this paper, our model can be extended in such a direction by adding a labor market congestion externality. This will be represented by a positive term in the utility function of a beneficiary and a negative term in the utility function of a non-beneficiary.

In this section, for clarity of exposition, we will suppose that for all j, $|A^j| = \overline{A}$ and $|B^j| = \overline{B}$. That is, all districts have the same mass of A workers and the same mass of B workers. We also let $\lambda_{F^j} = 0$ and thus ignore the firms' profits in the welfare objective.

The utility of a beneficiary will take the form

$$\begin{aligned} \tilde{u}_{B^j,t}(g(c), c) &= u_{B^j,t}(g(c), c) + \eta \\ &= \omega^j_t(g(c)) + \eta \end{aligned}$$

where η is a parameter measuring the magnitude of the allocation advantage in the labor market (e.g. the advantage of having a reserved slot in the labor market).

It is easy to see that the aggregate transfer of utility from non-beneficiaries to beneficiaries, due to labor market congestion, is simply

$$\int_{j\in J} \overline{B}\xi\sigma_t^j\eta dj = \overline{J}\ \overline{B}\xi\overline{\sigma}_t\eta,$$

recalling that $|J| = \overline{J}$ and that $\overline{\sigma}_t = \frac{1}{|J|}\int_{j\in J}\sigma_t^j dj$.

The utility of a non-beneficiary will then take the form

$$\begin{aligned}\tilde{u}_{G^j,t}(c,c) &= u_{G^j,t}(c,c) - K(\overline{\sigma}_t)\eta \\ &= \omega_t^j(c) - \gamma_{G^j}(c - \omega_t^j(c)) - K(\overline{\sigma}_t)\eta\end{aligned}$$

where $K(\overline{\sigma}_t) = \frac{\overline{J}\ \overline{B}\xi\overline{\sigma}_t}{\overline{J}\left(\overline{A}+\overline{B}(1-\xi\overline{\sigma}_t)\right)}$ is a term reflecting the congestion externality faced by a non-beneficiary of affirmative action in the labor market, due to certain slots being reserved for beneficiaries (the total mass of non-beneficiaries being given by $\overline{J}\big(\overline{A}+\overline{B}(1-\xi\overline{\sigma}_t)\big)$).

In what follows, call $W_{A^j,t}$ and $W_{B^j,t}$ the welfare of groups A^j and B^j at time t, *absent* the congestion externality. Also note that $W_{B^j,t} = W^{nb}_{B^j,t} + W^{b}_{B^j,t}$, the sum of the welfare of non-beneficiaries and beneficiaries of group B^j.

With congestion, the welfare of non-beneficiaries of group A^j at time t is

$$\begin{aligned}\tilde{W}_{A^j,t} &= \overline{A}\int_0^1 \tilde{u}_{A^j,t}(c,c) f_{A^j} dc \\ &= \overline{A}\int_0^1 \big(u_{A^j,t}(c,c) - K(\overline{\sigma}_t)\eta\big) f_{A^j}(c) dc \\ &= W_{A^j,t} - \overline{A}K(\overline{\sigma}_t)\eta. \qquad (4)\end{aligned}$$

The welfare of non-beneficiaries of group B^j at time t is

$$\begin{aligned}\tilde{W}^{nb}_{B^j,t} &= \overline{B}(1-\xi\sigma_t^j)\int_0^1 \tilde{u}_{B^j,t}(c,c) f_{B^j,n_t^j} dc \\ &= \overline{B}(1-\xi\sigma_t^j)\int_0^1 \big(u_{B^j,t}(c,c) - K(\overline{\sigma}_t)\eta\big) f_{B^j,n_t^j}(c) dc \\ &= W^{nb}_{B^j,t} - \overline{B}(1-\xi\sigma_t^j)K(\overline{\sigma}_t)\eta \qquad (5)\end{aligned}$$

while the welfare of beneficiaries of group B^j at time t is

$$\begin{aligned}\tilde{W}^{b}_{B^j,t} &= \overline{B}\xi\sigma_t^j\int_0^1 \tilde{u}_{B^j,t}(g(c),c) f_{B^j,n_t^j} dc \\ &= \overline{B}\xi\sigma_t^j\int_0^1 (u_{B^j,t}(g(c),c) + \eta) f_{B^j,n_t^j}(c) dc \\ &= W^{b}_{B^j,t} + \overline{B}\xi\sigma_t^j\eta. \qquad (6)\end{aligned}$$

Using Eqs. (4) to (6) and the fact that $W_{B^j,t} = W^{nb}_{B^j,t} + W^{b}_{B^j,t}$, we obtain that the welfare in district j at time t is

$$\tilde{W}_{A^j,t} + \lambda_{B^j}\tilde{W}_{B^j,t} = W_{A^j,t} + \lambda_{B^j}W_{B^j,t} - (\overline{A} + \lambda_{B^j}\overline{B})K(\overline{\sigma}_t)\eta + \lambda_{B^j}\overline{B}\xi\sigma_t^j\eta\big(K(\overline{\sigma}_t)+1\big). \quad (7)$$

It is the welfare at time t, *absent* the congestion externality, plus additional terms representing the welfare associated to the transfer of utility from non-beneficiaries to beneficiaries due to labor market congestion. We see from Eq. (7) that choosing $\sigma_t^j = 1$ results in an additional benefit. Indeed, the additional labor market allocation benefit to the beneficiaries of affirmative action in district j is positive, whereas there is no additional labor market congestion felt by non-beneficiaries since district j has measure zero and the decision $\sigma_t^j = 1$ therefore cannot influence $\overline{\sigma}_t$ (and $K(\overline{\sigma}_t)$).

A time-t policy maker's objective function (evaluated at some putative policy sequence σ) can now be written as

$$\sum_{s=t}^{\infty} \delta^{s-t} \Big(W_{A^j,s} + \lambda_{B^j} W_{B^j,s} - (\overline{A} + \lambda_{B^j} \overline{B}) K(\overline{\sigma}_s)\eta + \lambda_{B^j} \overline{B} \xi \sigma_s^j \eta \big(K(\overline{\sigma}_s) + 1 \big) \Big).$$

We therefore have the following analogue of Proposition 1.

Observation 2 (Equilibrium policy with congestion) *Proposition 1 (permanent affirmative action in equilibrium) holds in the presence of labor market congestion.*

We will now show that, in the presence of labor market congestion, the first-best policy also involves temporary affirmative action. For simplicity of exposition, we will suppose here that $\lambda_{B^j} = \lambda_B$ for all $j \in J$. That is, all district policy makers place the same weight on the welfare of group B relative to that of group A.

A centralized policy maker's objective function is now

$$\sum_{t=1}^{\infty} \delta^t \int_{j \in J} \Big(W_{A^j,t} + \lambda_B W_{B^j,t} - (\overline{A} + \lambda_B \overline{B}) K(\overline{\sigma}_s)\eta + \lambda_B \overline{B} \xi \sigma_t^j \eta \big(K(\overline{\sigma}_t) + 1 \big) \Big) dj,$$

which can be simplified as

$$\sum_{t=1}^{\infty} \delta^t \Big(\int_{j \in J} \big(W_{A^j,t} + \lambda_B W_{B^j,t} \big) dj - \overline{J} \big(\overline{A} + \lambda_B \overline{B} (1 - \xi \overline{\sigma}_t) \big) K(\overline{\sigma}_t)\eta + \lambda_B \overline{J}\ \overline{B} \xi \overline{\sigma}_t \eta \Big). \tag{8}$$

The first term in Eq. (8) is the welfare in the *absence* of congestion. It is easy to show that the second and third terms sum to 0 when $\lambda_B = 1$, since the congestion externality amounts to a transfer of utility between beneficiaries and non-beneficiaries. It is also easy to verify that, when $\lambda_B < 1$, the sum of the second and third terms is less than 0, as the welfare transfer to beneficiaries is weighted relatively less than what is taken from non-beneficiaries.

It then follows that when $\lambda_B \in [0, 1]$, labor market congestion represents an additional cost of implementing affirmative action in each period. The argument of Proposition 2 thus still holds and affirmative action will be stopped after a finite number of periods.

Observation 3 (First-best policy with congestion) *Proposition 2 (temporary affirmative action as first-best policy) holds in the presence of labor market congestion.*

5.1.9 Other elaborations

In an Appendix (Section 7.2), we formulate a generalized model where we allow for strategic behavior by workers, by which they can present a curriculum vitae of any chosen quality. We show that the wage chosen by employers is then a non-decreasing function of the curriculum vitae quality. This generalization formally removes the need for Assumption 1.

5.2 Comparisons with existing literature

We mainly depart from the existing literature on affirmative action by studying the incentives of decentralized policy makers to implement affirmative action policies. Indeed, most of the literature focuses on other incentives: those linked to hiring decisions made by employers or to investments in human capital made by workers (see Fang and Moro (2011) for a survey).

The existing literature on affirmative action is vast and often tries to describe or explain inequalities between groups. Early developments include taste-based theories of discrimination (e.g. Becker (1957)), which suppose that exogenous preferences generate wage differences between groups, although the latter are unlikely to persist in competitive markets. Statistical discrimination theories, on the other hand, mainly attempt to explain outcome differences using imperfect information about the workers' performance levels, which leads to different wages being rationally paid to workers of different groups (e.g. Phelps (1972), Arrow (1973), Lundberg and Startz (1983), Coate and Loury (1993a) or Coate and Loury (1993b)). Such models often also link these different wages to the workers' incentives to invest in human capital, thus sustaining a performance gap between groups. In a more recent development in this line of work, Echenique and Li (2022) borrow elements from both the Arrovian and Phelpsian frameworks—together with insights from the recent literature on rational inattention (Sims (2003))—to endogenize an employer's acquisition of information about the productivity of workers. In equilibrium, the workers' incentives to invest in skills are influenced by how they expect to be rewarded by the employer, which depends on the endogenously chosen information structure. They show that, with costly information acquisition, an employer may prefer to be rationally inattentive and to meticulously screen only some workers, thereby properly incentivizing them to invest in skills development, while rationally ignoring other workers, leading the latter to underinvest.

