Positive and Negative Affirmative Action

Abstract: Affirmative action continues to divide. My aim in this paper is to present participants in the debate with a new distinction, namely one between negative and positive affirmative action. Whereas positive affirmative action has to do with certain goods, such as a place at a prestigious university or a job at a prestigious company, negative affirmative action has to do with certain bads, such as a firing or a sentence. I then argue that some of the most prominent arguments in favor of affirmative action speak at least as much in favor of negative as positive affirmative action. At the same time, at least one of the most prominent arguments put forward against affirmative action speak less against negative affirmative action. Thus, the paper should redraw the battle lines in the affirmative action debate.

Keywords: affirmative action; goods; bads; injustice; mismatch; sentencing

Forthcoming: *Politics, Philosophy & Economics*

1. Introduction

Affirmative action continues to divide.[1] Whereas defenders maintain that affirmative action is necessary in the unjust societies in which we live, opponents maintain that affirmative action is unjust, for one because it is effectively a form of reverse discrimination. My aim in this paper is to redraw the battle lines in the affirmative action debate. I do so by putting forth a distinction between two forms of affirmative action, namely *negative* and *positive affirmative action*. Positive affirmative action has to do with certain goods, such as a place at a prestigious university, a job at a prestigious company or a seat in parliament. This is the type of affirmative action that is usually defended by proponents and objected to by opponents.

Negative affirmative action, on the other hand, has to do with certain bads, such as a firing or a sentence. Whereas positive affirmative action tries to secure, for some reason, that people from disadvantageous groups get more of these goods, negative affirmative action tries to secure, for some reason, that people from disadvantageous groups get fewer of these bads. It is surprising that negative affirmative action has not received (more) attention. After all, the disadvantage which makes it harder for members of disadvantaged groups to receive certain goods likely also makes it easier for them to receive certain bads. Indeed, I will argue that this is the case. I will also argue that some of the most prominent arguments in favor of affirmative action speak at least as much

in favor of negative as positive affirmative action. At the same time, at least one of the most prominent arguments put forward against affirmative action—the mismatch objection—speak less against negative affirmative action. This is significant for several reasons. It means that opponents of affirmative action, even if they succeed in objecting to positive affirmative action, might not establish that affirmative action *as such* is objectionable. It also means that proponents of affirmative action can provide more by way of argument for affirmative action as such— indeed, more by way of arguments they have already put forward—than has so far been acknowledged. In this way, the distinction I put forward should be of significance to both proponents and opponents of affirmative action.

To clarify, my primary aim in this paper is not to defend affirmative action.[2] My primary aim is to put forward the distinction between negative and positive affirmative action, and situate negative affirmative action in relation to the arguments usually put forward for and against affirmative action. However, the upshot of the latter will in some sense be an indirect defense of affirmative action: indirect in the sense that the arguments in favor of affirmative action speak at least as much in favor of negative affirmative action, and at least one of the arguments put forward against affirmative action has less force against negative affirmative action.

Here is the plan. In the next section (2), I define affirmative action. I start with a definition proposed by Fullinwider, point to a couple of ways in which this definition is too narrow, before I turn to Lippert-Rasmussen's broader definition. In Section 3, I put forward the distinction between negative and positive affirmative action and explain why we should expect that what makes positive affirmative action relevant—some sort of disadvantage—also makes negative affirmative action relevant. In Section 4, I situate negative affirmative action in relation to prominent arguments for and against affirmative action—the compensation argument, the equality of opportunity argument, the diversity argument, the stigma objection, the mismatch objection, and the merit objection. Section 5 responds to two objections to negative affirmative action, namely that it implausibly applies to sentencing, and that it incentivizes wrongdoing. Moreover, I consider an

empirical concern about my overall argument. I conclude in Section 6. If I succeed, the paper should redraw the battle lines in the affirmative action debate by, on the one hand, providing new ammunition to proponents of affirmative action, and, on the other hand, weakening the ammunition available to opponents of affirmative action. But most importantly perhaps, it should engage both sides of the debate.

2. What Is Affirmative Action?

A prominent understanding of affirmative action, proposed by Fullinwider, takes affirmative action to be,

> positive steps taken to increase the representation of women and minorities in areas of employment, education, and culture from which they have been historically excluded (Fullinwider, 2014).[3]

Suppose we reserve twenty spots at a prestigious university to members of a disadvantaged minority group. Insofar as we do this to increase the representation of members of this group at the university, and insofar as members of this group have historically been excluded from the university, we pursue affirmative action in relation to this disadvantaged group according to this definition. This is as it should be. It is a policy we would classify as affirmative action.

However, Lippert-Rasmussen (2020: ch. 1) argues that there are several respects in which Fullinwider's definition is *too narrow* (only some of which I will mention here).[4] First, Fullinwider's definition is *intention-based*, but it is not clear that *effects* do not suffice for something to count as affirmative action. If a scheme like the one mentioned above is put in place by politicians wanting to get re-elected, the intention is not to increase the representation of women and minorities. But it seems that we might still want to say that it is an affirmative action policy (Lippert-Rasmussen, 2020: 4).

Second, on Fullinwider's definition, the *site* of affirmative action is limited to "areas of employment, education, and culture from which women and minorities have been historically excluded." But it is not clear why something cannot count as affirmative action if it takes place in another area. Suppose we reserve twenty places at a prestigious private hospital for patients from a disadvantaged minority group to mitigate the fact that they are unjustly worse off. It is not clear why this should not count as affirmative action. Indeed, Segall (2013: 193-206) has recently defended a form of affirmative action in health. The same goes for other areas. Suppose we reserve twenty apartments in a particularly popular part of a city for members of a disadvantaged minority group, which they can rent at a discounted rate. It is not clear why this initiative should not count as affirmative action, but it does not fall within Fullinwider's definition (Lippert-Rasmussen, 2020: 9).

Third, on Fullinwider's definition, the recipients of affirmative action are limited to "women and minorities." Men and majority members thus cannot *qua* men and majority members be the recipients of affirmative action. Although men and majority members are advantaged *overall*, there might be *local* settings where they are disadvantaged, "e.g., most primary school teachers are women, and perhaps in some cases men might even have been subjected to exclusion in the form of gender policing, or, more likely, have endured the burden of 'sticking out'. It is unclear why a scheme to increase the proportion of male primary school teachers could not be seen as an affirmative action program" (Lippert-Rasmussen, 2020: 11).

Based on these, and related considerations, Lippert-Rasmussen proposes the following, wider definition of affirmative action:

> A policy, an act, etc. amounts to affirmative action if, and only if, in a particular site of justice (i) the agent of the policy, etc. ultimately aims at reasonably increasing the representation of minorities in the relevant area or aims at reasonably addressing the disadvantages they suffer in the relevant area in at least some, but presumably not all,

> ways other than by boosting their representation, or (ii) the relevant policy, etc. will in fact, or is believed to, address a disadvantage of a certain minority group in the relevant area using certain means, e.g., quotas, that go beyond eliminating direct discrimination against the group but not beyond eliminating the relevant disadvantages (Lippert-Rasmussen, 2020: 12).

As is clear, this definition is wider than Fullinwider's definition. For one thing, it allows that an "act", and not only a "policy," may amount to affirmative action. For another, it does not limit the site of affirmative action to "areas of employment, education, and culture," but says instead that it can take place within a particular "site of justice." One reason why Lippert-Rasmussen proposes this broader definition is that he wants to emphasize that affirmative action is not just one thing. Indeed, the definition "brings out nicely that affirmative action policies differ in relation to (1) who the recipients are; (2) who the agents of affirmative action policies are; (3) what the relevant means are; (4) the degree to which the effect is intended to affect an improvement in the recipients' situation; and (5) what the relevant baseline situation is" (Lippert-Rasmussen, 2020: 13). This is, as he says, an important point in itself. It shows that we cannot, at a general level, settle whether affirmative action is morally justified. For these reasons—the narrowness of Fullinwider's and similar definitions, and the important point that affirmative action is not just one thing—I employ Lippert-Rasmussen's definition of affirmative action in what follows.

3. Negative and Positive Affirmative Action

As we have seen, affirmative action is not just one thing. In this section, I want to put forward a distinction between two forms of affirmative action that has not gotten much attention. Let us start with standard cases of affirmative action. Typically, we employ affirmative action in relation to *goods*. We pursue affirmative action when it comes to admissions to prestigious universities (Anderson, 2010: 135; Appiah, 2011; *Grutter v. Bollinger*; *Regents of the University of California v. Bakke*).

