Rawlsian Affirmative Action*

Robert S. Taylor

INTRODUCTION

What implications does Rawls's justice as fairness have for affirmative-action policies? Surprisingly, Rawls never addressed this issue in his writings—apart from one indirect reference, which I will review below.[1] Samuel Freeman, however, reports that he spoke of it in his lectures and held the following views:

> So-called "affirmative action," or giving preferential treatment for socially disadvantaged minorities, is not part of FEO [Fair Equality of Opportunity] for Rawls, and is perhaps incompatible with it. This does not mean that Rawls never regarded preferential treatment in hiring and education as appropriate. In lectures he indicated that it may be a proper corrective for remedying the present effects of past discrimination. But this assumes it is temporary. Under the ideal conditions of a "well-ordered society," Rawls did not regard preferential treatment as compatible with fair equality of opportunity. It does not fit with the emphasis on individuals and individual rights, rather than groups or group rights, that is central to liberalism.[2]

Thomas Nagel largely concurs with Freeman's "reading" of Rawls, especially with its focus on FEO and the distinction between ideal and nonideal conditions.[3] These observations raise two questions, however. First, was Rawls correct in believing this, that is, are these conclusions

* I thank Amber Boydstun, Yvonne Chiu, Brad Jones, and Yuch Kono for helpful discussions and assistance. I am also grateful to the referees and editors of *Ethics* (especially David Miller) for their invaluable comments, suggestions, and support.

1. John Rawls, *Justice as Fairness: A Restatement*, ed. Erin Kelly (Cambridge, MA: Harvard University Press, 2001), 66.

2. Samuel Freeman, *Rawls* (London: Routledge, 2007), 90–91.

3. Thomas Nagel, "John Rawls and Affirmative Action," *Journal of Blacks in Higher Education* 39 (2003): 82–84, and "Rawls and Liberalism," in *The Cambridge Companion to Rawls*, ed. Samuel Freeman (Cambridge: Cambridge University Press, 2003), 62–85, here 84 n. 3.

Ethics 119 (April 2009): 476–506
 0014-1704/2009/11903-0003$10.00

really implied by justice as fairness? If so, what does this tell us about the appropriate forms of, as well as justifications for, affirmative-action programs? I will argue in this essay that these conclusions are indeed genuine implications of justice as fairness and that they offer us guidance in regard to the defense and design of affirmative-action policies. Additionally, insofar as other liberals share Rawls's commitments to individualism, proceduralism, and autonomy, the conclusions reached below about legitimate—and illegitimate—forms of affirmative action should have resonance outside his particular theoretical framework. In fact, by examining the controversial implications of these commitments in the case of affirmative action, we shall gain unanticipated insights into the structure of Rawls's political theory and liberalism more generally.

Perhaps even more remarkable than Rawls's silence on this issue is how little his chief interpreters have discussed it. Freeman's and Nagel's comments, for example, are rather brief, the former consisting of less than a paragraph (reproduced in full above), the latter a mere page and a half of text, plus a footnote in another essay. Granted, there is virtually no written material from which to work, but this absence has not stopped a handful of others, including Edwin Goff and Elisabeth Rapaport, who have published the only extended treatments.[4] These contributions are problematic, however, in light of Rawls's reported lecture comments. Rapaport, for example, gives us a four-page sketch of a Rawlsian defense of affirmative action that makes no distinction between ideal and nonideal conditions, and Goff builds a Rawlsian partial-compliance argument that bears no relation to, and makes no mention of, Rawls's own partial-compliance applications in *Theory*, including those of section 35 (on tolerating the intolerant) and sections 55–59 (on civil disobedience and conscientious refusal). Consequently, these early contributions seem insufficiently Rawlsian, as they lack the requisite tight connection to Rawls's own nonideal theory. In contrast, this essay will assiduously try to establish just such a connection, relying on his own writings (and those of his former students, like Christine Korsgaard) to discover what implications his nonideal theory has for affirmative action. In the course of doing so, we shall learn a great deal about the proper relationship between ideal and nonideal theory, a topic that has been explored in the past decade by a number of prominent philosophers, including G. A. Cohen, Liam Murphy, and George Sher;