In other recent work, Kim and Loury (2018) argue that when a group is affected by negative reputational externalities, the group cannot escape a low skill investment trap. Empirical work has also documented the stigmatization effect of affirmative actions policies on the members of the target group (for example, Heilman et al. (1992), Heilman et al. (1997) or Leslie et al. (2014)). We make this a central component of our model, through the fact that workers' curriculum vitae are devalued in the eyes of employers, due to the inference the latter make about the possibility a curriculum vitae may be artificially enhanced by affirmative action.

Another relevant strand of literature focuses on the potential mismatch consequences of affirmative action (see namely Arcidiacono et al. (2011), or the widely debated Sander (2005)). While this mismatch aspect—just like the impact of affirmative action on human capital investment decisions mentioned previously—is relevant to examine, we have chosen not to include such traditional ingredients in our model so as to highlight a novel source of potential inefficiency induced by affirmative action when implemented over many periods at a decentralized level. In summary, our inefficiency is based on a novel moral hazard consideration on the policy makers' part, and it complements other potential inefficiencies such as those emphasized in the above literature.

6 Proofs

Proof of Lemma 1 (Wage function).

Note that $\omega_t^j(\overline{c})$ is the conditional expectation of a district-j worker's actual performance level at time t when declaring a curriculum vitae of quality $\overline{c}$, given a putative policy sequence $\sigma = \{\{\sigma_t^j\}_{j\in J}\}_{t=1}^{\infty}$ assumed by employers and given observed aggregate (average) policy statistics $\{\overline{\sigma}_s\}_{s=1}^t$ consistent with σ (which is the case on the equilibrium path). Thus,

$$\begin{aligned}\omega_t^j(\overline{c}) &= \mathbb{E}_t[c|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] \\ &= \mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma) \cdot g^{-1}(\overline{c}) + \left(1 - \mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma)\right) \cdot \overline{c}\end{aligned}$$

Now to express $\mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma)$, we first express $\mathbb{P}_t(\{aa\}|\tilde{c} \in N(\overline{c}, \epsilon), j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma)$, where $N(\overline{c}, \epsilon)$ is an ϵ-neighborhood of $\overline{c}$:

$$\begin{aligned}\mathbb{P}_t(\{aa\}|\tilde{c} \in N(\overline{c}, \epsilon), j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma) &= \frac{\mathbb{P}_t(\{\tilde{c} \in N(\overline{c}, \epsilon)\} \bigcap \{aa\}|j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma)}{\mathbb{P}_t(\tilde{c} \in N(\overline{c}, \epsilon)|j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma)} \\ &= \frac{\mathbb{P}_t(\tilde{c} \in N(\overline{c}, \epsilon) \bigcap B^j|j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma) \cdot \xi\sigma_t^j}{\mathbb{P}_t(\tilde{c} \in N(\overline{c}, \epsilon)|j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma)} \\ &= \frac{\mathbb{P}_t(\tilde{c} \in N(\overline{c}, \epsilon)|B^j, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma) \cdot \mathbb{P}(B^j) \cdot \xi\sigma_t^j}{\mathbb{P}_t(\tilde{c} \in N(\overline{c}, \epsilon)|j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma)} \\ &= \frac{\mathcal{C}\frac{|B^j|}{|A^j|+|B^j|}\xi\sigma_t^j}{\mathcal{A}\frac{|A^j|}{|A^j|+|B^j|} + \mathcal{B}\frac{|B^j|}{|A^j|+|B^j|}(1-\xi\sigma_t^j) + \mathcal{C}\frac{|B^j|}{|A^j|+|B^j|}\xi\sigma_t^j} \\ &= \frac{\mathcal{C}|B^j|\xi\sigma_t^j}{|A^j|\mathcal{A} + |B^j|(1-\xi\sigma_t^j)\mathcal{B} + |B^j|\xi\sigma_t^j\mathcal{C}}\end{aligned}$$

where $\mathcal{A} = \int_{\tilde{c}\in N(\overline{c},\epsilon)} f_{A^j}(\tilde{c})d\tilde{c}$, while $\mathcal{B} = \int_{\tilde{c}\in N(\overline{c},\epsilon)} f_{B^j,n_t^j}(\tilde{c})d\tilde{c}$ and $\mathcal{C} = \int_{\tilde{c}\in N\left(g^{-1}(\overline{c}), \frac{\epsilon}{(g^{-1})'(\overline{c})}\right)} f_{B^j,n_t^j}(g^{-1}(\tilde{c}))d\tilde{c}$.

Then, we take the limit as $\epsilon \to 0$:

$$\begin{aligned}\mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma)) &= \lim_{\epsilon\to 0} \mathbb{P}_t(\{aa\}|\tilde{c} \in N(\overline{c}, \epsilon), j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma)) \\ &= \lim_{\epsilon\to 0} \frac{|B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))\frac{2\epsilon}{(g^{-1})'(\overline{c})}}{|A^j|f_{A^j}(\overline{c})2\epsilon + |B^j|(1-\xi\sigma_t^j)f_{B^j,n_t^j}(\overline{c})2\epsilon + |B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))\frac{2\epsilon}{(g^{-1})'(\overline{c})}} \\ &= \frac{|B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c})}{|A^j|f_{A^j}(\overline{c}) + |B^j|(1-\xi\sigma_t^j)f_{B^j,n_t^j}(\overline{c}) + |B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c})}\end{aligned}$$

■

Proof of Lemma 2 (Non-decreasing wage function).

First note that when ξ is high enough, then $\mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma)$ is non-increasing in $\overline{c}$.

Indeed, dividing the numerator and the denominator of the expression for $\mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma)$ in Lemma 1 by $|B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c})$, we obtain

$$\begin{aligned}\mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma) &= \frac{|B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c})}{|A^j|f_{A^j}(\overline{c}) + |B^j|(1-\xi\sigma_t^j)f_{B^j,n_t^j}(\overline{c}) + |B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c})} \\ &= \frac{1}{\frac{|A^j|f_{A^j}(\overline{c}) + |B^j|(1-\xi\sigma_t^j)f_{B^j,n_t^j}(\overline{c})}{|B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c})} + 1}.\end{aligned}$$

The first term at the denominator can be expressed as

$$\frac{|A^j|f_{A^j}(\overline{c})+|B^j|(1-\xi\sigma_t^j)f_{B^j,n_t^j}(\overline{c})}{|B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c})}=\frac{|A^j|f_{A^j}(\overline{c})}{|B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))}(g^{-1})'(\overline{c})+\frac{(1-\xi\sigma_t^j)f_{B^j,n_t^j}(\overline{c})}{\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))}(g^{-1})'(\overline{c}), \tag{9}$$

which is non-decreasing when the likelihood ratio $\frac{f_{A^j}(\overline{c})}{f_{B^j,n_t^j}(g^{-1}(\overline{c}))}$ is non-decreasing in $\overline{c}$ (Assumption 1(i)) and when ξ is high enough (as the second term in Eq. (9) becomes negligible), noting also that $(g^{-1})'(\overline{c})$ is increasing in $\overline{c}$ since $g^{-1}(\overline{c})$ is increasing and convex when $g(\overline{c})$ is increasing and concave (Assumption 1(ii)).

It follows that $\mathbb{P}_t(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)$ is non-increasing in $\overline{c}$.

Now note that when $\mathbb{P}_t(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)$ is non-increasing in $\overline{c}$, then $\omega_t^j(\overline{c})$ is non-decreasing in $\overline{c}$.

Indeed,

$$\begin{aligned}\omega_t^j(\overline{c}) &= \mathbb{P}_t(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)\cdot g^{-1}(\overline{c})+\big(1-\mathbb{P}_t(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)\big)\cdot\overline{c}\\ &= \mathbb{P}_t(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)\cdot\big(g^{-1}(\overline{c})-\overline{c}\big)+\overline{c}.\end{aligned}$$

and taking the derivative of $\omega_t^j(\overline{c})$ with respect to $\overline{c}$ yields

$$\begin{aligned}(\omega_t^j)'(\overline{c}) &= \mathbb{P}_t'(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)\cdot\big(g^{-1}(\overline{c})-\overline{c}\big)+\mathbb{P}_t(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)\cdot\big((g^{-1})'(\overline{c})-1\big)+1\\ &\geq \mathbb{P}_t'(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)\cdot\big(g^{-1}(\overline{c})-\overline{c}\big)+\mathbb{P}_t(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)\cdot\big(0-1\big)+1\\ &= \mathbb{P}_t'(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)\cdot\big(g^{-1}(\overline{c})-\overline{c}\big)+1-\mathbb{P}_t(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)\\ &\geq 0.\end{aligned}$$

The first inequality follows from the fact that $(g^{-1})'(\overline{c})>0$ as $g^{-1}(\overline{c})$ is increasing. The second inequality follows from the fact that $\mathbb{P}_t'(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)\cdot\big(g^{-1}(\overline{c})-\overline{c}\big)\geq 0$ since $g^{-1}(\overline{c})\leq\overline{c}$ and $\mathbb{P}_t'(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)\leq 0$ and from the fact that $1-\mathbb{P}_t(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)\geq 0$.

It follows that $\omega_t^j(\overline{c})$ is non-decreasing in $\overline{c}$. ■

Proof of Lemma 3 (Wage versus performance level).

From Lemma 1 we know that

$$\omega_t^j(\overline{c})=\mathbb{P}_t(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)\cdot g^{-1}(\overline{c})+\big(1-\mathbb{P}_t(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)\big)\cdot\overline{c}$$

Part (i): When $\sigma_t^j=1$, then $\mathbb{P}_t(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)>0$. Since $g^{-1}(\overline{c})<\overline{c}$, it follows immediately that $g^{-1}(\overline{c})<\omega_t^j(\overline{c})<\overline{c}$. Thus, if the worker does not benefit from affirmative action (i.e. $c=\overline{c}$), then $\omega_t^j(\overline{c})<c$ and he gets a wage lower than his performance level. On the other hand, if the worker benefits from affirmative action (i.e. $c=g^{-1}(\overline{c})$), then $c<\omega_t^j(\overline{c})$ and he gets a wage higher than his performance level.

Part (ii): When $\sigma_t^j=0$, then $\mathbb{P}_t(\{aa\}|\overline{c},j,\{\overline{\sigma}_s\}_{s=1}^t,\sigma)=0$. Thus, $\omega_t^j(\overline{c})=\overline{c}$ and $\overline{c}=c$ since no one benefits from affirmative action. ■

Proof of Proposition 1 (Equilibrium policy).