When it comes to hiring (Anderson, 2010: 135; Fullinwider, 2014; *Sheet Metal Workers v. EEOC*). And when it comes to being elected to parliament (Gulzar et. al. 2020). In short, we standardly pursue affirmative action in relation to goods: a place at a prestigious university, a job, or a seat in parliament. This is, in some sense, natural. After all, an aim has been, as Fullinwider's definition illustrates, to get women and disadvantaged minorities into beneficial social institutions from which they have been unjustly excluded—to make sure that their opportunity to receive such goods are as good as others' opportunity to receive them. We may refer to this form of affirmative action as

> *Positive affirmative action*: Making it, in some respect, easier for members of disadvantaged groups to receive certain goods (compared to members of non-disadvantaged groups).

"In some respect easier" is to be understood in a broad sense. It captures situations in which we, say, provide bonus points to a person from a disadvantaged minority group when we evaluate their university application. But it is also meant to capture situations such that, if, say, we reserve twenty seats in parliament for members of a disadvantaged group, then we make it easier for persons from this group to receive seats in parliament than for members of a non-disadvantaged group. In short, in some respect, we make it easier for members of disadvantaged groups to receive goods than if they had been members of a non-disadvantaged group (e.g., because it has been more difficult for them to earn the relevant qualifications to begin with because of the disadvantages from which they have suffered).

However, not only is it more difficult for members of disadvantaged groups to receive goods from social institutions. For the same reasons, it will be easier for members of disadvantaged groups to receive *bads* from social institutions (than if they had been members of a non-disadvantaged group). A bad in this sense could be, for instance, to receive a fine, a criminal conviction, or

a prison sentence. One way of seeing that it is easier for members of disadvantaged groups to receive bads is through Goldberg's (2022) argument that, in sexist, racist, and other -ist societies—societies in which some groups are unjustly disadvantaged in relation to others—the evidence will be stacked against members of the groups who are targeted by these -isms. Consider:

> *Disadvantageous Evidence*: Suppose that the news sources one trusts tend to focus on stories of Black men as criminals or athletes or rappers, and rarely feature Black men as intellectuals or scientists or business moguls; that the TV news stories and documentaries one watches tend to reflect these same stereotypes; that the adults in one's community traffic in these very stereotypes without batting an eye; that one's exposure to the Black community is limited, so that one tends not to see Blacks flourishing in intellectually demanding careers or to hear Blacks challenge unfair treatment; and that few if any within one's community draw attention to the injustice of any of this. If you can imagine all of this, you can see that it is possible for members of such a community to come to have evidence that supports skewed views as to the Black community, where their own total evidence does not enable them to recognize this. Now imagine that, on the basis of this sort of total evidence, such a person were to come to have doubts as to the trustworthiness of Black men, and that on the basis of this evidence were to disbelieve a Black male speaker on a given occasion on which he spoke from knowledge (Goldberg, 2022: 391).

In such a situation, Goldberg explains, it is the evidence itself which is stacked against members from the disadvantaged group, in this case Black people. As he says, "it can happen that, unbeknownst to one, the body of evidence one has was itself shaped by the distorting factors of racism or sexism (or some other pernicious -ism) prevalent in one's community [as is the case in Disadvantageous Evidence], such that to disbelieve (or reject) a Black or female speaker's say-so on the

basis of that evidence is to treat them unjustly" (Goldberg 2022: 387). The evidence will be such that it is easier for members of disadvantaged groups to receive negative judgments than members of non-disadvantaged groups. And it is easy to see why this may make it easier for members of such groups to receive certain bads. For instance, it is harder for them to avoid blame partly because it is harder for them to appear trustworthy to begin with. And it is harder for them to avoid being criminally convicted partly because it is harder for them to appear trustworthy, say, before a jury, to begin with (Fricker, 2007).

Another angle from which we can shed light on how it is easier for members of disadvantaged groups to receive bads from social institutions is as follows. As Tadros (2020: 226) explains, "social and political decisions, in conjunction with other facts, determine the rate and distribution of responsibility for wrongdoing." To get an intuitive grasp of this idea, consider:

> *Seating Arrangement*: Billy and Bobby start a new school at age 7, and are very similar. On their first day, Teacher sits them in two free seats in class. In World 1, where Teacher sits Billy next to Jack, Jack becomes Billy's best friend. In a nearby world, World 2, Teacher sits Billy next to John and John becomes Billy's best friend. Things are *vice versa* for Bobby. In World 1, Jack is a bad influence on Billy. When they are 20, he persuades Billy to commit a single crime—a serious insult on Jeff—which Billy does intentionally and without excuse. This does not happen in World 2: Billy does not commit any serious wrongs in that world, as John is a good influence on Billy. Again, things are *vice versa* for Bobby. Other things are equal (Tadros, 2020: 229).

In this case, Teacher's decision determines responsibility for wrongdoing: it determines whether Billy or Bobby later commits a crime. Of course, this is a hypothetical case that has been cleaned of confounding factors to illustrate that Teacher's choice, in conjunction with other factors, determines responsibility for wrongdoing. But the important point is that the same is true of social and

political decisions in the real world (being Billy and Bobby can be seen as being born into an advantaged and a disadvantaged social group). Indeed, Tadros points to Hinton (2016) who shows that two distinct approaches were used against young offenders in the US in the 1970s. Whereas a rehabilitative approach were typically used for white young offenders, a punitive approach were typically used for black young offenders. This made a difference. Significant evidence shows that the rehabilitative approach "resulted in less recidivism and escalation of criminal activity, partly because the punitive approach involved incarceration, which created communities of offenders. This may well have resulted in black young offenders in the 1970s committing more crimes than white young offenders later in life" (Tadros, 2020: 227). The social and political decision of deciding in which cases to use the rehabilitative approach, and in which cases to use the punitive approach, made a difference for responsibility for wrongdoing. It meant that black young offenders would be more likely to commit other wrongs later in life.

As Tadros explains, examples like this one are widespread and familiar,

> Social policies are often designed to prevent wrongdoing or are criticized for failing to do so. Erosion of educational and social facilities for young people are criticized because they make a difference to whether young people offend; transitional processes, practices, and institutions for military personnel, as well as those who have been incarcerated, are needed in part because these people are especially likely to offend or reoffend; urban environments and school buildings need to be restored and protected, because erosion of those environments causes crime (Tadros, 2020: 227).[5]

And this is not only the case when it comes to people acting badly. It is also true when it comes to people acting well. Indeed, "the distribution of educational resources, support to parents, and accolades determines the rate of valuable acts, and who will perform them" (Tadros, 2020: 227). Now, it was probably not just by chance that the punitive approach was chosen for black young

offenders, and the rehabilitative approach was chosen for white young offenders, in the 1970s. More generally, and taking into account Goldberg's remarks from above, it is likely that such social and political decisions, which distribute responsibility for wrongdoing (and "good-doing"), are similarly affected by sexism, racism and other -isms in sexist, racist, and other -ist societies.

In short, it is likely that what makes it harder for members of disadvantaged groups to receive goods from social institutions (such as a place at a prestigious university, a job, etc.) also makes it easier for them to receive bads from social institutions (such as a firing, a prison sentence, etc.) (see also Lippert-Rasmussen, 2010: 173; for a lot of empirical evidence on how members of disadvantaged groups are disadvantaged in such respects, see Alexander, 2010; Anderson, 2010).[6]

Now, if this is true,[7] positive affirmative action is not our only option when it comes to pursuing affirmative action. We could also pursue

> *Negative affirmative action*: Making it, in some respect, harder for members of disadvantaged groups to receive certain bads (compared to members of non-disadvantaged groups).

Negative affirmative action is different from positive affirmative action in being concerned with bads. And whereas, when we pursue positive affirmative action, we make it, in some respect, easier for members of disadvantaged groups to receive certain goods, we make it more difficult, in some respect, for members of disadvantaged groups to receive certain bads when we pursue negative affirmative action.

As far as I am aware, this distinction between positive and negative affirmative action has received almost no attention.[8] As we will see, it can do a lot of work in the affirmative action debate. Before moving forward, however, it is important for me to be clear on why I describe one as "positive," and the other as "negative." Clearly, both positive and negative affirmative action aim to make members of disadvantaged groups better off. Getting a good and avoiding a bad both

make the person better off. So, if that were how I drew the distinction between positive and negative affirmative action, one might be worried that any instance of positive affirmative action could be redescribed as an instance of negative affirmative action, and vice versa. But this is not what I have in mind when I say that one is positive and the other is negative. Instead, it is simply a matter of the item targeted by the affirmative action policy. Some items are goods, such as a place at a prestigious university. Other items are bads, such as a fine or a place in prison. Positive affirmative action is "positive" in the sense that it has to do with those items that are goods. And negative affirmative action is "negative" in the sense that it has to do with those items that are bads. I take it that, for many items, we can somewhat easily group them as goods and bads.[9] In any case, going forward, I will only be concerned with items that pretty clearly are either good or bad.