4. See Edwin L. Goff, "Affirmative Action, John Rawls, and a Partial Compliance Theory of Justice," *Cultural Hermeneutics* 4 (1976): 43–59; and Elizabeth Rapaport, "Ethics and Social Policy," *Canadian Journal of Philosophy* 11 (1981): 285–308.

specifically, my essay will argue that any deontological ideal theory must not just guide but also constrain its complementary nonideal theory, lest they suffer from a fatal tension.[5]

Before continuing, I should offer a definition of the essay's central concept: I understand "affirmative action" to be a class of public policies focused on achieving equality of opportunity, especially in the realms of tertiary education and employment, for certain historically oppressed groups (e.g., African Americans and women). My approach will therefore be "forward-looking" (motivated by the egalitarian political ideal of a color-blind and gender-blind society) rather than "backward-looking" (focused on reparations for past injuries) or diversity oriented.[6] Moreover, within this class of public policies, some will be stronger or more aggressive than others. Nagel provides a useful taxonomy, which I will adopt but modify for my own purposes, consisting of five affirmative-action categories ranging from weakest to strongest:

Category 1. Formal Equality of Opportunity: "careers open to talents," requiring inter alia the elimination of legal barriers to persons of color, women, and so forth as well as the punishment of private discrimination against them.

Category 2. Aggressive Formal Equality of Opportunity: self-conscious impartiality achieved through sensitivity training, external monitoring and enforcement (e.g., by the Equal Employment Opportunity Commission), outreach efforts, and so forth as a possible supplement to category 1.

Category 3. Compensating Support: "special training programs, or financial backing, or day-care centers, or apprenticeships, or tutoring," all designed to compensate for color- or gender-based disadvantages in preparation, social support, and so forth and by doing so to help recipients compete more effectively for university admission or employment.

Category 4. Soft Quotas: "compensatory discrimination in the selection process," such as adding "bonus points" to the selection indices of persons of color or women in the college-admissions or hiring processes, but without the use of explicit quotas.

Category 5. Hard Quotas: "admission [or hiring] quotas," perhaps "pro-

5. G. A. Cohen, *If You're an Egalitarian, How Come You're So Rich?* (Cambridge, MA: Harvard University Press, 2000); Liam Murphy, *Moral Demands in Nonideal Theory* (Oxford: Oxford University Press, 2000); George Sher, *Approximate Justice: Studies in Non-Ideal Theory* (Lanham, MD: Rowman & Littlefield, 1997).

6. For critiques of these alternative approaches to affirmative action, see George Sher's articles "Groups and Justice," *Ethics* 87 (1977): 174–81, and "Diversity," *Philosophy & Public Affairs* 28 (1999): 85–104, respectively.

> portional to the representation of a given [historically oppressed] group in the population."[7]

As I will later discuss, categories 1 and 2 involve (more or less aggressive) interventions into the admissions and hiring processes to eliminate bias, that is, promote color-blind, gender-blind selection; category 3 involves compensating interventions in preparation, financial and social support, and so forth prior to candidates entering selection processes; category 4 involves interventions into selection processes in order to predispose them in favor of persons of color, women, and so forth; finally, category 5 involves interventions into the outcomes of selection processes, mandating particular racial and gender mixes, for example.

The article will proceed in two stages. First, I will show that under ideal conditions justice as fairness demands category 1 interventions and (under special circumstances) even category 2 interventions but prohibits categories 3, 4, and 5. Second, I will demonstrate that under nonideal conditions justice as fairness also allows category 3 interventions but nearly always continues to disallow those under categories 4 and 5, which are tough to square with justice as fairness under any conditions. In the conclusion, I discuss the implications of these surprising results for liberal theory more widely and for public law and policy in this area, which include both Supreme Court decisions in the *Bakke*, *Gratz*, and *Grutter* cases and ballot initiatives in California, Washington, and Michigan banning affirmative action in state hiring, contracting, university admissions, and so on.