We first show that $\{\{\sigma_s^{j*}\}_{j\in J}\}_{s=1}^{\infty}=\{\{1\}_{j\in J}\}_{s=1}^{\infty}$ is an equilibrium.

Given some equilibrium decision profile $\sigma^*=\{\{\sigma_s^{j*}\}_{j\in J}\}_{s=1}^{\infty}=\{\{1\}_{j\in J}\}_{s=1}^{\infty}$, any deviation $\sigma_t^{j\prime}$

at some time t has no impact on the wage function since this deviation is unobserved by employers. Indeed, employers form a wage $\omega_t^{j*}(\bar{c}) = \mathbb{E}_t[c|\bar{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma^*\}]$ using their understanding of the strategy played in equilibrium, σ^*, and after observing the aggregate (average) policy statistics $\{\overline{\sigma}_s\}_{s=1}^t$. A district j's actual policy $\sigma_t^{j\prime}$ cannot be traced back to it from observing this aggregate statistic. Moreover, by virtue of having measure zero, an individual district's policy $\sigma_t^{j\prime}$ does not influence the aggregate statistic $\overline{\sigma}_t$.

Therefore, $\sum_{s=t}^{\infty} \delta^{s-t}(W_{A^j,s} + \lambda_{F^j}\Pi_{A^j,s})$, the discounted future welfare of group A^j and weighted profits of the employers (i.e. firms) hiring A^j workers are completely unaffected by an unobserved deviation to $\sigma_t^{j\prime}$. Indeed,

$$\begin{aligned} W_{A^j,s} + \lambda_{F^j}\Pi_{A^j,s} &= |A^j| \int_0^1 \big(u_{A^j,s}(c,c) + \lambda_{F^j}\pi_s(c,c)\big) f_{A^j}(c)dc \\ &= |A^j| \int_0^1 \Big(\big(\omega_s^{j*}(c) - \gamma_{A^j}(c - \omega_s^{j*}(c))\big) + \lambda_{F^j}(c - \omega_s^{j*}(c))\Big) f_{A^j}(c)dc \end{aligned}$$

where the density function $f_{A^j}(c)$ is constant through time and thus not impacted by $\sigma_t^{j\prime}$, while the wage $\omega_t^{j*}(c)$ is unaffected by an unobserved deviation to $\sigma_t^{j\prime}$.

On the other hand, $\sum_{s=t}^{\infty} \delta^{s-t}(\lambda_{B^j}W_{B^j,s} + \lambda_{F^j}\Pi_{B^j,s})$, the discounted weighted sum of the future welfare of group B^j and profits of the employers (i.e. firms) hiring B^j workers is strictly lower following an unobserved deviation from $\sigma_t^{j*} = 1$ to $\sigma_t^{j\prime} = 0$. Indeed, at any time $s \geq t$,

$$\begin{aligned} \lambda_{B^j}W_{B^j,s} + \lambda_{F^j}\Pi_{B^j,s} &= |B^j| \int_0^1 \Bigg(\lambda_{B^j}\Big(\xi\sigma_s^j u_{B^j,s}(g(c),c) + (1-\xi\sigma_s^j)u_{B^j,s}(c,c)\Big) \\ &\quad + \lambda_{F^j}\Big(\xi\sigma_s^j \pi_s(g(c),c) + (1-\xi\sigma_s^j)\pi_s(c,c)\Big)\Bigg) f_{B^j,n_s^j}(c)dc \\ &= |B^j| \int_0^1 \Bigg(\lambda_{B^j}\Big(\xi\sigma_s^j\omega_s^{j*}(g(c)) + (1-\xi\sigma_s^j)\big(\omega_s^{j*}(c) - \gamma_{B^j}(c - \omega_s^{j*}(c))\big)\Big) \\ &\quad + \lambda_{F^j}\Big(\xi\sigma_s^j(c - \omega_s^{j*}(g(c))) + (1-\xi\sigma_s^j)\big(c - \omega_s^{j*}(c)\big)\Big)\Bigg) f_{B^j,n_s^j}(c)dc \quad (10) \end{aligned}$$

where σ_s^j is the "actual" policy decision (not necessarily equal to the equilibrium one σ_t^{j*}).

Let us first examine what happens at time $s = t$.

From Eq. (10), we can see that at time $s = t$, B^j workers lose out from a deviation to $\sigma_t' = 0$, since the beneficiaries of affirmative action no longer get the boost $g(c)$ to their curriculum vitae. Indeed, $\omega_s^{j*}(g(c)) > \omega_s^{j*}(c) - \gamma_{B^j}(c - \omega_s^{j*}(c))$, recalling that $\omega_s^{j*}(g(c)) > \omega_s^{j*}(c)$, that $c > \omega_s^{j*}(c)$ by Observation 1(i) and thus the first term of Eq. (10) (multiplied by λ_{B^j}) decreases following a deviation to $\sigma_t' = 0$.

On the contrary, at time $s = t$, the employers gain from that deviation to $\sigma_t' = 0$ since they no longer pay a boosted wage $\omega_s^{j*}(g(c))$ to the beneficiaries of the policy in group B^j. Indeed, $c - \omega_s^{j*}(g(c)) < c - \omega_s^{j*}(c))$ in the the second term of Eq. (10) (multiplied by λ_{F^j}).

On the other hand, since $\lambda_{B^j} \geq \lambda_{F^j}$, the loss of group B^j workers outweighs the gains of the employers. To see this, note that Eq. (10) can be rearranged as

$$\lambda_{B^j} W_{B^j,s} + \lambda_{F^j}\Pi_{B^j,s} = |B^j| \int_0^1 \Bigg((\lambda_{B^j} - \lambda_{F^j})\Big(\xi\sigma_s^j\omega_s^{j*}(g(c)) + (1-\xi\sigma_s^j)\omega_s^{j*}(c)\Big) - (1-\xi\sigma_s^j)\lambda_{B^j}\gamma_{B^j}\big(c - \omega_s^{j*}(c)\big) + \lambda_{F^j} c \Bigg) f_{B^j,n_s^j}(c)dc, \quad (11)$$

which is clearly increasing in σ_s^j. We conclude that a deviation from $\sigma_t^{j*} = 1$ to $\sigma_t^{j\prime} = 0$ decreases the welfare at time $s = t$.

Let us now examine what happens at times $s > t$.

At times $s > t$, $f_{B^j,n_s^j|\sigma_t^{j\prime}=0}(c) \prec f_{B^j,n_s^j|\sigma_t^{j*}=1}(c)$ since a deviation to $\sigma_t^{j\prime} = 0$ has the effect of not improving the distribution of performance at time $t+1$ compared to the previous period t. Note that both workers and employers can potentially loose from a worsening of the performance distribution. To analyze this more carefully, let us look at Eq. (11). The first term in the large brackets of the right-hand side of Eq. (11) (the one associated with $(\lambda_{B^j} - \lambda_{F^j})$) is non-decreasing in c, since the equilibrium wage function ω_s^{j*} is non-decreasing in c by Assumption 1 and since $\lambda_{B^j} \geq \lambda_{F^j}$. The third term (associated with λ_{F^j}) is also clearly increasing in c for any $\lambda_{F^j} > 0$. It is however unclear whether the second term (associated with $-(1-\xi\sigma_s^j)$) is increasing in c. Indeed, this term is associated to the feeling of injustice felt by non-beneficiaries of group B^j and the later could increase if the performance c were to rise faster than the wage following an improvement in the performance distribution. A sufficient (but not always necessary) condition for the expression in the large brackets on the right-hand side of Eq. (11) to be increasing in c is when γ_{B^j} is small enough, i.e. $\gamma_{B^j} < \overline{\gamma}_{B^j}$ for some appropriately chosen $\overline{\gamma}_{B^j}$, or also when ξ is large enough, i.e. $\xi > \underline{\xi}$ for some appropriately chosen $\underline{\xi}$.

Since $f_{B^j,n_s^j|\sigma_t^{j\prime}=0}(c) \prec f_{B^j,n_s^j|\sigma_t^{j*}=1}(c)$, it then follows by first-order stochastic dominance that $\lambda_{B^j} W_{B^j,s} + \lambda_{F^j}\Pi_{B^j,s}$ decreases (at least weakly) at times $s > t$ following a deviation from $\sigma_t^{j*} = 1$ to $\sigma_t^{j\prime} = 0$.

It follows that as long as $\lambda_{B^j} > 0$, then $\sigma_t^{j*} = 1$ for all t and j will be an equilibrium.

To show that this is the unique equilibrium, we now have to show that a deviation to $\sigma_t^{j\prime} = 1$, from a putative equilibrium in which $\sigma_t^{j*} = 0$, is always desirable for a district-j policy maker at time t. For that purpose, suppose that $\sigma_t^{j*} = 0$ for some j, t. By the same argument as before, $\sum_{s=t}^{\infty} \delta^{s-t}(W_{A^j,s} + \lambda_{F^j}\Pi_{A^j,s})$ is unaffected by any deviation $\sigma_t^{j\prime}$. Thus, we must show that $\sum_{s=t}^{\infty} \delta^{s-t}(\lambda_{B^j} W_{B^j,s} + \lambda_{F^j}\Pi_{B^j,s})$ is strictly higher following a deviation from $\sigma_t^{j*} = 0$ to $\sigma_t^{j\prime} = 1$.

Consider first the effect of this deviation on the welfare at time $s = t$. Since the right-hand side of Eq. (11) is increasing in the actual policy decision σ_t^j, the same argument as before allows us to conclude that $\lambda_{B^j} W_{B^j,t} + \lambda_{F^j}\Pi_{B^j,t}$ increases following a deviation to $\sigma_t^{j\prime} = 1$ from a putative equilibrium in which $\sigma_t^{j*} = 0$.

Consider now the effect of this deviation on the welfare, at any future time $s > t$. We know that $f_{B^j,n_s^j|\sigma_t^{j*}=0}(c) \prec f_{B^j,n_s^j|\sigma_t^{j\prime}=1}(c)$ for all $s > t$ since a deviation to $\sigma_t^{j\prime} = 1$ has the effect of shifting (in a strict first-order stochastic dominance sense) the future performance distributions of group B^j.