Interestingly, negative affirmative action amounts to affirmative action on Lippert-Rasmussen's definition. The second part of the definition, recall, says that "the relevant policy, etc. will in fact, or is believed to, address a disadvantage of a certain minority group in the relevant area using certain means, e.g., quotas, that go beyond eliminating direct discrimination against the group but not beyond eliminating the relevant disadvantages" (Lippert-Rasmussen, 2020: 12). Negative affirmative action may precisely address a disadvantage of a certain minority group in the relevant area: the disadvantage that it will be easier for them to receive certain bads, such as punishment, from social institutions than members of advantaged groups. Moreover, some of the examples of negative affirmative action that I will discuss later in the paper, e.g., in sentencing, also satisfy the first disjunct of Lippert-Rasmussen's definition in that they may block discrimination that disadvantaged groups face in this context. So, in one sense, negative affirmative action is just the other side of the coin of positive affirmative action.

Some may oppose Lippert-Rasmussen's definition of affirmative action. They may, for instance, support Fullinwider's definition (which I discussed in Section 2). According to this definition, recall, affirmative action amounts to "positive steps taken to increase the representation of women and minorities in areas of employment, education, and culture from which they have been

historically excluded." Negative affirmative action might also count as affirmative action on this definition. Suppose we pursue negative affirmative action with the ultimate aim of increasing the representation of women and minorities in employment, education, etc., e.g., because we know that a long prison sentence will drastically reduce their chances in employment and education. As long as the ultimate aim is to increase their representation in these areas, it might count as affirmative action on Fullinwider's definition. Even if not, such that some would want to restrict affirmative action to positive affirmative action (or something along those lines, cf. Fullinwider's definition), I am happy with then calling negative affirmative action *shmaffirmative action* (cp. Lippert-Rasmussen, 2020: 245). The reason is that my aims in this paper are not primarily definitional. I want to show, *inter alia*, that at least one of the objections posed against (positive) affirmative action does not apply to the same extent to negative affirmative action, and that prominent arguments in favor of (positive) affirmative action speak at least as much in favor of negative affirmative action. So whether we refer to it as negative affirmative action or shmaffirmative action does not make any substantial difference. The interesting question is how negative affirmative action, or shmaffirmative action, is situated in relation to common arguments in favor of, and objections against, affirmative action (thus, I will continue to refer to it as negative affirmative action).

Before I turn to this question, I want to separate the distinction between negative and positive affirmative action from another distinction that has been drawn in the affirmative action literature. As Lippert-Rasmussen (2020: 18) argues, we may distinguish between *entry-based* and *exit-based* affirmative action. Take the context of jobs. Entry-based affirmative action pertains to recruitment, e.g., providing a benefit to the candidate from a disadvantaged minority group. This is how we usually think of affirmative action. But, as Lippert-Rasmussen points out, there is nothing in principle that precludes us from pursuing affirmative action when it comes to lay-offs, e.g., firing an employee from an advantaged group (instead of not hiring an employee from an advantaged group, or instead of firing the employee from a disadvantaged group). As he argues, it is not clear why we do not focus more on exit-based affirmative action (instead of entry-based affirmative

action). It is not because of losses. After all, the 63-year-old white male professor who gets fired loses out on a few years of employment, whereas the 26-year-old white male applicant might lose out on a life doing what he likes (Lippert-Rasmussen, 2020: 19; Sterba in Cohen and Sterba, 2003: 268-269). In any case, whereas entry-based affirmative action, in the context of jobs, pertains to hirings, exit-based affirmative action pertains to firings. The distinction between entry-based and exit-based affirmative action cuts across the distinction between negative and positive affirmative action. We can pursue entry-based negative affirmative action, e.g., provide a less harsh sentence than if the wrongdoer belonged to an advantaged majority group. But we can also pursue exit-based negative affirmative action, e.g., requiring worse qualifications before firing a person from a disadvantaged minority group (than a person from an advantaged majority group).[10] Thus, the distinction I pose between negative and positive affirmative action should not be confused with the distinction between entry-based and exit-based affirmative action.

4. Negative affirmative action and pro et contra affirmative action

I have put forward the distinction between negative and positive affirmative action. Whereas the former pertains to bads, the latter pertains to goods. In this section, I want to situate negative affirmative action in relation to common pro et contra affirmative action arguments. That is, I want to show that some of the most prominent arguments in favor of affirmative action speak at least as much in favor of negative as positive affirmative action; and that at least one of the most prominent arguments put forward against affirmative action—the mismatch objection—speak less against negative than positive affirmative action.

4.A Compensation

A common argument in favor of affirmative action is the *compensation argument* (Anderson, 2010; Lippert-Rasmussen, 2020: ch. 2; Sher, 2002; *United Steel Workers of America v. Weber*). According to this argument, affirmative action seeks to benefit members of groups that are victims of past

injustice, either directly (where they themselves have been subject to injustice in the past) or indirectly (where their ancestors have suffered from injustice). A paradigm example is slavery in the United States. My aim here is not to evaluate this argument.[11] My aim here is to point out that, *if* this argument speaks in favor of positive affirmative action, it also speaks in favor of negative affirmative action. The reason is simply that both forms of affirmative action can be seen as compensation for past injustice (cp. Butler, 1997). Positive affirmative action can be seen as compensation for the fact that it has been harder for members of the group having suffered from past injustice to obtain the qualifications necessary to obtain certain goods, such as a place at a prestigious university or a job at a prestigious company. Negative affirmative action can be seen as compensation for the fact that it has been easier for members of the group having suffered from past injustice to obtain the "qualifications" necessary to obtain certain bads, such as a firing or a prison sentence.[12]

4.B Equality of opportunity

Another prominent argument in favor of affirmative action is the *equality of opportunity argument*. Sher (2002: 61) puts forward this argument,

> the key to an adequate justification of reverse discrimination [affirmative action] [is] to see that practice, not as the redressing of *past* privations, but rather as a way of neutralizing the *present* competitive disadvantage *caused* by those past privations and thus as a way of restoring equal access to those goods which society distributes competitively (see also Beauchamp, 2002, 214; Cohen and Sterba, 2003: 231; Harris and Narayan, 2014; Lippert-Rasmussen, 2020: ch. 4; Sotomayor's dissent in *Schuette v. Coalition to Defend Affirmative Action*; Taylor, 2009: 478).

Thus, according to this argument, affirmative action is needed because the playing field is not level. Members of disadvantaged groups have, in general, worse opportunities than members of advantaged groups, where we understand equality of opportunity such that

> X and Y enjoy substantive equality of opportunity (vis-á-vis one another) with regard to a certain position, P, if, and only if, when X and Y have the same native talent required for the position and the same ambition (i.e., they commit the same level of efforts to achieve the relevant position), they enjoy equal chances of getting it (Lippert-Rasmussen, 2020: 78; see also Rawls, 2001: 44).

We will assume that the position, P, can both be an advantageous and a disadvantageous position (in the latter case, it might be read, for instance, such that if the majority and the minority person are equally unqualified, they enjoy equal chances of getting fired). Now, if my discussion in Section 3 is sound, then the equality of opportunity argument should speak at least as much in favor of negative affirmative action as positive affirmative action. Recall Goldberg's argument that in racist, sexist and other -ist societies, the evidence will be stacked against members of the groups suffering from these -isms. In general, this should be the case as much when it comes to evidence pertaining to receiving bads as when it comes to receiving goods. Indeed, if Goldberg is right, the evidence will be stacked against members of disadvantaged groups in the sense that they will appear more "qualified" in relation to receiving bads and less qualified in relation to receiving goods than members of an advantaged group. Even if the minority and the majority individual put in the same level of effort, they do not enjoy equal chances of getting it, irrespective of whether it is an advantageous or a disadvantageous position, as long as the evidence itself is skewed, assuming that the evidence will not be more skewed when it comes to those who are worst off within the disadvantaged group (if it would, the equality of opportunity argument would speak more in favor of negative than positive affirmative action, as we will see in relation to the mismatch objection). And note,

importantly, that this is the case even if the evaluator is not biased in any sense, but simply decides in accordance with the evidence. Thus, if the equality of opportunity argument provides a reason to pursue positive affirmative action, it provides at least a reason of the same strength to pursue negative affirmative action.[13]

4.C Diversity

A third prominent argument in favor of affirmative action is the *diversity argument*. In court, Justice Powell put forward this argument in relation to attaining a diverse student body:

> This [a diverse student body] clearly is a constitutionally permissible [438 U.S. 265, 312] goal for an institution of higher education . . . Ethnic diversity, however, is only one element in a range of factors a university properly may consider in attaining the goal of a heterogeneous student body. Although a university must have wide discretion in making the sensitive judgments as to who should be admitted, constitutional limitations protecting individual rights may not be disregarded (*Regents of the University of California v. Bakke 438 U.S. 265*, 1978; see also *Grutter v. Bolinger 539 U.S. 306*, 2003).