IDEAL THEORY

Briefly, the special conception of justice as fairness, with its lexically ordered principles of justice—the equal-liberty principle (EL), FEO, and the difference principle (DP)—holds only under ideal conditions, namely, "strict compliance" (i.e., no ongoing injustices) as well as the absence of particular "historical contingencies" (e.g., an authoritarian political culture or severe economic underdevelopment).[8] Postponing for a moment further discussion of the precise meaning of these conditions, let us assume that they hold and that we are consequently in the realm of ideal theory. In this case, FEO applies with full force and

7. Thomas Nagel, "Equal Treatment and Compensatory Discrimination," *Philosophy & Public Affairs* 2 (1973): 348–63, here 349–51, 356; also see Alan H. Goldman, "Affirmative Action," *Philosophy & Public Affairs* 5 (1976): 178–95, here 181, 185 (examples of category 2 and 3 interventions).

8. John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1999), 54–55, 132, 214–20, 474–76, *Political Liberalism* (New York: Columbia University Press, 1993), 7, 297, and *The Law of Peoples* (Cambridge, MA: Harvard University Press, 1999), pt. 3, especially sec. 15 ("Burdened Societies").

may not be compromised for the sake of either the DP or any considerations beyond justice as fairness, including concerns for social welfare or military efficiency.[9]

FEO has two discrete components. First, FEO demands formal equality of opportunity or "careers open to talents," that is, it forbids arbitrary discrimination (on grounds of race, gender, etc.) by either the state or private agents and condemns all monopolistic privileges (including barriers to entry in labor markets, like closed-shop unionism and exclusionary occupational licensing).[10] Second, FEO requires substantive equality of opportunity: all citizens must be guaranteed a fair chance to compete for offices and positions in the basic structure of society, regardless of social circumstances (e.g., class status or family background); as Rawls says, "those with similar skills and talents should have similar life chances." In order to achieve fairness, the state must prevent "excessive accumulations of property and wealth" and sustain "equal opportunities of education for all." More specifically, the state might impose inheritance and gift taxes, restrict the right of bequest, and subsidize education (whether directly through public schools—including so-called charter schools—or indirectly through vouchers, tuition tax credits, loans, etc.).[11]

What kinds of affirmative action does FEO permit under ideal conditions? Because FEO includes formal equality of opportunity as one of its components, it not only permits but requires category 1 interventions. Punishing private discrimination might seem unnecessary under ideal conditions, as such discrimination appears prima facie inconsistent with "strict compliance," but Rawls assumes that low-level criminality is an ineliminable characteristic of human societies, so antidiscrimination laws would presumably be violated occasionally and would therefore need to be enforced.[12] Category 2 interventions, on the other hand, would seem permissible only under nonideal conditions, as their ag-

9. For example, insofar as the United States' "don't ask, don't tell" policy for homosexuals in military service is grounded on a concern for military efficiency—unit esprit de corps might be hampered by a revelation of unconventional sexual orientation—it would be inconsistent with FEO, at least under ideal conditions.

10. Rawls, *Theory of Justice*, 62, 64, 243; John Rawls, "Distributive Justice," in *Collected Papers*, ed. Samuel Freeman (Cambridge, MA: Harvard University Press, 1999), 130–53, here 141, *Justice as Fairness*, 43 (where formal equality of opportunity is defined as "careers open to talents"), 67 n. 35, and *Political Liberalism*, 6.

11. Rawls, *Theory of Justice*, 63, 245, "Distributive Justice," 141, and *Justice as Fairness*, 51, 161. This list of policies is just suggestive, of course: one must go through the four-stage sequence to determine which policy mix is required in any given time and place (*Theory of Justice*, sec. 31).