Since, as shown before, the expression in large brackets on the right-hand side of Eq. (11) is increasing in c (at least under the above-stated sufficient conditions), it then follows by first-order stochastic dominance that $\lambda_{B^j} W_{B^j,s} + \lambda_{F^j}\Pi_{B^j,s}$ increases (at least weakly) at times $s > t$ following a deviation from $\sigma_t^{j*} = 0$ to $\sigma_t^{j\prime} = 1$.

This concludes the proof.

We will now comment on when the improvement in the performance distribution is crucial to drive this equilibrium result. The wages are effectively a transfer of welfare from the employers (i.e. firms) to the workers. To make this explicit, consider the case when $\lambda_{B^j} = \lambda_{F^j}$. This case is interesting since it includes namely the utilitarian welfare objective (i.e. when $\lambda_{B^j} = \lambda_{F^j} = 1$). In such a case, Eq. (11) becomes

$$\lambda_{B^j} W_{B^j,s} + \lambda_{F^j} \Pi_{B^j,s} \quad = \quad |B^j| \int_0^1 \Bigg(-(1-\xi\sigma_s^j)\lambda_{B^j}\gamma_{B^j}\big(c - \omega_s^{j*}(c)\big) + \lambda_{F^j} c \Bigg) f_{B^j,n_s^j}(c)dc.$$

We see that the effect of the boosted wage $\omega_s^{j*}(g(c))$ has disappeared from the equation, since wages are mere transfers from employers to workers. What is left is the productivity component c of the employers' profits and the feeling of injustice felt by non-beneficiaries of the policy: $\gamma_{B^j}(c - \omega_s^{j*}(c))$.

Note that $\lambda_{F^j} > 0$ (since $\lambda_{B^j} > 0$) and thus, under the appropriate sufficient conditions on γ_{B^j} or ξ, the bracketed term $-(1-\xi\sigma_s^j)\lambda_{B^j}\gamma_{B^j}\big(c - \omega_s^{j*}(c)\big) + \lambda_{F^j} c$ on the right-hand side of the above equation is increasing in c. Choosing $\sigma_t^j = 1$ instead of 0 thus improves the welfare *entirely* through the improvement in the performance distribution $f_{B^j,n_s^j}(c)$ by the first-order stochastic dominance argument. ■

Proof of Proposition 2 (First-best policy).

Note that $\Pi_{F^j,t} = 0$ for all j, t, since we have perfectly competitive markets (see Section 7.1). Thus $\Pi_{F^j,t}$ does not affect the welfare function and can be neglected from the analysis in the first-best case[18].

We start with the following lemma.

Lemma 4 *Consider some particular district j. Let $\sigma^{j\prime} = \{\sigma_t^{j\prime}\}_{t=1}^{\infty}$ be a policy plan with $\sigma_\tau^{j\prime} = 0$ and $\sigma_{\tau+1}^{j\prime} = 1$ for some τ. Let $\sigma^j = \{\sigma_t^j\}_{t=1}^{\infty}$ be another policy plan with $\sigma_\tau^j = 1$, $\sigma_{\tau+1}^j = 0$ and $\sigma_t^{j\prime} = \sigma_t^j$ for all other t. Then there exists $\overline{\delta} \geq 0$ such that for all $\delta \in (\overline{\delta}, 1)$, σ^j yields a strictly higher welfare for district j than $\sigma^{j\prime}$.*

Proof of Lemma 4.

First note that for any group $G^j \in \{A^j, B^j\}$,

$$\sum_{t=1}^{\infty} \delta^t W_{G^j,t} = \sum_{t=1}^{\tau-1} \delta^t W_{G^j,t} + \delta^\tau W_{G^j,\tau} + \delta^{\tau+1} W_{G^j,\tau+1} + \sum_{t=\tau+2}^{\infty} \delta^t W_{G^j,t},$$

where only the terms $\delta^\tau W_{G^j,\tau}$ and $\delta^{\tau+1} W_{G^j,\tau+1}$ are different under policies σ^j versus $\sigma^{j\prime}$. We thus only need to compare these two terms under the two policies.

Suppose for now that $\delta = 1$.

For group A^j, the sum $\delta^\tau W_{A^j,\tau} + \delta^{\tau+1} W_{A^j,\tau+1}$ is the same under policies σ^j and $\sigma^{j\prime}$.

For group B^j, on the other hand, $\delta^\tau W_{B^j,\tau} + \delta^{\tau+1} W_{B^j,\tau+1}$ is strictly greater under plan σ^j than

[18] We had to consider it in the equilibrium case (proof of Proposition 1) since we were considering potential deviations from the equilibrium and such deviations could affect the profits considered in a district's welfare objective.

under $\sigma^{j\prime}$. To see this, note that under σ^j

$$\delta^\tau W_{B^j,\tau}+\delta^{\tau+1}W_{B^j,\tau+1}=\delta^\tau|B^j|\int_0^1[\xi\omega_\tau^j(g(c))+(1-\xi)\big(\omega_\tau^j(c)-\gamma_{B^j}(c-\omega_\tau^j(c))\big)]f_{B^j,n_\tau^j}(c)dc+\delta^{\tau+1}|B^j|\int_0^1\omega_{\tau+1}^j(c)f_{B^j,n_{\tau+1}^j}(c)dc$$

while under $\sigma^{j\prime}$

$$\delta^\tau W'_{B^j,\tau}+\delta^{\tau+1}W'_{B^j,\tau+1}=\delta^\tau|B^j|\int_0^1\omega_\tau^{j\prime}(c)f_{B^j,n_\tau^{j\prime}}(c)dc+\delta^{\tau+1}|B^j|\int_0^1[\xi\omega_{\tau+1}^{j\prime}(g(c))+(1-\xi)\big(\omega_{\tau+1}^{j\prime}(c)-\gamma_{B^j}(c-\omega_{\tau+1}^{j\prime}(c))\big)]f_{B^j,n_{\tau+1}^{j\prime}}(c)dc.$$

The fact that $\delta^\tau W_{B^j,\tau}+\delta^{\tau+1}W_{B^j,\tau+1}>\delta^\tau W'_{B^j,\tau}+\delta^{\tau+1}W'_{B^j,\tau+1}$, when $\delta=1$, follows from the facts that $\omega_{\tau+1}^j(c)=\omega_\tau^{j\prime}(c)=c$, that $\xi\omega_\tau^j(g(c))+(1-\xi)\big(\omega_\tau^j(c)-\gamma_{B^j}(c-\omega_\tau^j(c))\big)=\xi\omega_{\tau+1}^{j\prime}(g(c))+(1-\xi)\big(\omega_{\tau+1}^{j\prime}(c)-\gamma_{B^j}(c-\omega_{\tau+1}^{j\prime}(c))\big)$, that $f_{B^j,n_\tau^j}(c)=f_{B^j,n_{\tau+1}^{j\prime}}(c)$ and that $f_{B^j,n_{\tau+1}^j}(c)\succ f_{B^j,n_\tau^{j\prime}}(c)$.

By continuity, it then follows that there exists $\overline{\delta}\in(0,1)$ such that for all $\delta\in(\overline{\delta},1)$, the total welfare is also higher under plan σ^j than under $\sigma^{j\prime}$. ∎

Therefore, when δ is high enough, it follows by iterative application of Lemma 4 that the optimal policy in district j has a threshold form $\hat{\sigma}_t^j=1$ for $t<\overline{T}^j$ and $\hat{\sigma}_t^j=0$ for $t\geq\overline{T}^j$ for some $\overline{T}^j\in\mathbb{N}\bigcup\infty$.

We will now rule out the case where $\overline{T}^j$ could be infinite and thus show that $\overline{T}^j\in\mathbb{N}$.

Let us thus compare the welfare of some (large) $\overline{T}^j<\infty$ to that of the case $\overline{T}^{j\prime}=\infty$. In what follows, the quantities with a prime ($'$) will be the ones associated to $\overline{T}^{j\prime}=\infty$.

We need to show that

$$\sum_{t=1}^\infty\delta^t(W_{A^j,t}+\lambda_{B^j}W_{B^j,t})>\sum_{t=1}^\infty\delta^t(W'_{A^j,t}+\lambda_{B^j}W'_{B^j,t}). \tag{12}$$

Equivalently, it will be convenient to multiply the welfare by the constant $\frac{1}{|A^j|+|B^j|}$ and verify that

$$\frac{1}{|A^j|+|B^j|}\Big(\sum_{t=1}^\infty\delta^t(W_{A^j,t}+\lambda_{B^j}W_{B^j,t})-\sum_{t=1}^\infty\delta^t(W'_{A^j,t}+\lambda_{B^j}W'_{B^j,t})\Big)>0$$

$$\begin{aligned}\sum_{t=1}^\infty\frac{\delta^t}{|A^j|+|B^j|}\big((W_{A^j,t}+\lambda_{B^j}W_{B^j,t})-(W'_{A^j,t}+\lambda_{B^j}W'_{B^j,t})\big) \quad = \quad & \sum_{t=1}^\infty\delta^t\Big(\frac{|A^j|}{|A^j|+|B^j|}\int\omega_t^j(c)f_{A^j}(c)dc\\ &+\frac{\lambda_{B^j}|B^j|}{|A^j|+|B^j|}\int[\xi\sigma_t^j\omega_t^j(g(c))+(1-\xi\sigma_t^j)\omega_t^j(c)]f_{B^j,n_t^j}(c)dc\\ &-\frac{|A^j|}{|A^j|+|B^j|}\int\omega_t^{j\prime}(c)f_{A^j}(c)dc\\ &-\frac{\lambda_{B^j}|B^j|}{|A^j|+|B^j|}\int[\xi\omega_t^{j\prime}(g(c))+(1-\xi)\omega_t^{j\prime}(c)]f_{B^j,n_t^{j\prime}}(c)dc\Big)\\ &+\sum_{t=1}^\infty\delta^t\Big(\frac{|A^j|}{|A^j|+|B^j|}\gamma_{A^j}\int(\omega_t^j(c)-\omega_t^{j\prime}(c))f_{A^j}(c)dc\\ &+\frac{\lambda_{B^j}|B^j|}{|A^j|+|B^j|}\int(1-\xi\sigma_t^j)\gamma_{B^j}[\omega_t^j(c)-c]f_{B^j,n_t^j}(c)dc\\ &-\frac{\lambda_{B^j}|B^j|}{|A^j|+|B^j|}\int(1-\xi)\gamma_{B^j}[\omega_t^{j\prime}(c)-c]f_{B^j,n_t^{j\prime}}(c)dc\Big)\end{aligned} \tag{13}$$

The case $\lambda_{B^j}=1$ is interesting and worth examining first. In that case, note that the first two terms of the right-hand side of Eq. (13) rewrite as

$$\Big(\frac{|A^j|}{|A^j|+|B^j|}\int \omega_t^j(c) f_{A^j}(c)dc + \frac{|B^j|}{|A^j|+|B^j|}\int [\xi\sigma_t^j \omega_t^j(g(c)) + (1-\xi\sigma_t^j)\omega_t^j(c)] f_{B^j,n_t^j}(c)dc\Big) = \mathbb{E}_t[c|j],$$

since the time t average wage in district j under policy $\overline{T}^j < \infty$ is equal to the time t average performance level in district j under policy $\overline{T}^j < \infty$ (here denoted by $\mathbb{E}_t[c|j]$).