As in this example, most often the diversity argument is put forward in an educational context, i.e., that having a diverse student body is good (see also Anderson, 2010: 141; Appiah, 2011: 278; Bowen and Bok, 2002: 179). But the scope of the diversity argument is not limited to educational contexts. Indeed, it has also been pointed out that a diverse work force promotes profitability and efficiency (Lippert-Rasmussen, 2020: 124). What is integral to these different versions of the diversity argument is that they take diversity to be valuable, not in itself, but because of something else that diversity brings, such as better education and productivity.

Now, we might believe that the diversity argument does not speak in favor of negative affirmative action. After all, why would diversity be valuable when it comes to those who receive

bads? But this might be too quick. Suppose we pursue negative affirmative action when it comes to prison sentences such that we reduce the sentence, to some extent and in relation to some types of crime, if the convicted is a member of one of several disadvantaged groups. Suppose also that members of disadvantaged groups are overrepresented in prison (in relation to their share of the population) compared to members of advantaged groups. If so, pursuing negative affirmative action in this sense might actually increase diversity in the prison population. Is there a reason to think that a diverse prison population would be good?

Suppose Appiah (2011: 276; see also, e.g., Alexander, 2010; Cholbi and Madva, 2018; Hunt, 2015; Levinson et. al., 2014; Lynch and Haney, 2011) is right that,

> on average, a black person enters most public contexts with a serious risk of paying higher psychic and material costs than otherwise identical white people … Police officers are more likely to stop you and more likely to arrest you after stopping you. Indeed, you are more likely to be racially profiled in criminal justice contexts. Prosecutors are likely to give you worse plea deals and ask for longer sentences. Juries are more likely to convict you and judges are likely to give you longer sentences than similarly accused whites.

If that is true, then securing a more diverse prison population through negative affirmative action might secure fairness, or at least less unfairness (cp. Butler, 1997: 853). In that case, diversity could be valuable because fairness is valuable, similar to how diversity can be valuable because educational achievement is valuable.[14] And the diversity argument would, insofar as it speaks in favor of positive affirmative action, also speak in favor of negative affirmative action, although the specifications of the argument would be different on the two. Of course, this only points to one particular reason for why diversity might be good when it comes to negative affirmative action. There might be others, e.g., perhaps diversity in the prison population is good because it leads to major social institutions being better integrated (cp. Anderson, 2010).[15] Or perhaps diversity in the prison

population is good because it increases diversity outside of prisons. The important point is that even if we might initially have thought so, the diversity argument is not restricted to positive affirmative action (insofar as it works in that case).

At this point, one might point out that I have been assuming a *proportional* understanding of diversity. This understanding is assumed by Sher: "it [the diversity argument] is the argument that preferential treatment is justified when, and because, it moves us closer to a situation in which the holders of every (desirable) type of job and position include representatives of all racial, sexual, and ethnic groups in rough proportion to their overall numbers" (Sher, 2003: 193). The case I discussed above—negative affirmative action in prison sentencing—is one where we can expect increased diversity in the proportional sense. As of February 10, 2024, 38,7 % of inmates in the US are black (Federal Bureau of Prisons, 2024), whereas only 13,6 % of the US population is black United States Census Bureau, 2023). Inasmuch as the negative affirmative action proposal would reduce the number of black inmates, it would secure more diversity in the proportional sense. But, one might point out, there are other understandings of diversity. One might support a *numerical* understanding of diversity. If a majority within a population has a particular trait, it does not increase diversity to add even more people with this trait. This makes a difference in the prison example. As of February 10, 2024, 57,2 % of inmates are white. So negative affirmative action in prison sentencing in favor of black people, expectedly leading to more white prisoners, will decrease diversity in the numerical sense. So it seems that my argument requires the proportional understanding of diversity.[16] However, it is important to keep the dialectics in mind. First, negative affirmative action in the prison context is not different from positive affirmative action when it comes to numerical diversity. Positive affirmative action in favor of black people when it comes to granting parole would, like negative affirmative action in the sentencing case, also decrease diversity in the numerical sense. Positive and negative affirmative are not differently situated in this respect. Second, even assuming the numerical understanding of diversity, some forms of negative affirmative action would increase diversity, e.g., negative affirmative action in relation to firings at

prestigious workplaces where white people are numerically overrepresented. Thus, there is no reason to think that the diversity argument is restricted to positive affirmative action nor that the diversity argument necessarily speaks less in favor of negative than positive affirmative action.

4.D Mismatch

We have now analyzed negative affirmative action in relation to three prominent arguments in favor of affirmative action.[17] Let us now turn to some prominent objections to affirmative action. One common objection is the *mismatch objection*. Cohen (Cohen and Sterba, 2003: 31) presents this objection,

> It is one of the great ironies of "affirmative action" that those among minority groups receiving its preferences are precisely those least likely to deserve them.

The mismatch objection points out that there is a mismatch when we pursue affirmative action between those who ought to benefit, because they have been disadvantaged, and those who will actually benefit (Fullinwider, 1980: 53-56; Justice Scalia in *Fisher v. University of Texas*; Mulligan, 2018; Pojman, 2014: 438; *Richmond v. J. A. Croson Co*; Sher, 2002). As Lippert-Rasmussen (2020: 190; see also Segall, 2013: 202) explains, "affirmative action at universities benefits the, comparatively speaking, privileged minority people who, relative to other minority members, are those who have suffered the least from various discrimination-related injustices." That is, typically, when we pursue positive affirmative action, those who will be the actual beneficiaries will be those who are well enough placed to actually apply and compete for a place at a prestigious university, or for a job at a prestigious company. And those will likely be the better off within the disadvantaged group (Khaitan, 2015: 224). In this way, affirmative action is *underinclusive* when it comes to the recipients of affirmative action. But this is not the only mismatch. The mismatch objection also points out that affirmative action is *overinclusive* when it comes to those who bear the costs of affirmative

action. The worst off within the advantaged group—such as "the poor white male from Appalachia (Lawrence and Matsuda, 1997: 190-191)—bear at least some of the costs, even though they may have suffered relevantly similar injustices as recipients of affirmative action (Lippert-Rasmussen, 2020: 191).

In relation to underinclusiveness, negative affirmative action is different from standard forms of positive affirmative action, i.e., those in relation to which the mismatch objection has been raised, such as affirmative action at prestigious universities and companies. This is so because, for many of the bads—such as prison sentences—the worst off within the disadvantaged group are more likely to be among the recipients in the first place (even if we take into account, as pointed to by Appiah, the harms suffered by every member of the group qua group member). And this means that it is not the best off within the disadvantaged group who will be among the most likely recipients of such forms of negative affirmative action. In general, we can expect less of a mismatch with such forms of affirmative action than with standard forms of positive affirmative action. Of course, this is not to say that there will be no mismatch at all. It is only to say that it is likely that the mismatch will be smaller.

But there are some forms of negative affirmative action where it is likely that we are within the range of the best off within the disadvantaged group. Suppose that we pursue negative affirmative action at prestigious universities and companies, e.g., by requiring worse qualifications before we "fire" them (in the university case, by, say, requiring worse grades before we throw them out of the program). Here the recipients are likely to be among the better off within the disadvantaged groups. But this does not change the fact that since negative affirmative action is concerned with bads, we can in general expect less underinclusiveness than in the case of positive affirmative action.

The overinclusiveness part in relation to negative affirmative action is trickier. It is trickier because it is not clear what the costs are in some cases of negative affirmative action. Take again the example with reducing prison sentences for some types of crime for members of disadvantaged

groups. That a minority member has to serve less time (than they otherwise would) does not entail that a majority member has to serve more time. This is different when it comes to at least standard forms of positive affirmative action: if the minority member is a recipient of affirmative action in relation to hiring, or the place at the university, it reduces the majority member's chances of getting the job or the place. Of course, there will be some forms of negative affirmative action—e.g., in cases of firings to reduce costs—where the majority member's risk of being fired increases when we pursue negative affirmative action (and where it will likely be the worst off within the best off of the advantaged group who will pay the costs). And there will also still be the problem that some of the most disadvantaged majority members may have suffered the same injustices as the minority members, but will not receive affirmative action.[18] But there at least seems to be some cases of negative affirmative action where it is unclear that costs will be borne by disadvantaged majority members.[19] In any case, the mismatch objection is in general less of an objection to negative affirmative action than positive affirmative action, at least because underinclusiveness can be expected to be less of a problem with the former than the latter.