12. Rawls says that "we need an account of penal sanctions, however limited, *even for ideal theory*. Given the normal conditions of human life such arrangements are necessary" (*Theory of Justice*, 212; emphasis added).

gressive "social-engineering" quality appears to presume citizen recalcitrance with respect to both attitudes and behavior: why would sensitivity training, external monitoring and enforcement, and so forth be necessary unless many citizens were still under the sway of racism and sexism and therefore prone to systematic violations of antidiscrimination laws? While category 2 interventions are easiest to justify under nonideal conditions, one can imagine at least two situations where they might be justifiable under ideal conditions, both related to stability: in the first, ideal conditions have only recently been achieved and the possibility of "backsliding" is nontrivial, given recent historical experiences of racism and sexism (though interventions in such a case would presumably be impermanent and prophylactic); in the second, ideal conditions have been attained but continuing large-scale immigration combined with ethnic/racial "clumping" in particular neighborhoods and/or industries threaten to reignite stereotyping and discrimination.[13] In the absence of such special circumstances, though, category 2 interventions would admittedly be hard to justify under ideal conditions.

FEO would appear to rule out the other, stronger categories of affirmative action, at least under ideal conditions, because they straightforwardly violate formal equality of opportunity and the associated ideals of color-blindness and gender-blindness. Even category 3 interventions allot "compensating support" on the basis of race and gender, preventing those who lack the requisite "markers" from even competing for special training programs, financial support, and so forth. Categories 4 and 5 offend even more blatantly against these ideals, as they distribute selection-index points (category 4) or even actual positions (category 5) in a racially and/or gender-exclusive manner and thereby balkanize academic and occupational space. This is presumably why Rawls believed that "under . . . ideal conditions . . . preferential treatment [is not] compatible with fair equality of opportunity. It does not fit with the emphasis on individuals and individual rights, rather than groups or group rights, that is central to liberalism."[14]

13. Consider, for example, the remarkable surge in Indian ownership of economy hotels in the United States over the last three decades. Such "clumping" (which acts as a catalyst for the stereotyping and discrimination that often follows from it) occurs for innocent reasons: previous immigrants accumulate industry-specific expertise and capital that they can pass on to newcomers, especially family members, easing their transition into the economic life of their host country. It is unclear to me, at least, that this phenomenon is inevitably the consequence of nonideal conditions either domestically or internationally. One might also maintain that the instability evident in my two examples above means that ideal conditions have not been achieved yet. I do not think that the bar for ideal conditions should be set quite that high, but if one were inclined to do so, then category 2 interventions would likely be limited to nonideal conditions.

14. Freeman, *Rawls*, 91.

One might reasonably object here that FEO has both formal and substantive components and that insofar as race and gender are treated as social constructs and consequently regulable by FEO—as they typically are by Rawls's interpreters—rectificatory action to effect substantive EO and "level the playing field" with respect to race and gender is as easy to justify as similar action taken with respect to family income and social class.[15] In other words, just as FEO is designed to compensate for the social disadvantages of family and class, so it should compensate for those of race and gender, even though this demands race and gender consciousness in apparent violation of formal EO. As I shall show in the next section, such a tension between formal and substantive EO indeed exists under nonideal conditions, where legacies of racism and sexism continue in the form of systematic discrimination sustained by hateful doctrines and stereotypes, all of which act to further disadvantage historically burdened groups and make a mockery of "strict compliance." Under ideal conditions, however, such legacies have been overcome, and no disadvantages in the domains of race and gender remain to be corrected by substantive EO; discrimination might still occur, as I noted above, but it will be unsystematic and idiosyncratic, like discrimination against the red-haired or gray-eyed, and can be remedied by the enforcement of antidiscrimination laws alone. Implicit then in the inclusion of race and gender in formal EO as suspect classifications is the assumption that a race- and gender-blind world is possible—a world where the interventions of categories 3–5 would be superfluous—whereas the inequalities of family and class cannot be eliminated in this way but only counterbalanced by substantive EO, even in ideal theory.[16]

Under ideal conditions, the relationship between formal and substantive EO is best seen as one of lexical priority, like that between the wider principles of justice. Rawls never explicitly says this, to be clear, but it is very strongly implied by two structural features of his theory. First, when initially interpreting his second principle of justice, Rawls recognizes two natural readings of each part of that principle, which