Likewise, the third and fourth terms rewrite as

$$-\Big(\frac{|A^j|}{|A^j|+|B^j|}\int \omega_t^{j\prime}(c) f_{A^j}(c)dc + \frac{|B^j|}{|A^j|+|B^j|}\int [\xi \omega_t^{j\prime}(g(c)) + (1-\xi)\omega_t^{j\prime}(c)] f_{B^j,n_t^{j\prime}}(c)dc\Big) = -\mathbb{E}_t'[c|j],$$

since the time t average wage in district j under policy $\overline{T}^{j\prime} = \infty$ is equal to the time t average performance level in district j under policy $\overline{T}^{j\prime} = \infty$ (here denoted by $\mathbb{E}_t'[c|j]$).

We then have that the right-hand side of Eq. (13) can be written as

$$\begin{aligned}\sum_{t=1}^{\infty}\delta^t\big(\mathbb{E}_t[c|j] - \mathbb{E}_t'[c|j]\big) + \quad & \sum_{t=1}^{\infty}\delta^t\Big(\frac{|A^j|}{|A^j|+|B^j|}\gamma_{A^j}\int(\omega_t^j(c) - \omega_t^{j\prime}(c)) f_{A^j}(c)dc + \\ & \frac{\lambda_{B^j}|B^j|}{|A^j|+|B^j|}\int(1-\xi\sigma_t^j)\gamma_{B^j}[\omega_t^j(c) - c] f_{B^j,n_t^j}(c)dc \\ & -\frac{\lambda_{B^j}|B^j|}{|A^j|+|B^j|}\int(1-\xi)\gamma_{B^j}[\omega_t^{j\prime}(c) - c] f_{B^j,n_t^{j\prime}}(c)dc\Big)\end{aligned}$$

We must now verify if this is greater than 0. We first make the following observations:

- The first summation term is negative and converges to 0 as $\overline{T}^j \to \infty$. Indeed, $\mathbb{E}_t[c|j] < \mathbb{E}_t'[c|j]$ for $t \geq \overline{T}^j$, since the time t average performance level keeps increasing as affirmative action gets implemented for more periods. This term converges to 0 as $\overline{T}^j \to \infty$ since $\mathbb{E}_t[c|j] = \mathbb{E}_t'[c|j]$ for $t < \overline{T}^j$ and $\sup_{t\geq\overline{T}^j}|\mathbb{E}_t[c|j] - \mathbb{E}_t'[c|j]| \xrightarrow[\overline{T}^j\to\infty]{} 0$, reflecting the fact that the improvements in the performance distribution of group B^j become marginal after a while.

- The second summation term is positive and bounded away from 0 as $\overline{T}^j \to \infty$. This captures the gain to the non-beneficiaries (of both groups A^j and B^j) of stopping affirmative action after a finite number of periods. Indeed, under a policy of permanent affirmative action $\overline{T}^{j\prime} = \infty$,

$$\begin{aligned}\omega_t^{j\prime}(c) &= \mathbb{P}_t(\{aa\}|c,j,\{\overline{\sigma}_s\}_{s=1}^t,\hat{\sigma}^{j\prime})g^{-1}(c) + (1-\mathbb{P}_t(\{aa\}|c,j,\{\overline{\sigma}_s\}_{s=1}^t,\hat{\sigma}^{j\prime}))c \\ &< c\end{aligned}$$

  since $\mathbb{P}_t(\{aa\}|c,j,\{\overline{\sigma}_s\}_{s=1}^t,\hat{\sigma}^{j\prime}) > 0$ for all t. Thus, for $t \geq \overline{T}^j$, $\omega_t^j(c) - \omega_t^{j\prime}(c) = c - \omega_t^{j\prime}(c) > \Delta$ for some $\Delta > 0$, while $\omega_t^j(c) - c = c - c = 0$.

From the above observations, we can formally state that $\forall\epsilon > 0$, there exists $\overline{T}^j < \infty$ large enough and $\overline{\delta}(\overline{T}^j) \in (0,1)$ such that $\forall\delta \in (\overline{\delta}(\overline{T}^j), 1)$

$$\sum_{t=1}^{\infty}\delta^t|\mathbb{E}_t[c|j] - \mathbb{E}_t'[c|j]| < \epsilon,$$

and

$$\sum_{t=1}^{\infty}\delta^t\Big(\frac{|A^j|}{|A^j|+|B^j|}\gamma_{A^j}\int(\omega_t^j(c)-\omega_t^{j\prime}(c))f_{A^j}(c)dc+$$
$$\frac{\lambda_{B^j}|B^j|}{|A^j|+|B^j|}\int(1-\xi\sigma_t^j)\gamma_{B^j}[\omega_t^j(c)-c]f_{B^j,n_t^j}(c)dc$$
$$-\frac{\lambda_{B^j}|B^j|}{|A^j|+|B^j|}\int(1-\xi)\gamma_{B^j}[\omega_t^{j\prime}(c)-c]f_{B^j,n_t^{j\prime}}(c)dc\Big)>2\epsilon$$

from which it follows that the right-hand side of Eq. (13) is positive and thus that Eq. (12) is verified.

To complete the proof, we now turn to the case when $\lambda_{B^j}<1$.

First note that when δ is high enough, unsurprisingly, group A^j gains from stopping affirmative action whereas at least a fraction of group B^j loses. Thus, rearranging the left-hand side of Eq. (13) as follows

$$\sum_{t=1}^{\infty}\frac{\delta^t}{(|A^j|+|B^j|)}\big((W_{A^j,t}-W'_{A^j,t})+\lambda_{B^j}(W_{B^j,t})-W'_{B^j,t})\big),$$

we notice that decreasing the weight λ_{B^j} placed on the welfare of group B^j to values strictly smaller than 1 keeps this quantity positive. We can thus conclude that it will still be worth stopping affirmative action after $\overline{T}^j<\infty$ periods as opposed to continuing it forever. The first-best optimal policy $\overline{T}^j_{\lambda_{B^j}}$ for some $\lambda_{B^j}<1$ will thus be such that $\overline{T}^j_{\lambda_{B^j}}\le\overline{T}^j_{\lambda_{B^j}=1}<\infty$.

Since optimal policies are separable across districts, it follows that the above is true for any $j\in J$. This completes the proof. ■

References

Antonovics, K. and B. Backes (2014): “The effect of banning affirmative action on human capital accumulation prior to college entry,” *IZA Journal of Labor Economics*, 3, 1–20.

Arcidiacono, P., E. M. Aucejo, H. Fang, and K. I. Spenner (2011): “Does affirmative action lead to mismatch? A new test and evidence,” *Quantitative Economics*, 2, 303–333.

Arrow, K. (1973): “The theory of discrimination. In: Discrimination in labor markets,” *Orley Achenfelter and Albert Rees (eds.), Princeton–New Jersey.*

Bagwell, K. (1995): “Commitment and observability in games,” *Games and Economic Behavior*, 8, 271–280.

Bar, T., V. Kadiyali, and A. Zussman (2009): “Grade information and grade inflation: The Cornell experiment,” *Journal of Economic Perspectives*, 23, 93–108.

Beaman, L., E. Duflo, R. Pande, and P. Topalova (2012): “Female leadership raises aspirations and educational attainment for girls: A policy experiment in India,” *Science*, 335, 582–586.

Becker, G. S. (1957): *The economics of discrimination*, University of Chicago press.

Bisin, A., A. Lizzeri, and L. Yariv (2015): “Government policy with time inconsistent voters,” *American Economic Review*, 105, 1711–1737.

Chung, K.-S. (2000): "Role models and arguments for affirmative action," *American Economic Review*, 90, 640–648.

Coate, S. and G. Loury (1993a): "Antidiscrimination enforcement and the problem of patronization," *American Economic Review*, 83, 92–98.

Coate, S. and G. C. Loury (1993b): "Will affirmative-action policies eliminate negative stereotypes?" *American Economic Review*, 1220–1240.

Echenique, F. and A. Li (2022): "Rationally inattentive statistical discrimination: Arrow meets Phelps," *arXiv preprint arXiv:2212.08219*.

Fang, H. and A. Moro (2011): "Theories of statistical discrimination and affirmative action: A survey," *Handbook of social economics*, 1, 133–200.

Fershtman, D. and A. Pavan (2021): ""Soft" Affirmative Action and Minority Recruitment," *American Economic Review: Insights*, 3, 1–18.

Fryer Jr, R. G. and G. C. Loury (2005): "Affirmative action and its mythology," *Journal of Economic Perspectives*, 19, 147–162.

Heilman, M. E., C. J. Block, and J. A. Lucas (1992): "Presumed incompetent? Stigmatization and affirmative action efforts." *Journal of Applied Psychology*, 77, 536.

Heilman, M. E., C. J. Block, and P. Stathatos (1997): "The affirmative action stigma of incompetence: Effects of performance information ambiguity," *Academy of Management Journal*, 40, 603–625.

Kahneman, D. and A. Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica*, 47, 263–292.

Kim, Y.-C. and G. C. Loury (2018): "Collective reputation and the dynamics of statistical discrimination," *International Economic Review*, 59, 3–18.