4.E Stigma

Another prominent objection to affirmative action is the *stigma objection*. Cohen (Cohen and Sterba, 2003: 121; see also Beauchamp, 2002: 216; Lippert-Rasmussen, 2020: ch. 9; Thernstrom and Thernstrom, 2002: 187) dramatically puts forward this objection,

> If some demon had sought to concoct a scheme aimed at undermining the credentials of minority scholars, professionals, and students, to stigmatize them permanently and humiliate them publicly, no more ingenious plan could have been devised than the system of preferences now defended as a social need and great favor to minorities.

In essence, the stigma objection points out that affirmative action stigmatizes its recipients, e.g., it may result in people questioning whether females or minority candidates got hired for a prestigious job because of their gender or race, and it may even lead recipients to pose this question as well, with threats of damages to their self-esteem (Lippert-Rasmussen, 2020: 173).[20] Moreover, that affirmative action stigmatizes its recipients has some empirical backing (Deshpande, 2019; Lippert-Rasmussen, 2020: 174; but see Anderson, 2010).

It might be that, in many instances, positive affirmative action stigmatizes the recipients—those, say, admitted to prestigious universities and hired to prestigious jobs—but actually benefits the rest of the group of which the recipients are members. This might be because of the *sheer numbers effect*: simply seeing women and minority groups appear as experts, CEOs, professors, etc. might lead people to believe that people from these groups are more talented and competent than they would believe if they only saw white men as experts, CEOs, professors, etc. (Lippert-Rasmussen, 2020: 182). If so, it might benefit members of the group as a whole, even though the recipients of affirmative action might be stigmatized.

Now, whether negative affirmative action stigmatizes its recipients is ultimately an empirical question. Since negative affirmative action is not pursued, we do not have empirical data on this. But I suspect that negative and positive affirmative action is similarly situated when it comes to the stigma objection. Just as some forms of positive affirmative action—say, in admissions to university—may lead to stigma, both from those who lose out and the recipients, some forms of negative affirmative action may also lead to stigma. An illustration may be helpful. One way of pursuing negative affirmative action, recall, is when it comes to firings. Suppose that we pursue negative affirmative action when it comes to firing at a prestigious workplace, such that we fire the majority candidate, who is slightly more qualified (but not so qualified that it would be unreasonable to fire them), than the minority candidate because the latter belongs to a disadvantaged group. After the firing has taken place, other people at the workplace might stigmatize the minority candidate, e.g., thinking "the only reason they were not fired was because of their group membership,"

and if the minority candidate believes, justifiably or not, that the other candidate was more qualified than them, they may suffer a loss of self-esteem. In short, hiring as a form of positive affirmative action and firing as a form of negative affirmative action seem to be equally vulnerable to the stigma objection. And I suspect that the same will be true of other instances of positive and negative affirmative action as well. But, again, it is ultimately an empirical question. It seems, though, that we have no particular reason to think that the stigma objection should speak more against negative than positive affirmative action (irrespective of whether it speaks against positive affirmative action or not, cf. Anderson, 2010).

4.F Merit

A third prominent objection to affirmative action is the *merit objection*. According to this objection, "affirmative action clashes with the principle that positions should be open for competition and the best qualified candidate selected" (Lippert-Rasmussen, 2020: 230; see also Cavanagh, 2002: 33; Pojman, 2014: 440-441).[21] Thus, the objection focuses on the entitlements of the best qualified, e.g., to the spot at the university or the job. Suppose the best qualified applicant is the majority candidate. Suppose we pursue positive affirmative action, such that we hire the minority candidate who is a bit less qualified than the majority candidate. According to the merit objection, the majority candidate has been wronged: as they were the best qualified candidate, they were entitled to the position.

The parallel situation in negative affirmative action is one in which we must fire someone. The minority candidate is the least qualified, and the majority candidate is a bit more qualified. Suppose we pursue negative affirmative action, such that we fire the majority candidate instead of the minority candidate. If the merit objection travels to cases of negative affirmative action, then this majority candidate has been wronged: as they were not the least qualified candidate, they were entitled not to be fired. In principle, it seems that, if the best qualified has an entitlement in cases of positive affirmative action, the next-to-least-qualified candidate has an entitlement in cases of

negative affirmative action. In that regard, the objection seems to apply equally to the two forms of affirmative action (which is not necessarily to say that it is a good objection to any form of affirmative action).[22]

5. Objections to negative affirmative action

We have now seen that prominent arguments in favor of affirmative action speak at least as much in favor of negative as positive affirmative action. Moreover, we have seen that whereas some objections speak equally against negative and positive affirmative action, at least the mismatch objection seems to speak more against positive than negative affirmative action. Now, there might also be objections to negative affirmative action that do not speak, or at least not to the same extent, against positive affirmative action. In this section, I would like to consider two such objections. I will also consider a concern that one might have about the empirical assumptions underlying my arguments.

First, I have, at several places, referred to negative affirmative action in sentencing. Some might be particularly wary when it comes to this form of affirmative action. They might think that sentencing is not an area in relation to which affirmative action should take place because it is a particularly serious unfairness to those who receive a longer sentence compared to an affirmative action recipient. I have several responses to this objection. First, suppose it is true that negative affirmative action in sentencing is unfair to those from the majority group who will not receive affirmative action for committing identical crimes as recipients of affirmative action. Still, in terms of unfairness, negative affirmative action in sentencing might still be better than the status quo.[23] Take the case of the US. "A wide body of studies indicate," Cholbi and Madva (2018: 517) explain, "that (*a*) black capital defendants are more likely to be subject to execution than defendants of other races and (*b*) those who murder blacks are less likely to be subject to execution than are those who murder members of other races." As this shows, the status quo in the US is filled with unfairness in sentencing (remember also the arguments by Goldberg and Tadros that we considered

in Section 2). Because of this, introducing negative affirmative action might lead to less unfairness in sentencing overall. We must remember not to compare the situation in which we pursue negative affirmative action with a situation in which there is no unfairness, or injustice more broadly (cp. Lippert-Rasmussen, 2020: 171). After all, affirmative action, according to most, only becomes relevant once we are in an unjust society (see, e.g., Adams, 2021; Taylor, 2009; but see Meshelski, 2016 for an opposing view).

Second, even if the unfairness in sentencing objection succeeds, it does not speak against negative affirmative action *as such*. Sentencing is merely one example of negative affirmative action. We can imagine many others. One which we have discussed at several points in the paper is negative affirmative action when it comes to firings. The objection does not apply to this variant of negative affirmative action.

Third, in this paper, I have mostly been interested in the relationship between negative and positive affirmative action, as opposed to justifying negative affirmative action as such. This is why I have focused on situating negative affirmative action in relation to the arguments usually posed for and against affirmative action. Interestingly, as we have seen, the mismatch objection speaks less against negative affirmative action in prison sentencing than against standard forms of positive affirmative action. This suggests that if one believes that negative affirmative action in sentencing is particularly objectionable, one should be wary of the mismatch objection. One will instead have to rely on another argument.

Fourth, it is not even clear that negative affirmative action in sentencing is special in the sense suggested by the objection, at least not if we compare it to a particular form of positive affirmative action. Sometimes, positive affirmative action is practiced in relation to political office. This is, for instance, the case in India where local political office in the Scheduled Areas is reserved for the historically disadvantaged Scheduled Tribes (Gulzar et. al., 2020). Suppose we pursue positive affirmative action by reserving some seats in parliament for members of disadvantaged groups. It is not clear why negative affirmative action in sentencing should be deemed more special

than this form of positive affirmative action. After all, those who will receive the seats in parliament can partake in deciding the laws (including those involving punishment) (cp. Kolodny, 2014: 305-307). In that sense, it seems that negative affirmative action in sentencing is not relevantly different from positive affirmative action in parliament. At least, we would need an argument for why this should be so. And, in any case, and as I said in my third response, my primary aim has not been to justify negative affirmative action, but to situate it in relation to positive affirmative action and common arguments for and against affirmative action.