15. Both Freeman (*Rawls*, 90–91) and Nagel ("John Rawls and Affirmative Action," 84) treat FEO as the relevant principle in judging affirmative action, and FEO is designed to compensate for social contingencies such as those of "social class" and "family income" (*Justice as Fairness*, 44), which implies that they consider race and gender to be social constructs, albeit ones that are grounded upon certain natural "markers." Thomas Pogge agrees that they are closer to being social contingencies than natural ones—see his *Realizing Rawls* (Ithaca, NY: Cornell University Press, 1989), 164–65.

16. The DP explicitly allows economic inequalities, perhaps even large ones, and is therefore consistent with a class system, albeit one tightly regulated by the principles of justice as fairness. Rawls also recognizes that the family is a source of continuing inequalities of opportunity, ones that FEO must try (not wholly successfully) to counterbalance, but that this is not a sufficient reason to abolish the family—see Rawls, *Theory of Justice*, 265, 448.

states that socioeconomic inequalities should be regulated so that they are both (*a*) to "everyone's advantage" and (*b*) linked to jobs "open to all."[17] Part *a* is given two readings—the principle of efficiency and the DP—the former being more permissive of inequality, the latter less permissive but nonetheless incorporating the former: as Rawls says, the DP is a principle that "singles out one of these efficient distributions as also just. [It moves] beyond mere efficiency yet in a way compatible with it."[18] In other words, the DP first identifies those distributions consistent with efficiency, then picks the one that is to the greatest advantage of the least-advantaged person; in this sense, the principle of efficiency is prior to the DP. To be sure, this priority does not necessarily hold under nonideal conditions: Rawls argues that "if the basic structure is unjust . . . changes that are not efficient" may be required, because in a nonideal world "justice is prior to efficiency"; under ideal conditions, however, "justice is *defined* so that it is consistent with efficiency."[19]

In analogous fashion, part *b* is also given two readings—"careers open to talents" and FEO—the former being more permissive of inequality, the latter less permissive but nonetheless incorporating the former: FEO counters social disadvantages by "*adding* to the requirement of careers open to talents the *further* condition of the principle of fair equality of opportunity. The idea here is that positions are to be *not only* open in a formal sense, but that all should have a fair chance to attain them."[20] In other words, FEO first checks to see that distributions are consistent with careers open to talents, then picks the one that counteracts social contingencies; in this way, formal EO is prior to substantive EO. As we shall see in the next section, this priority again does not necessarily hold under nonideal conditions.

This textual evidence could not establish the lexical priority of formal to substantive EO on its own, however, because the passages cited can be reasonably interpreted in other ways, and the priority identified in them is more methodological than substantive and may not be lexical. If we turn to *Justice as Fairness*, though, a second structural feature provides the needed additional evidence: Rawls treats formal (but not substantive) EO there as a "constitutional essential," that is, "those crucial

17. Ibid., 53.

18. Ibid., 61.

19. Ibid., 69 (emphasis added); cf. "Distributive Justice," 136: "Now we shall assume that this [efficiency] principle would be chosen in the original position." I am offering a strongly Paretian reading of the DP here: it is Paretian first, egalitarian second. For a discussion of different ways of reading the DP—ranging from strongly Paretian to strongly egalitarian—and the textual evidence that is available to each, see Philippe Van Parijs, "Difference Principles," in Freeman, *The Cambridge Companion to Rawls*, 200–240, especially 205–8.