Kydland, F. E. and E. C. Prescott (1977): "Rules rather than discretion: The inconsistency of optimal plans," *Journal of political economy*, 85, 473–491.

Leslie, L. M., D. M. Mayer, and D. A. Kravitz (2014): "The stigma of affirmative action: A stereotyping-based theory and meta-analytic test of the consequences for performance," *Academy of Management Journal*, 57, 964–989.

Lundberg, S. J. and R. Startz (1983): "Private discrimination and social intervention in competitive labor market," *American Economic Review*, 73, 340–347.

Phelps, E. S. (1972): "The statistical theory of racism and sexism," *American Economic Review*, 62, 659–661.

Reardon, S. (2021): "Fewer citations for female authors of medical research." *Nature*.

Sabot, R. and J. Wakeman-Linn (1991): "Grade inflation and course choice," *Journal of Economic Perspectives*, 5, 159–170.

Sander, R. H. (2005): "The racial paradox of the corporate law firm," *NCL Rev.*, 84, 1755.

Sims, C. A. (2003): "Implications of rational inattention," *Journal of monetary Economics*, 50, 665–690.

Tyner, A. and S. Gershenson (2020): "Conceptualizing grade inflation," *Economics of Education Review*, 78, 102037.

Van Damme, E. and S. Hurkens (1997): "Games with imperfectly observable commitment," *Games and Economic Behavior*, 21, 282–308.

7 Supplementary appendix

7.1 Micro-foundations: Wage setting with Bertrand competition

We suppose that each employer (which we can identify with a firm) produces a numeraire good of price equal to 1 with a constant return to scale technology and using labor as the input. The quantity of the numeraire good produced by a unit mass of workers of performance level c is thus simply c. The profit generated by a unit mass of workers of performance level c, when they are paid a wage $\omega_t^j(\overline{c})$, is thus

$$\pi_t(\overline{c}, c) = c - \omega_t^j(\overline{c}).$$

Since a firm only observes the curriculum vitae quality $\overline{c}$ of a district-j worker it hires, the expected profit generated by a district-j worker with such curriculum vitae is then

$$\mathbb{E}_t[\pi_t(\overline{c}, c)|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] = \mathbb{E}_t[c|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] - \omega_t^j(\overline{c})$$

where, as we know, $\mathbb{E}_t[c|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$ is the expected performance level of a district-j worker presenting a curriculum vitae $\overline{c}$, given some conjectured affirmative action policy sequence σ consistent with the observed statistics $\{\overline{\sigma}_s\}_{s=1}^t$.

If the firm hires a mass q^j of district-j workers with curriculum vitae qualities having a density function $f_t^j(\overline{c})$, then its expected profit from the activity of those workers is

$$\begin{aligned}
\Pi_{q^j,t} &= q^j \mathbb{E}_t[\pi_t(\overline{c}, c)|j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] \\
&= q^j \int_{\overline{c}} \mathbb{E}_t[\pi_t(\overline{c}, c)|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] f_t^j(\overline{c}) d\overline{c} \\
&= q^j \int_{\overline{c}} \left(\mathbb{E}_t[c|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] - \omega_t^j(\overline{c})\right) f_t^j(\overline{c}) d\overline{c} \qquad (14)
\end{aligned}$$

where $\Pi_{q^j,t}$ is also the realized profit, since each worker has zero measure.

A firm will thus maximize this profit by choosing an optimal wage function ω_t^j. Note that the profit in Eq. (14) is additively separable across $\overline{c}$. A firm thus chooses, for each curriculum vitae quality $\overline{c}$, the wage $\omega_t^j(\overline{c})$ that maximizes

$$\mathbb{E}_t[\pi_t(\overline{c}, c)|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] = \mathbb{E}_t[c|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] - \omega_t^j(\overline{c}).$$

Since we consider a perfectly competitive Bertrand setting, it follows that the optimal wage will be equal to a worker's expected performance level, i.e. $\omega_t^j(\overline{c}) = \mathbb{E}_t[c|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$, which is the worker's marginal productivity. Indeed, giving a wage higher than $\mathbb{E}_t[c|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$ would result in a negative profit from hiring workers of that curriculum vitae quality, while giving a wage lower than $\mathbb{E}_t[c|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$ would result in another employer hiring the workers away with a slightly higher wage.

It also follows that a firm's profit is zero, i.e. $\Pi_{q^j,t} = 0$, on the equilibrium path. It should be mentioned, however, that including profits in the policy maker's objective function (i.e. in Eq. (3)) still affects the assessment of deviations, as discussed following Proposition 1, and is thus important. Indeed, since $\omega_t^j(\overline{c}) = \mathbb{E}_t[c|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$, a more productive population commands higher salaries, which increases the policy maker's welfare objective.

7.2 A more general model allowing for strategic behavior by workers

We present here a more general model, of which the model presented in the main part of the paper is a particular case. We show that in equilibrium, this more general model endogenously generates a wage function that is non-decreasing in the curriculum vitae quality, thus formally removing the need for Assumption 1.

Here, we allow workers to choose the curriculum vitae quality that they present to employers. This allows us to treat the more general case where the conditional expectation $\mathbb{E}_t[c|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$ may not be monotone. We illustrate that the results presented in the main part of the paper still hold, since they are just a particular case of this more general setting (i.e. the case when workers truthfully declare their curriculum vitae quality).

In this general model, a wage function $\omega_t^j(\hat{c})$ set by employers is the wage the worker earns when declaring a curriculum vitae of quality $\hat{c} \in [0, 1]$ to the employer. Here, we see that a worker can declare a curriculum vitae of quality not necessarily equal to his actual quality $\overline{c}$. This is formalized in the following definition.

Definition 2 *A wage function $\omega_t^j : [0, 1] \to [0, 1]$ determines the wage a worker earns when declaring a curriculum vitae of quality $\hat{c}$ to the employer.*

The utility of a type $(c, \overline{c}, G^j)$ worker, when presenting a curriculum vitae of quality $\hat{c} \in [0, 1]$, is thus

$$u_{G^j,t}(\hat{c}, c) \quad = \quad \omega_t^j(\hat{c}) - \gamma_G \max\{c - \omega_t^j(\hat{c}), 0\} - \kappa \max\{\hat{c} - \overline{c}, 0\} \tag{15}$$

where $\kappa \max\{\hat{c}-\overline{c}, 0\}$, with $\kappa > 0$, is a penalty suffered for cheating (i.e. presenting a curriculum vitae quality higher than the actual one $\overline{c}$). Note that no penalty is suffered for presenting a curriculum vitae of lower quality than $\overline{c}$.

A worker thus chooses to present a curriculum vitae of quality $\hat{c}$ such that

$$\hat{c} \in \underset{\tilde{c} \in [0,1]}{\operatorname{argmax}}\ u_{G^j,t}(\tilde{c}, c)$$

Definition 3 *Given a wage function* $\omega_t^j : [0,1] \to [0,1]$*, a curriculum vitae declaration function* $\mu_t^j : [0,1] \to [0,1]$ *assigns a* declared *curriculum vitae quality* $\hat{c}$ *to an actual curriculum vitae quality* $\overline{c}$*, that is* $\hat{c} = \mu_t^j(\overline{c})$*.*

Definition 4 *Given a putative policy sequence* σ*, a labor market equilibrium* $(\omega_t^{j*}, \mu_t^{j*})$ *is a continuous wage function and a curriculum vitae declaration function such that*

$$\omega_t^{j*}(\hat{c}) = \mathbb{E}_t[c|\hat{c}, j, \mu_t^{j*}, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$$

and

$$\mu_t^{j*}(\overline{c}) \in \underset{\tilde{c}\in[0,\overline{c}]}{\text{argmax}}\ u_{G^j,t}(\tilde{c}, c).$$

Recall from Eq.(15) that the utility $u_{G^j,t}(\hat{c}, c)$ depends on the wage $\omega_t^{j*}(\hat{c})$.

If κ is high enough, a continuous wage function $\omega_t^j(\hat{c})$ will prevent cheating since the marginal penalty of presenting a curriculum vitae quality greater than $\overline{c}$ will exceed the marginal benefit in terms of increased wage. A sufficient condition for this to hold is that $\kappa > \frac{\omega_t^j(\hat{c})-\omega_t^j(\overline{c})}{\hat{c}-\overline{c}}$ for any $\hat{c} > \overline{c}$.

We thus have the following lemma.

Lemma 5 *Suppose* κ *is high enough. Given a putative policy sequence* σ*, there exist intervals* $\{(c_l^L, c_l^H)\}_{l=1}^{\overline{l}}$ *with* $\overline{l} \geq 0$*, so that the (weakly) increasing wage function*

$$\omega_t^{j*}(\hat{c}) = \begin{cases} \mathbb{E}_t[c|\overline{c} = \hat{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] & \text{if } \hat{c} \notin \bigcup_l (c_l^L, c_l^H) \\ \mathbb{E}_t[c|\overline{c} \in (c_l^L, c_l^H), j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] & \text{if } \hat{c} \in (c_l^L, c_l^H) \end{cases} \tag{16}$$

and the curriculum vitae declaration strategy

$$\mu_t^{j*}(\overline{c}) = \begin{cases} \overline{c} & \text{if } \overline{c} \notin \bigcup_l (c_l^L, c_l^H) \\ c_l^L & \text{if } \overline{c} \in (c_l^L, c_l^H) \end{cases} \tag{17}$$

constitute a labor market equilibrium.