A second, and related, objection is that negative affirmative action *incentivizes wrongdoing*. For instance, if a minority candidate knows that, because of negative affirmative action, it will be harder for them to get fired, they might have less of an incentive not to do wrong at work. And the same might be true of negative affirmative action in sentencing: if a minority candidate knows that they, because of affirmative action, might receive a less harsh sentence than they otherwise would have received, they might have a stronger incentive to do wrong. A policy which incentivizes wrongdoing is objectionable for that reason. Thus, negative affirmative action is objectionable. I have the following responses. First, arguably, policies that actually lead to wrongdoing are worse than policies that incentivize wrongdoing, all else equal. This might make a difference in this context. We know, as we have seen previously, that there is plenty of wrongdoing in the status quo. And even if negative affirmative action incentivizes wrongdoing, it might not lead to much wrongdoing since "agents do not always respond so straightforwardly to the law's incentives" (Cholbi and Madva, 2018: 525, referring to Glaser, 2014: ch. 5). Even if it does, there might still be more wrongdoing, and incentive to do wrong, in the status quo than in the situation where we pursue negative affirmative action.

Second, clearly there are other considerations than wrongdoing when we determine which policies to pursue. In this sense, we might have a *pro tanto* reason to pursue negative affirmative action, even if that is not what we should do *all things considered*. So if negative affirmative action incentivizes wrongdoing, we might still have a pro tanto reason to implement it—e.g., because the

status quo is unfair to members of disadvantaged groups—even if, perhaps, we should not implement it all things considered.

Third, a related objection can be posed, and has been posed, to positive affirmative action, namely that it incentivizes slacking, or less "good-doing." Since minority candidates know that they will receive affirmative action, they will have more incentive to slack and not do as much good as they otherwise would (Cohen and Sterba, 2003: 127; Lippert-Rasmussen, 2020: 249; Loury, 2003). We must not forget here that it might, at the same time, incentivize non-recipients of affirmative action to put in even more effort and do even more good, because they know that it will be harder for them to be selected (Lippert-Rasmussen, 2020: 249); similar to how the status quo, with a lack of negative affirmative action, might incentivize individuals to commit wrongdoing against members of disadvantaged groups because they know they will receive less punishment (Cholbi and Madva, 2018). So these objections do not seem to be good objections to negative and positive affirmative action. And, in any case, since I am primarily interested in how the two forms of affirmative action are situated in relation to each other, what is important is that a similar objection can be raised against positive affirmative action.

Let me finally end by responding to an empirical concern that some may have. One might object that my paper gives the sense that current circumstances are always unjust in a way that could be corrected, at least in part, by affirmative action. But is this really the case universally? Even if it is true in the US when it comes to criminal justice, is it true in, say, Sweden in all spheres of social life? I would like to emphasise that this paper has been meant primarily as a philosophical exploration. In that sense, I do not mean to lean too much on what actually happens to be the case. Although I suspect that the empirical picture is leaning in the direction that at least some of the injustice could be corrected by affirmative action (Adams, 2021; Gulzar et. al., 2020; Khaitan, 2015: 8), I do not want to rely too much on this fact (so in that sense, I am happy to acknowledge that the empirical picture may be muddier). I have intended my argument to be at least partially independent from actual empirical circumstances, in the sense that I have explored the distinction

between negative and positive affirmative action under certain empirical circumstances that are in some sense favorable for affirmative action.[24] At the same time, I have also explored objections to affirmative action under empirical circumstances that may not be conducive to affirmative action, e.g., I have assumed that affirmative action recipients may be stigmatized and that there might be merit concerns. So I have mostly tried to conduct a philosophical discussion which has been informed by what I have taken to be somewhat realistic empirical circumstances. At the end of the day, I am a philosopher, and not a social scientist. And the distinction between positive and negative affirmative action is theoretically interesting, even if the empirical picture is somewhat muddy.

6. Conclusion

My aim in this paper has been to redraw the battle lines in the affirmative action debate. I have done so by providing a distinction between two forms of affirmative action, namely negative and positive. Whereas prominent arguments in favor of affirmative action speak at least as much in favor of negative as positive affirmative action, at least one of the most prominent arguments against affirmative action—the mismatch objection—speak less against negative affirmative action than positive affirmative action. (It is important to notice that even if you disagree with some of the conclusions I reached in relation to particular affirmative action arguments, the distinction between negative and positive affirmative action should still be of interest.) This is, among other things, dialectically important. It means that opponents of affirmative action, even if they succeed in objecting to positive affirmative action, might not establish that affirmative action *as such* is objectionable. It also means that proponents of affirmative action can provide more by way of argument for affirmative action as such.

My argument may also, at least to some extent, speak to contemporary political and legal issues. In the recent US Supreme Court case, *Students for Fair Admissions, Inc v. President and Fellows of Harvard College*, affirmative action has been struck down as unconstitutional qua violating the Equal Protection Clause of the 14th Amendment. The verdict seriously threatens affirmative action

in American universities. But what has been tried in court is positive affirmative action. If my arguments in this paper are correct, there may be some reason to treat negative affirmative action differently. This becomes more evident when we consider the well-established legal practice of *mitigation*. As Atiq and Miller (2018: 169) point out, "it is well-established law, since *Eddings v. Oklahoma*, that evidence of "severe environmental deprivation" (SED)-such as egregious child abuse, neglect, or poverty-must be "considered" by judges as a mitigating factor during the penalty phase of capital trials."[25] A disadvantageous background is, in this sense, a mitigating factor. Negative affirmative action in sentencing can be seen as a way of expanding legal mitigation to include race as a mitigating factor.[26] If negative affirmative action may be seen as a form of legal mitigation, this seems to raise two possibilities with regard to positive affirmative action. Either positive affirmative action may, like negative affirmative action, be seen as a form of legal mitigation. If so, this puts some pressure on the verdict that positive affirmative action is necessarily unconstitutional. Or positive affirmative action is, legally speaking, relevantly different from negative affirmative action in not being a form of legal mitigation, in which case the recent Supreme Court case may be seen to threaten positive affirmative action. But then it is not clear that it necessarily threatens negative affirmative action. Space precludes me from going further into these issues, but my brief discussion illustrates how the distinction laid out in this paper between negative and positive affirmative action may usefully inform discussions of contemporary political and legal issues. We must remember that affirmative action does not have to do with goods. It may also have to do with bads.

Notes

[1] As the recent Supreme Court case in the US— *Students for Fair Admissions, Inc v. President and Fellows of Harvard College*—clearly illustrates.
[2] For arguments to this effect, see, e.g., Adams (2021); Anderson (2010); Appiah (2011); Cahn (2002); Cohen and Sterba (2003); Lippert-Rasmussen (2020).

[3] Another common definition which is, in many respects, in line with Fullinwider's definition is Anderson's (2010: 135) definition according to which affirmative action is "any policy that aims to increase the participation of a disadvantaged social group in mainstream institutions, either through 'outreach' (targeting the group for publicity and invitations to participate) or 'preference' (using group membership as criteria for selecting membership." It may be wider than Fullinwider's definition in that a social group may be disadvantaged now, even if it has not been historically excluded. It may be narrower than Fullinwider's definition insofar as only political agents can implement a policy (she refers to "any policy").

[4] If we assume that Fullinwider is after a fully worked out definition, stating necessary and sufficient conditions for affirmative action (he might not be) (Lippert-Rasmussen, 2020: 2).

[5] To be clear, Tadros's concern is with the distribution of responsibility, and not with whether it is appropriate to hold wrongdoers responsible in the first place. What I mean to illustrate by pointing to Tadros's argument is that responsibility is distributed both when it comes to good-doing and wrongdoing, such that we should expect this to ultimately affect distributions of both goods and bads. I do not mean to suggest that Tadros's argument by itself has any clear upshot when it comes to affirmative action for members of disadvantaged groups.

[6] This leads Tadros (2020: 224) to conclude, "in distributive justice, there is at least some pressure to allocate welfare or resources to those who are responsible for wrongdoing, and away from those who are responsible for good deeds."

[7] For my purposes, it does not even have to be the same disadvantages doing the work when it comes to goods and bads. Moreover, even if matters are more complicated empirically speaking, the distinction that I draw between negative and positive affirmative action is still interesting from a conceptual point of view (see also the empirical concern that I address in Section 5).