20. Rawls, *Theory of Justice*, 63 (emphasis added).

matters about which, given the fact of pluralism, working political agreement is most urgent."[21] Constitutional essentials like formal EO are realized in the second, constitutional stage of Rawls's four-stage sequence, while nonessentials like substantive EO are attained in the third, legislative stage, which is constrained by the constitution chosen in the prior stage.[22] If we scrutinize Rawls's list of constitutional essentials—basic liberties, "a social minimum providing for the basic needs of all citizens," and formal EO—we will notice that the first two are grounded on principles lexically prior to the second principle, namely, the first principle and a prior basic-needs principle, respectively.[23] This grouping strongly implies that formal EO is itself lexically prior to the second principle. In fact, at times in the text Rawls seems to "promote" formal EO to the first principle of justice, which would explain its priority over the second principle.[24] I think a better way to interpret him here is to think of formal and substantive EO as being in a lexical-priority relation within the second principle of justice, a priority relation reflected in their realization at the different, ordered stages of his four-stage sequence.

Although Rawls never really offers a defense of formal EO, his inclusion of it among the constitutional essentials along with the basic liberties hints at one. Rawls defends formal equality of the basic liberties as a necessary support for self-respect: even a just society will be marked by socioeconomic inequalities, which threaten the self-respect of citizens of low socioeconomic status; to guard against this threat, society guarantees the formal equality of basic liberties, thereby asserting equality of status along the key dimension of political citizenship; failure to do this would be tantamount to treating some adult citizens as minors, marking them with an official stamp of inferiority, which would undermine their self-respect.[25] In a similar way, formal EO asserts equality of status along the crucial dimension of social citizenship, ensuring that

21. Rawls, *Justice as Fairness*, 46–47.

22. Ibid., 48; cf. *Theory of Justice*, sec. 31.

23. Rawls, *Justice as Fairness*, 44 n. 7, 46–48, and *Political Liberalism*, 7: "The first principle . . . may easily be preceded by a lexically prior principle requiring that citizens' basic needs be met. . . . Certainly any such principle must be assumed in applying the first principle."

24. For example, Rawls, *Justice as Fairness*, 47: "The first principle, as explained by its interpretation, covers the constitutional essentials," which include formal but not substantive EO. I think that "promoting" formal EO in this way would be a mistake, as it does not serve the same set of purposes that the basic liberties do in the first principle of justice, namely, to protect the development and exercise of the second moral power of rationality as well as its political preconditions. Richard Arneson seems to agree with this assessment: see his "Against Rawlsian Equality of Opportunity," *Philosophical Studies* 93 (1999): 77–112, here 102–3.

25. Rawls, *Theory of Justice*, sec. 82.

ascriptive traits such as race and gender will play no role in the assignment of offices and positions in the basic structure of society; to do otherwise would again be to mark some citizens as inferiors on the basis of these traits, which as long experience has shown is difficult if not impossible to square with the self-respect of those so marked. The elimination of all formal, public status hierarchies—whether based on race, gender, caste, or aristocratic birth—is perhaps the signal achievement of liberalism, one that promises an end to the mutual degradations of mastery and servitude. These considerations help explain why formal EO has the priority that it does, at least under ideal conditions; whether it retains such priority under nonideal conditions is the question to which we now turn.[26]

NONIDEAL THEORY

Rawlsian nonideal theory is triggered by specific conditions, namely, partial compliance (i.e., ongoing, systematic injustices carried out by private and/or public agents) and/or the presence of adverse "historical contingencies," be they economic (e.g., severe underdevelopment) or cultural (e.g., authoritarian political mores).[27] Under such nonideal conditions, the lexical priorities of EL and FEO might be temporarily suspended, in which case the general (not the special) conception of justice would apply; this conception maintains that "all social values—liberty and opportunity, income and wealth, and the social bases of self-respect—are to be distributed equally unless an unequal distribution of any, or all, of these values is to everyone's advantage," which effectively allows the social primary goods to be traded off against one another.[28] In Rawls's one very brief mention of "existing discrimination and distinctions based on gender and race," he indicates that the partial-compliance branch of nonideal theory would be the right venue for dealing with them, but he declines to do so himself, saying that his focus is instead on ideal theory, though he admits here that justice as fairness would indeed be at fault if it "lack[ed] the resources to articulate the political values essential to justify the legal and social institutions needed

26. For further discussion of the strengths—and weaknesses—of this approach to defending the priority of formal EO, see Arneson, "Against Rawlsian Equality of Opportunity," 103–8.