In the above,

$$\mathbb{E}_t[c|\overline{c} = \hat{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] = \mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma) \cdot g^{-1}(\overline{c}) + \big(1 - \mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma)\big) \cdot \overline{c},$$

with

$$\mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma) = \frac{|B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c})}{|A^j|f_{A^j}(\overline{c}) + |B^j|(1-\xi\sigma_t^j)f_{B^j,n_t^j}(\overline{c}) + |B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c})},$$

$\{aa\}$ *being the event that a worker benefited from affirmative action and* $(g^{-1})'(\overline{c})$ *the derivative of* g^{-1} *evaluated at* $\overline{c}$*, while*

$$\mathbb{E}_t[c|\overline{c} \in (c_l^L, c_l^H), j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] = \int_{\overline{c}=c_l^L}^{c_l^H} \mathbb{E}_t[c|\overline{c} = \hat{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] f_t^j(\overline{c}) d\overline{c}$$

where

$$f_t^j(\overline{c}) = \frac{1}{|A^j| + |B^j|}\Big(|A^j|f_{A^j}(\overline{c}) + |B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c}) + |B^j|(1-\xi\sigma_t^j)f_{B^j,n_t^j}(\overline{c})\Big)$$

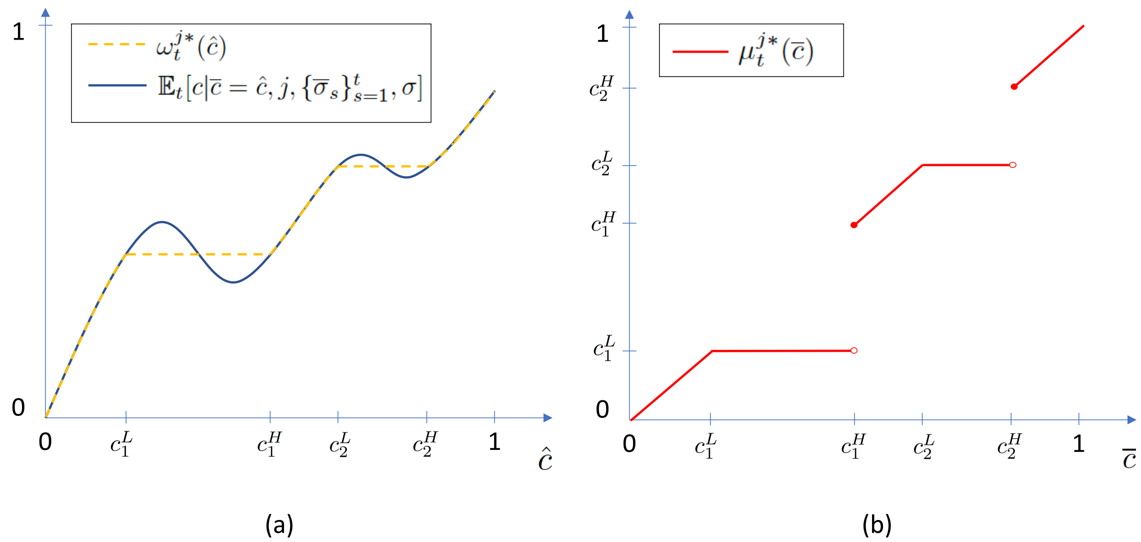

Figure 4: Equilibrium wage function ω_t^{j*} (panel (a)) and curriculum vitae declaration function μ_t^{j*} (panel (b)).

is the overall population density for the curriculum vitae quality at time t in district j.

The equilibrium wage function stated in Lemma 5 has the form described in Figure 4(a). We see that it is weakly increasing, but strictly increasing in certain sections. In the particular case when $\bar{l} = 0$, then it can be strictly increasing over the whole domain, as in the case presented earlier in the main part of the paper. The equilibrium curriculum vitae declaration function has the form described in Figure 4(b). It is such that a worker truthfully declares his curriculum vitae quality, i.e. $\hat{c} = \bar{c}$, when $\bar{c}$ is in an interval where the wage function is strictly increasing, since declaring anything lower would yield a lower salary. On the other hand, when $\bar{c}$ is in an interval where the wage function is flat, the worker declares the lowest curriculum vitae quality $\hat{c}$ on that flat interval, i.e. $\hat{c} = c_l^L$. Indeed, declaring such a curriculum vitae quality $\hat{c} \leq \bar{c}$ provides the worker with the same salary as he would get when declaring the actual one: $\omega_t^j(\hat{c}) = \omega_t^j(\bar{c})$. In the particular case where $\bar{l} = 0$ and the wage function is strictly increasing, then all workers would always declare their true curriculum vitae quality ($\mu_t^{j*}(\bar{c}) = \bar{c}$, as in the case presented earlier in the main part of the paper). Note that, as required by the equilibrium definition, an employer correctly sets the wage equal to the conditional expectation of a worker's performance (i.e. $\omega_t^{j*}(\hat{c}) = \mathbb{E}_t[c|\hat{c}, j, \mu_t^{j*}, \{\bar{\sigma}_s\}_{s=1}^t, \sigma]$).

Lemma 6 *The equilibrium wage function $\omega_t^{j*}(\hat{c})$ is weakly increasing, but strictly increasing at least on some regions of the support*[19] $[0, 1]$.

Lemma 7 *Let $h(c)$ be any weakly increasing function that is strictly increasing at least on some opened subinterval of its support $[0, 1]$ and is differentiable almost everywhere. If $f \succ \tilde{f}$, where f and $\tilde{f}$ are probability density functions on $[0, 1]$ and $\succ$ indicates strict first-order stochastic dominance, then $\int_0^1 h(c)f(c)dc > \int_0^1 h(c)\tilde{f}(c)dc$.*

[19]This is actually stronger than needed. For Propositions 1 and 2 to hold in this more general model, $\omega_t^{j*}(\hat{c})$ only needs to have these properties for the $\hat{c}$'s being played in equilibrium (i.e. $\hat{c} = \mu_t^{j*}(\bar{c})$).

The next lemma is simply a more general version of Lemma 3(i) of the main part of the paper, adapted to the labor market equilibrium concept defined in Definition 4.

Lemma 8 *If a worker benefits from affirmative action (i.e. $c = g^{-1}(\overline{c})$), then he gets a wage higher than his performance level (i.e. $c < \omega_t^{j*}(\mu_t^{j*}(\overline{c}))$). If a worker does not benefit from affirmative action (i.e. $c = \overline{c}$), then he gets a wage lower than his performance level (i.e. $c > \omega_t^{j*}(\mu_t^{j*}(\overline{c}))$).*

Using Lemma 8, we can make the same observations as in the main part of the paper, namely that non-beneficiaries of affirmative action (of either group A or B) suffer a feeling of injustice, while beneficiaries do not.

Lemmas 6, 7 and 8 are all the ingredients needed to confirm that Proposition 1 (permanent affirmative action in equilibrium) and Proposition 2 (temporary affirmative action in the first-best case) of the main part of the paper hold in this more general model. The proofs are otherwise identical.

7.3 Proofs of results in Section 7.2

Proof of Lemma 5. Throughout this proof, we suppose κ is high enough to prevent cheating. A sufficient condition for this to hold is that $\kappa > \frac{\omega_t^j(\hat{c}) - \omega_t^j(\overline{c})}{\hat{c} - \overline{c}}$ for any $\hat{c} > \overline{c}$. In such a case, the marginal penalty of presenting a curriculum vitae quality greater than $\overline{c}$ will exceed the marginal benefit in terms of increased wage.

Step I: Compute the wage $\tilde{\omega}_t^j$ assuming truthful declaration of $\overline{c}$.

Suppose first that workers truthfully declare their curriculum vitae quality, i.e. $\hat{c} = \mu_t^j(\overline{c}) = \overline{c}$. Under such a declaration function μ, call $\tilde{\omega}_t^j(\hat{c}) = \mathbb{E}_t[c|\hat{c}, j, \mu_t^j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$ the conditional expectation of the actual performance level when declaring a curriculum vitae of quality $\hat{c}$. Then,

$$\begin{aligned}
\tilde{\omega}_t^j(\hat{c}) &= \mathbb{E}_t[c|\hat{c}, j, \mu_t^j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] \\
&= \mathbb{E}_t[c|\hat{c} = \overline{c}, j, \mu_t^j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] \\
&= \mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma) \cdot g^{-1}(\overline{c}) + \left(1 - \mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma)\right) \cdot \overline{c}
\end{aligned}$$

and the same argument as in the proof of Lemma 1 of the main part of the paper allows us to state that

$$\mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma) = \frac{|B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c})}{|A^j|f_{A^j}(\overline{c}) + |B^j|(1-\xi\sigma_t^j)f_{B^j,n_t^j}(\overline{c}) + |B^j|\xi\sigma_t^j f_{B^j,n_t^j}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c})},$$

$\{aa\}$ being the event that a worker benefited from affirmative action and $(g^{-1})'(\overline{c})$ the derivative of g^{-1} evaluated at $\overline{c}$.

Step II: Such a wage function $\tilde{\omega}_t^j$ cannot in general be part of an equilibrium.

Suppose that $\tilde{\omega}_t^j$ is increasing for $c \in [0, c_1]$ and decreasing over some interval $[c_1, c_1']$. If the wage function is $\tilde{\omega}_t^j$, then a worker with an actual curriculum vitae quality $\overline{c} \in (c_1, c_1']$ will choose to declare a curriculum vitae quality $\hat{c} < \overline{c}$ since he can obtain a higher wage $\tilde{\omega}_t^j(\hat{c}) > \tilde{\omega}_t^j(\overline{c})$ by doing so. It follows that $\mu_t^j(\overline{c}) = \overline{c}$ cannot be part of an equilibrium since $\mu_t^j(\overline{c}) \notin \underset{\tilde{c} \in [0,\overline{c}]}{\text{argmax}}\, u_{G^j,t}(\tilde{c}, c)$ for such $\overline{c}$.

Since $\mu_t^j(\overline{c}) = \overline{c}$ is not part of an equilibrium, it follows that $\tilde{\omega}_t^j(\hat{c}) = \mathbb{E}_t[c|\overline{c} = \hat{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$ is not equal to the correct conditional expectation $\mathbb{E}_t[c|\hat{c}, j, \mu_t^{j*}, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$ where μ_t^{j*} is an equilibrium declaration function. Thus, $\tilde{\omega}_t^j(\hat{c})$ cannot in general be the equilibrium wage function.

Step III: Building a weakly increasing wage function $\omega_t^{j*}(\hat{c})$ using $\tilde{\omega}_t^j(\hat{c})$.