[8] But as two anonymous reviewers have pointed out to me, Butler (1997; see also 2021) defends affirmative action in criminal law. As he says, "in addressing the problems of African Americans, affirmative action largely has been limited to the contexts of education, employment, and voting. Affirmative action has ignored one of the most troubling disparities between the white majority and the black minority in the United States. The purpose of this article is to make the case for affirmative action in criminal law" (Butler, 1997: 843). And Butler further points out that whereas affirmative action in the traditional sites has to do with benefits, affirmative action in criminal law has to do with burdens (ibid.: 868). In this sense, Butler's paper may be seen as a forerunner to my paper. But our arguments are still different. Butler's argument is focused on a particular site of justice, namely the criminal law. He wants to extend affirmative action to this site. My argument is not site-focused in this sense. My concern is with extending affirmative action to burdens, and not just benefits; that is, to draw the distinction between negative and positive affirmative action. In this sense, my argument speaks to any site of justice in which benefits and burdens are distributed. This is also why I discuss burdens outside of criminal law, such as firings. (Note, also, that not only burdens, but also benefits, are distributed in criminal law, such as early parole). In this sense, my argument is broader than Butler's argument. Moreover, Butler is mostly speaking to those who already accept affirmative action in civil law (ibid.: 845), whereas my arguments speak directly to opponents of affirmative action by showing that their arguments might not establish that affirmative action as such is objectionable. Additionally, whereas Butler's argument is situated mostly within a legal framework, my argument is situated mostly within a philosophical framework. But despite these differences, it is important to acknowledge Butler's work since my paper is very much within the spirit of his work.

[9] Some items might be neither good nor bad, but simply neutral. If such items exist, I set them aside and focus on goods and bads.

[10] The following illustration with examples may be helpful:

	Positive AA	Negative AA

Entry-focused AA	Hiring	Criminal sentencing
Exit-focused AA	Early release from prison	Firing

[11] For discussion, see, e.g., Lippert-Rasmussen (2020, ch. 2) and Sher (2005).
[12] I am aware that the *non-identity problem* challenges the claim that the potential recipients of affirmative action are worse off because of past injustice because, had it not been for the past injustice, different people would have existed instead (Lippert-Rasmussen, 2020: 32-36; see also Parfit, 1984). But if this is a challenge, it is as much a challenge to positive affirmative action as negative affirmative action.
[13] A closely related argument in favor of affirmative action is the *mitigating discrimination argument* (Anderson, 2010: 136; Lippert-Rasmussen, 2020: ch. 3; Scanlon, 2018: 48; *Sheet Metal Workers v. EEOC*). According to this argument, affirmative action is needed to eliminate, or at least mitigate, present discrimination and its negative effects on members of disadvantaged groups. Because these two arguments are closely related, my arguments in relation to the equality of opportunity argument apply, *mutatis mutandis*, to the mitigating discrimination argument as well. For this reason, I will not further discuss this argument.
[14] Note also that if the diversity argument proposes that diversity is valuable *in itself*, then diversity in the prison population should also be valuable in itself.
[15] Thus, this might speak to Anderson's (2010) *integrationist* argument in favor of affirmative action. According to this argument, affirmative action is justified when and because it reduces stigmatization and segregation in society, i.e., when it promotes integration (while Anderson focuses on African Americans, her argument applies to any group which is segregated and stigmatized, cf. Lippert-Rasmussen, 2020: 144). As far as I can see, the integrationist argument should speak at least as much in favor of negative as positive affirmative action. After all, those who are stigmatized and segregated can be expected both to receive fewer goods and more bads, although which members within the group who are likely to receive which will differ. Indeed, negative and positive affirmative action seem to address two equally important parts of disadvantaged groups in relation to integration: the former is likely to primarily address the worse off within the disadvantaged group, whereas the latter is likely to primarily address the better off within the disadvantaged group. I say more about this in the next section when I discuss the mismatch objection to affirmative action. Perhaps one might even say that, from the point of view of the integrationist argument, negative affirmative action is more apt than positive affirmative action, assuming that the worst off members within the group are likely to be more disintegrated than the better off members. But I will not pursue this argument further here.
[16] I thank an anonymous reviewer for raising this objection.
[17] Might there be a prominent argument in the literature that speaks more in favor of positive than negative affirmative action? The most obvious candidate seems to be the *role model argument*, according to which we must pursue affirmative action to secure group-identical role models for members of disadvantaged social groups (Lippert-Rasmussen, 2020: 104-105; see also Allen, 2002; Appiah and Gutmann, 1996; Sher, 2002). We might think that this speaks more in favor of positive than negative affirmative action. After all, why would members of disadvantaged groups need the presence of role models in, say, prisons? But this is to conceive of negative affirmative action too narrowly. First, parents are among the most important role models for their children. Pursuing negative affirmative action in relation to sentencing gives parents from disadvantaged groups better opportunities to serve as role models for their children (because they will spend less time in prison). Second, we could also, as we have seen, pursue negative affirmative action in relation to firings. Not firing someone from a disadvantaged group (negative affirmative action) might be as important in securing role models as hiring someone from a disadvantaged group (positive affirmative action). Third, positive affirmative action might lead to more role model

benefits to those who are better off within the disadvantaged groups (because they will be more likely to apply to a prestigious university or a job to begin with), whereas negative affirmative action might lead to more role model benefits (or fewer role model bads) to those who are worst off within the disadvantaged group (think, again, of reduced prison sentences). In sum, it is by no means clear that the role model argument speaks more in favor of positive than negative affirmative action.

[18] Although, as Lippert-Rasmussen (2020: 200) reminds us, "there is no reason in principle why one could not also have affirmative action programs for majority people who have been subjected to injustices comparable to those that standard affirmative action programs address."

[19] One might think that the important difference between negative and positive affirmative action is actually that the latter has to do with limited spots for which the candidates compete (e.g., job hirings and a spot at a university), whereas this is not the case with the former, and that this is actually what is doing the work in the argument. However, the distinction between limited/non-limited spots cuts across the negative/positive affirmative action distinction. Some forms of negative affirmative action have to do with limited spots, e.g., suppose a company, for budget reasons, has to fire one employee. If they pursue negative affirmative action in relation to the minority candidate, that increases the chances that the majority candidate gets fired. Some forms of positive affirmative action do not have to do with limited spots, e.g., suppose we pursue affirmative action when it comes to granting parole, that is, we make it easier (e.g., by requiring less in terms of good behavior) for the minority person to get early parole. But it still seems that there is a difference when it comes to the degree of mismatch: when it comes to goods, whether limited or not, the better off within the disadvantaged group will often be more likely to receive them, for one because they have more resources to make use of the opportunity (e.g., to get hired (limited) or to get an early parole (non-limited)), whereas, when it comes to bads, whether limited or not, the worst off within the disadvantaged group will often be more likely to receive them in the first place (e.g., to get fired (limited) or to get a sentence (non-limited)). But with that said, I think it is true that the limited/non-limited distinction can do some work in the affirmative action debate, e.g., merit concerns, standardly raised against affirmative action (see Section 4.F), are less pressing when it comes to non-limited goods. I explore this issue in Bengtson (forthcoming).

[20] Justice Clarence Thomas says in his (2007) autobiography that, because of affirmative action, his Yale law degree "bore the taint of racial preference" (see also his argument in *Adarand Constructors, Inc. v. Peña*).

[21] For more general discussion of merit and meritocracy, see, e.g., Cavanagh (2002); Daniels (1978); Mason (2006); Mulligan (2018); Segall (2012).

[22] For criticism, see Cavanagh (2002: 33-82); Lippert-Rasmussen (2020: ch. 12); Meshelski (2016). For a recent study that affirmative action in politics might increase the qualifications of parliamentarians, see Aldrich and Daniel (2024).

[23] I know that matters are complicated here, i.e., in terms of how to understand (un)fairness (but space unfortunately does not allow me to dive too deep here). What I have in mind is what Temkin (2011: 62) calls *equality as comparative fairness*. He describes it as follows: "If I give one piece of candy to Andrea, and two to Rebecca, Andrea will immediately assert 'unfair!' This natural reaction suggests an intimate connection between equality and fairness. I believe that there is one central conception of equality—I do not claim it is the only one—that focuses on how people fare *relative to each other*, where the concern for equality is not separable from our concern for a certain aspect of fairness; they are part and parcel of a single concern. On this conception we say that certain inequalities are bad, or objectionable, when, and because, they are *comparatively unfair*; but by the same token, we say that there is a certain kind of comparative unfairness in certain kinds of undeserved inequalities." In any case, even if there is a way to understand unfairness

such that negative affirmative action is unfair, this objection does not threaten negative affirmative action *as such*, as I explain below.

[24] I thank an anonymous reviewer for raising this and for useful discussion.

[25] This is also the case in other countries, see, e.g., *Bugmy v. The Queen* (2013) in Australia, and *R. v. Gladue* (1999) in Canada. For discussion, see, e.g., Atiq and Miller (2018) and Tonry (2020: ch. 4).