27. Rawls, *Theory of Justice*, 215. I set aside one important aspect of nonideal theory: dealing with "natural limitations," including the temporary and permanent immaturities of childhood and severe mental retardation, respectively. See Rawls's brief discussions of justified paternalism in ibid., 183, 218–20.

28. Ibid., 54–55.

to secure the equality of women and minorities."[29] Are the stronger varieties of affirmative action (categories 3–5) among these "legal and social institutions"?

Arguably, at least, the United States is under conditions of partial compliance with respect to race and gender. Although state-sanctioned discrimination against minorities and women is a thing of the past,[30] and private discrimination has declined (probably significantly) over time, there is still substantial and systematic private discrimination on the basis of race and gender, behavior that is motivated by beliefs—be they conscious or subconscious—in the mental, physical, and/or ethical inferiority of minorities and women. Moreover, and perhaps more importantly, the legacy of past discrimination private and public can easily be seen in the socioeconomic deprivations, festering resentments, and dysfunctional identities born of oppression, ones that keep the affected citizens from participating as equals in our society. In the remainder of this essay, I shall simply assume that existing discrimination and the legacies of past discrimination constitute a violation of strict compliance and therefore of ideal conditions.

Maintaining the internal priority of formal over substantive EO under these conditions would make a mockery of both the equal-opportunity ideal and justice as fairness more broadly, because such priority would prevent us from addressing those underlying disadvantages faced by women, blacks, and so forth in open competition for offices and positions in the basic structure. To keep such priority under nonideal conditions would be even less justifiable than a failure to counteract the disadvantages of family and class under ideal conditions: even classical-liberal supporters of what Rawls calls the "system of natural liberty" would regard the disadvantages wrought by past and present discriminatory behavior as great injustices because they are the result of violations of formal EO, a principle that (unlike substantive EO) classical liberals themselves accept.[31] One is reminded here of the leftist critique of liberalism, first offered by Karl Marx, that liberal equality is

29. Rawls, *Justice as Fairness*, 66. He does go on in sec. 50 ("The Family as a Basic Institution") of this book to discuss the injustice of the gendered division of labor within the household and to offer suggestions for reform.

30. Exceptions exist, of course, including inter alia many types of official discrimination against homosexuals (with regard to military service, marriage, and adoption) and the exclusion of women from infantry and artillery units.

31. For Rawls's account of the system of natural liberty, see *Theory of Justice*, 57, 62–63. The more extreme classical liberals, such as libertarians, may reject antidiscrimination laws as violations of the property rights of employers and landlords—e.g., Murray Rothbard, *For a New Liberty: The Libertarian Manifesto* (New York: Libertarian Review Foundation, 1985), 206–7.

merely formal, insensitive to the deeper inequalities that open competition simply reproduces in its results.[32]

We can, of course, pursue the aggressive, category 2 form of formal EO consistent with the internal FEO priority, and under nonideal conditions we would most likely be obligated to do so. However, even if such interventions were perfectly efficacious, which is unlikely, they would simply eliminate present and future (systematic) discrimination, leaving the numerous legacies of past discrimination untouched. In order to realize the equal-opportunity ideal most fully, we must pursue policies that attack such legacies root and branch and thereby strive to counterbalance and ultimately eliminate the social disadvantages of gender, race, and so on. In short, we must entertain the adoption of stronger kinds of affirmative action (categories 3–5), all of which violate the letter of formal EO but appear more consistent with the spirit, at least, of FEO.