On the other hand, there exist $c_1^L < c_1$ and $c_1^H \geq c_1'$ such that a wage

$$\omega_t^{j*}(\hat{c}) = \begin{cases} \tilde{\omega}_t^j(\hat{c}), & \text{if } \hat{c} \in [0, c_1^L] \\ \tilde{\omega}_t^j(c_1^L) & \text{when } \hat{c} \in (c_1^L, c_1^H] \end{cases} \tag{18}$$

corresponds to $\mathbb{E}_t[c|\hat{c}, j, \mu_t^{j*}, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$, where μ_t^{j*} is as in the statement of the lemma. Such a pair $\{c_1^L, c_1^H\}$ satisfies

$$\tilde{\omega}_t^j(c_l^L) = \int_{\overline{c}=c_l^L}^{c_l^H} \tilde{\omega}_t^j(\overline{c}) f_t^j(\overline{c}) d\overline{c} \tag{19}$$

$$\tilde{\omega}_t^j(c_l^H) = \int_{\overline{c}=c_l^L}^{c_l^H} \tilde{\omega}_t^j(\overline{c}) f_t^j(\overline{c}) d\overline{c} \tag{20}$$

and

$$\int_{\overline{c}=c_l^H}^{1} \tilde{\omega}_t^j(\overline{c}) f_t^j(\overline{c}) d\overline{c} > \tilde{\omega}_t^j(c_l^H). \tag{21}$$

where

$$f_t^j(\overline{c}) = \frac{1}{|A^j| + |B^j|}\Big(|A^j| f_{A^j}(\overline{c}) + |B^j| \xi \sigma_t^j f_{B^j, n_t^j}(g^{-1}(\overline{c}))/(g^{-1})'(\overline{c}) + |B^j|(1 - \xi\sigma_t^j) f_{B^j, n_t^j}(\overline{c})\Big)$$

is simply the overall population density for the curriculum vitae quality $\overline{c}$ at time t in district j.

By construction, $\omega_t^{j*}(\hat{c})$ is strictly increasing for $\hat{c} \in [0, c_1^L]$ and flat for $\hat{c} \in (c_1^L, c_1^H]$. This is pictured in Figure 4(a). We will generalize this in Step V below.

Step IV: Verifying that $(\omega_t^{j*}, \mu_t^{j*})$ is a labor market equilibrium for $\overline{c} \in [0, c_1^H]$.

For any worker with an actual curriculum vitae quality $\overline{c} \in [0, c_1^L]$, the best response to such a wage function is $\mu_t^{j*}(\overline{c}) = \overline{c} = \operatorname*{argmax}_{\tilde{c} \in [0,\overline{c}]} u_{G^j,t}(\tilde{c}, c)$ since $\omega_t^{j*}(\hat{c})$ is strictly increasing over that range and thus the worker chooses to declare $\hat{c} = \overline{c}$ to maximize his wage. Therefore, $\omega_t^{j*}(\hat{c}) = \mathbb{E}_t[c|\hat{c}, j, \mu_t^{j*}, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] = \mathbb{E}[c|\hat{c} = \overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] = \tilde{\omega}_t^j(\hat{c})$ for $\overline{c} \in [0, c_1^L]$. It follows that ω_t^{j*} and μ_t^{j*} satisfy the labor market equilibrium condition for $\overline{c} \in [0, c_1^L]$.

Moreover, for any worker with an actual curriculum vitae quality $\overline{c} \in (c_1^L, c_1^H]$, the best response set to a such a wage function is $[c_1^L, \overline{c}] = \operatorname*{argmax}_{\tilde{c} \in [0,\overline{c}]} u_{G^j,t}(\tilde{c}, c)$. A worker is indeed indifferent about declaring any $\hat{c} \in [c_1^L, \overline{c}]$, since it yields a salary $\omega_t^{j*}(\hat{c}) = \tilde{\omega}_t^j(c_1^L)$, which is the maximum the worker can obtain. It follows that $\mu_t^{j*}(\overline{c}) = c_1^L \in \operatorname*{argmax}_{\tilde{c} \in [0,\overline{c}]} u_{G^j,t}(\tilde{c}, c)$. Since, $\omega_t^{j*}(c_1^L) = \mathbb{E}_t[c|\hat{c}, j, \mu_t^{j*}, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] = \mathbb{E}_t[c|\hat{c} = c_1^L, j, \mu_t^{j*}, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] = \mathbb{E}_t[c|\overline{c} \in [c_1^L, c_1^H], j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] = \tilde{\omega}_t^j(c_1^L)$, it follows that ω_t^{j*} and μ_t^{j*} satisfies the labor market equilibrium condition for $\overline{c} \in (c_1^L, c_1^H]$.

Step V: Generalizing to $\overline{c} \in [0, 1]$.

If $c_1^H < 1$ and $\tilde{\omega}_t^j(\overline{c})$ is decreasing over some range(s) in $[c_1^H, 1]$, then an iterative application of conditions (19), (20) and (21) allows to find other pairs $\{c_l^L, c_l^H\}$ such that

$$\omega_t^{j*}(\hat{c}) = \begin{cases} \mathbb{E}_t[c|\overline{c} = \hat{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] & \text{if } \hat{c} \notin \bigcup_l (c_l^L, c_l^H) \\ \mathbb{E}_t[c|\overline{c} \in (c_l^L, c_l^H), j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma] & \text{if } \hat{c} \in (c_l^L, c_l^H) \end{cases}$$

and the analysis of Steps II, III and IV generalizes to the rest of the support. ■

Proof of Lemma 6. This is a corollary of Lemma 5.

Lemma 5 states that $\omega_t^{j*}(\hat{c}) = \mathbb{E}_t[c|\overline{c} \in (c_l^L, c_l^H), j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$ for any $\hat{c} \in (c_l^L, c_l^H)$, implying that $\omega_t^{j*}(\hat{c})$ is flat for such $\hat{c}$ (since $\mathbb{E}_t[c|\overline{c} \in (c_l^L, c_l^H), j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$ is a constant).

On the other hand, Lemma 5 states that $\omega_t^{j*}(\hat{c}) = \mathbb{E}_t[c|\overline{c} = \hat{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma]$ when $\hat{c} \notin \bigcup_l (c_l^L, c_l^H)$ and Steps III and V of the proof of Lemma 5 show that $\omega_t^{j*}(\hat{c})$ is constructed so as to be strictly increasing over such intervals. ■

Proof of Lemma 7. The inequality rewrites

$$\int_0^1 h(c) \left[f(c) - \tilde{f}(c)\right] dc > 0.$$

After integrating by parts, this can be written as

$$\left[h(c) \left[F(c) - \tilde{F}(c)\right]\right] |_0^1 - \int_0^1 h'(c) \left[F(c) - \tilde{F}(c)\right] dc$$

where F and $\tilde{F}$ are the CDFs associated with the PDFs f and $\tilde{f}$. The first term is equal to 0 since $F(0) = \tilde{F}(0) = 0$ and $F(1) = \tilde{F}(1) = 1$. Moreover, since $h'(c) \geq 0$ almost everywhere with $h'(c) > 0$ on non-trivial parts of the support, the last term is strictly greater than 0 if $F(c) < \tilde{F}(c)$ for all $c \in (0, 1)$, i.e. if $f \succ \tilde{f}$. ■

Proof of Lemma 8. When $\overline{c} \notin \bigcup_l (c_l^L, c_l^H)$, then from Lemma 5 we know that a worker truthfully declares a curriculum vitae quality $\hat{c} = \overline{c}$ and gets a wage

$$\omega_t^{j*}(\overline{c}) = \mathbb{P}_t(\{aa\}|\overline{c}, j, \{\overline{\sigma}_s\}_{s=1}^t, \sigma) \cdot g^{-1}(\overline{c}) + \left(1 - \mathbb{P}_t(\{aa\}|\overline{c}, j\{\overline{\sigma}_s\}_{s=1}^t, \sigma)\right) \cdot \overline{c}$$

Since $g^{-1}(\overline{c}) < \overline{c}$, it follows immediately that $g^{-1}(\overline{c}) < \omega_t^{j*}(\overline{c}) < \overline{c}$.

Thus, if the worker does not benefit from affirmative action (i.e. $c = \overline{c}$), then $\omega_t^{j*}(\overline{c}) < c$ and he gets a wage lower than his performance level. On the other hand, if the worker benefits from affirmative action (i.e. $c = g^{-1}(\overline{c})$), then $c < \omega_t^{j*}(\overline{c})$ and he gets a wage higher than his performance level.

We now show that this is also true when $\overline{c} \in \bigcup_l (c_l^L, c_l^H)$.

Recall from Lemma 5 that the wage function is flat over $[c_l^L, c_l^H]$ and equal to $\omega_t^{j*}(c_l^L)$. Thus, a worker of performance level c_l^L who does not benefit from affirmative action gets a wage $\omega_t^{j*}(c_l^L)$ with $\omega_t^{j*}(c_l^L) < c_l^L$ and a worker of performance level c_l^H who does not benefit from affirmative action also gets a wage $\omega_t^{j*}(c_l^L)$ and $\omega_t^{j*}(c_l^L) < c_l^H$. Consider now a worker who does not benefit from affirmative action and $\overline{c} \in (c_l^L, c_l^H)$. Then, $c = \overline{c}$ with $c_l^L < c < c_l^H$ and the worker gets a wage $\omega_t^{j*}(c_l^L)$. It follows that $\omega_t^{j*}(c_l^L) < c$ and he gets a wage lower than his performance level. This applies to any $\overline{c} \in \bigcup_l (c_l^L, c_l^H)$.

Now again, recall from Lemma 5 that the wage function is flat over $[c_l^L, c_l^H]$ and equal to $\omega_t^{j*}(c_l^L)$. Thus, a worker of performance level $g^{-1}(c_l^L)$ who benefits from affirmative action gets a wage $\omega_t^{j*}(c_l^L)$

with $g^{-1}(c_l^L) < \omega_t^{j*}(c_l^L)$ and a worker of performance level $g^{-1}(c_l^H)$ who benefits from affirmative action also gets a wage $\omega_t^{j*}(c_l^L)$ and $g^{-1}(c_l^H) < \omega_t^{j*}(c_l^L)$. Consider now a worker who benefits from affirmative action and $\bar{c} \in (c_l^L, c_l^H)$. Then, $c = g^{-1}(\bar{c})$ with $g^{-1}(c_l^L) < g^{-1}(\bar{c}) < g^{-1}(c_l^H)$ and the worker gets a wage $\omega_t^{j*}(c_l^L)$. It follows that $c = g^{-1}(\bar{c}) < \omega_t^{j*}(c_l^L)$ and he gets a wage higher than his performance level. This applies to any $\bar{c} \in \bigcup_l (c_l^L, c_l^H)$. ■