[26] I thank an anonymous reviewer for making this point.

Acknowledgements: For comments, I am grateful to Lauritz Munch Aastrup, Hugo Cossette-Lefebvre, Alex Madva, Søren Flinch Midtgaard, Viki Møller Lyngby Pedersen, an audience at MeST, University of Copenhagen (a particular thanks to Ji Young Lee and Ezio Di Nucci for inviting me), and several anonymous reviewers. For funding, I am grateful to the Danish National Research Foundation (DNRF144).

References

Adams M (2021) Nonideal Justice, Fairness, and Affirmative Action. *Journal of Ethics and Social Philosophy* 20(3): 310–341.

Adarand Constructors, Inc v. Peña 515 U.S. 200 (1995).

Aldrich AS, Daniel WT (2024) Gender Quota Adoption and the Qualifications of Parliamentarians. *The Journal of Politics* first view: 1-6.

Alexander M (2010) *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. New York: New Press.

Allen A (2002) The Role Model Argument and Faculty Diversity. In: Cahn S (ed) *The Affirmative Action Debate*. New York: Routledge, 153–167.

Anderson E (2010) *The Imperative of Integration*. Princeton, NJ: Princeton University Press.

Appiah KA (2011) 'Group Rights' and Racial Affirmative Action. *The Journal of Ethics* 15(3): 265–280.

Appiah KA, Gutmann A (1996) *Color Conscious: The Political Morality of Race*. Princeton, NJ: Princeton University Press.

Atiq EH, Miller EL (2018) The Limits of Law in the Evaluation of Mitigating Evidence. *American Journal of Criminal Law* 45: 167-202.

Beauchamp TL (2002) In Favor of Affirmative Action. In: Cahn S (ed) *The Affirmative Action Debate*. New York: Routledge, 209–223.

Bengtson A (forthcoming) Affirmative Action without Competition. *American Journal of Political Science*.

Bowen WG, Bok D (2002) The meaning of "merit". In: Cahn S (ed) *The Affirmative Action Debate*. New York: Routledge, 176-182.

Bugmy v. The Queen S99 NSW (2013).

Butler P (1997) Affirmative Action and the Criminal Law. *University of Colorado Law Review* 68(4): 841-890.

Butler P (2021) Foreword to the Republication of Affirmative Action and the Criminal Law. *University of Colorado Law Review* 92(5): 1443-1448.

Cahn SM (2002) *The Affirmative Action Debate*. New York: Routledge.

Cavanagh M (2002) *Against Equality of Opportunity*. Oxford: Clarendon Press.

Cholbi M, Madva A (2018) Black Lives Matter and the Call for Death Penalty Abolition. *Ethics* 128(3): 517-544.

Cohen C, Sterba J (2003) *Affirmative Action and Racial Preference: A Debate*. New York: Oxford University Press.

Daniels N (1978) Merit and Meritocracy. *Philosophy & Public Affairs* 7(3): 206-223.

Deshpande A (2019) Double Jeopardy? Stigma of Identity and Affirmative Action. *The Review of Black Political Economy* 46(1): 38-64.

Federal Bureau of Prisons (2024) Statistics: Inmate Race. https://www.bop.gov/about/statistics/statistics_inmate_race.jsp

Fischer v. University of Texas 570 U.S. 297 (2013).

Fricker M (2007) *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford: Oxford University Press.

Fullinwider RK (2014) "Affirmative Action." *Stanford Encyclopedia of Philosophy*. http://plato.stanford.edu.proxy-ub.rug.nl/entries/affirmative-action/

Glaser J (2014) *Suspect Race: Causes and Consequences of Racial Profiling*. Oxford: Oxford University Press.

Goldberg SC (2022) What is a speaker owed? *Philosophy & Public Affairs* 50(3): 375-407.

Grutter v. Bolinger 539 U.S. 306 (2003).

Gulzar S, Haas N, Pasquale B (2020) Does Political Affirmative Action Work, and for Whom? Theory and Evidence on India's Scheduled Areas. *American Political Science Review* 114(4): 1230–1246.

Harris LC, Narayan U (2014) Affirmative Action as Equalizing Opportunity: Challenging the Myth of Preferential Treatment. In: LaFollette H (ed) *Ethics in Practice: An Anthology*. Hoboken, NJ: Wiley-Blackwell, 422–433.

Hinton E (2016) *From the War on Poverty to the War on Crime: The Making of Mass Incarceration in America*. Cambridge, MA: Harvard University Press.

Hunt JS (2015) Race, Ethnicity, and Culture in Jury Decision Making. *Annual Review of Law and Social Science* 11: 269-288.

Khaitan T (2015) *A Theory of Discrimination Law*. Oxford: Oxford University Press.

Kolodny N (2014) Rule Over None II: Social Equality and the Justification of Democracy. *Philosophy & Public Affairs* 42(4): 287-336.

Lawrence C, Matsuda MJ (1997) *We Won't Go Back*. Boston: Houghton Mifflin.

Levinson JD, Smith RJ, Young DM (2014) Devaluing Death: An Empirical Study of Implicit Racial Bias on Jury-eligible Citizens in Six Death Penalty States. *New York University Law Review* 89: 513-581.

Lippert-Rasmussen K (2020) *Making Sense of Affirmative Action*. New York: Oxford University Press.

Loury GC (2003) *The Anatomy of Racial Inequality*. Cambridge, MA: Harvard University Press.

Lynch M, Haney C (2011) Looking Across the Empathic Divide: Racialized Decision Making on the Capital Jury. *Michigan State Law Review*: 573-607.

Mason A (2006) *Levelling the Playing Field: The Idea of Equal Opportunity and Its Place in Egalitarian Thought*. Oxford: Oxford University Press.

Meshelski K (2016) Procedural Justice and Affirmative Action. *Ethical Theory and Moral Practice* 19: 425–443.

Mulligan T (2018) *Justice and the Meritocratic State*. New York: Routledge.

Parfit D (1984) *Reasons and Persons*. Oxford: Oxford University Press.

Pojman LP (2014) Against Affirmative Action. In: LaFollette H (ed) *Ethics in Practice: An Anthology*. Hoboken, NJ: Wiley-Blackwell, 433–442.

Rawls J (2001) *Justice as Fairness: A Restatement*. Cambridge, MA: Harvard University Press.

Regents of the University of California v. Bakke 438 U.S. 265 (1978).

R. v. Gladue 1 SCR 688 (1999).

Richmond v. J. A. Croson Co 488 U.S. 469 (1989).

Scanlon TM (2018) *Why Does Inequality Matter?* Oxford: Oxford University Press.

Schuette v. Coalition to Defend Affirmative Action 572 U.S. 291 (2014).

Segall S (2012) Should the Best Qualified Be Appointed? *Journal of Moral Philosophy* 9: 31-54.

Segall S (2013) *Equality and Opportunity*. Oxford: Oxford University Press.

Sheet Metal Workers v. EEOC 478 U.S. 421 (1986).

Sher G (2002) Justifying Reverse Discrimination in Employment. In: Cahn S (ed) *The Affirmative Action Debate*. New York: Routledge, 58–67.

Sher G (2005) Transgenerational Compensation. *Philosophy & Public Affairs* 33(2): 181-200.

Students for Fair Admissions, Inc. v. President and Fellows of Harvard College 600 U.S. 181 (2023).

Tadros V (2020) Distributing Responsibility. *Philosophy & Public Affairs* 48(3): 223-261.

Táíwò O (2022) *Elite Capture: How the Powerful Took Over Identity Politics (And Everything Else)*. La Vergne: Pluto Press.

Taylor R (2009) Rawlsian Affirmative Action. *Ethics* 119: 476-506.

Temkin L (2011) Justice, Equality, Fairness, Desert, Rights, Free Will, Responsibility, and Luck. In: Knight C, Stemplowska Z (ed) *Responsibility and Distributive Justice*, 51-76.

Thernstrom A, Thernstrom S (2002) Does Your 'Merit' Depend upon Your Race? A Rejoinder to Bowen and Bok. In: Cahn S (ed) *The Affirmative Action Debate*. New York: Routledge, 183-189.

Thomas C (2007) *My Grandfather's Son: A Memoir*. New York: Harper.

Tonry M (2020) *Doing Justice, Preventing Crime*. Oxford: Oxford University Press.

United States Census Bureau (2023) QuickFacts. https://www.census.gov/quickfacts/fact/table/US/IPE120222

United Steel Workers of America v. Weber 443 U.S. 193 (1979).