In order to determine more precisely which categories of affirmative action are allowable in what circumstances, however, we need to know much more about Rawls's nonideal theory, in regard to both its goals and the constraints under which it operates. The sketch offered above is simply inadequate to this task and was intended merely as a placeholder, one which has served to get us to this point in the argument but no further. In explicating Rawls's nonideal theory, I shall follow the lead of Christine Korsgaard, who has provided a concise, highly insightful overview of it.[33]

To begin, the goal of Rawls's nonideal theory is to achieve ideal conditions in order that the special conception of justice with its lexical-priority relations—both within and between the principles of justice—can be fully implemented. In short, the goal of nonideal theory is to create a world in which the ideal theory can be applied.[34] Hence, any proposed relaxation of the priority relations under nonideal conditions must be both temporary and instrumentally valuable.[35] When ideal con-

32. See Karl Marx, "On the Jewish Question," in *The Marx-Engels Reader*, ed. Robert C. Tucker (New York: Norton, 1978), 26–52. Rawls replies to this kind of critique in *Political Liberalism*, 324–31.

33. Christine Korsgaard, *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press, 1996), 147–51.

34. Thus Rawls says that "the complete realization of the two principles in serial order is the long-run tendency of this ordering, at least under reasonably fortunate conditions. . . . Their full achievement is, so to speak, the inherent long-run tendency of a just system" (*Theory of Justice*, 132, 218).

35. Thus Rawls says that "it is only when social circumstances do not allow the effective establishment of these basic rights that one can concede their limitation, and even then these restrictions can be granted only to the extent that they are necessary to prepare the way for the time when they are no longer justified. The denial of the equal liberties can be defended only when it is essential to change the conditions of civilization so that

ditions are attained, the lexical-priority relations are (re)established, and any relaxations therefore eliminated as temporary expedients. Hence, the stronger varieties of affirmative action can be warranted (if at all) on a provisional basis only, as when ideal conditions are achieved, the internal FEO priority is (re)established and they are then ruled out of bounds by formal EO.[36] In order for the relaxation of a priority relation to be instrumentally valuable, it must contribute to the achievement of ideal conditions. The stronger forms of affirmative action must consequently be valuable means to construct a world where ideal theory can be applied. I shall simply assume here that this is in fact the case. Many scholars have argued, of course, that affirmative action is inefficacious, possibly even counterproductive in this regard, but this raises some exceptionally complex empirical issues, ones that have yet to be resolved after much debate.[37] I therefore set them aside, at least for the purposes of this article.

If Rawls's nonideal theory consisted of nothing but a goal (ideal conditions) and a pair of conditions following immediately from it (provisionality and instrumentality), it would appear to be consequentialist in spirit if not in letter and would therefore sit uneasily with its deontological ideal-theory counterpart: to our question of how we should achieve ideal conditions, the nonideal theory would seemingly answer, "by any means necessary."[38] As Korsgaard points out, however, Rawls's "ideal [theory] will also guide our choice among nonideal alternatives, importing criteria for choice other than effectiveness."[39] There are at least three such criteria in the nonideal theory, all of them "imported"

---

in due course these liberties can be enjoyed" (ibid., 132). Notice that Rawls appears to say here that these restrictions must be not only instrumentally valuable but also "necessary" or "essential." I will stick to the weaker reading for reasons that will become clearer as the essay progresses: briefly, a strict-necessity requirement may rule out certain categories of affirmative action (namely, categories 4 and 5) that I will later criticize on other, less empirically contentious grounds.

36. I shall assume "good faith" (i.e., no hidden agendas) on the part of affirmative-action proponents when they argue, as they commonly do, that such programs are merely temporary. For one example of a scholar who does not assume good faith, see Carl Cohen's essay, "Why Race Preference Is Wrong and Bad," in *Affirmative Action and Racial Preference: A Debate*, by Carl Cohen and James P. Sterba (New York: Oxford University Press, 2003), 147.

37. For example, see the debate between Cohen and Sterba in ibid., 109–29, 260–61, 269–72.

38. It is not consequentialist in letter for a reason that Korsgaard perspicuously identifies: "The goal set by the ideal is not just one of good consequences, but of a just state of affairs. If a consequentialist view is one that defines right action entirely in terms of good consequences (which are not themselves defined in terms of considerations of rightness or justice), then nonideal theory is not consequentialist" (*Creating the Kingdom of Ends*, 157 n. 15).

39. Ibid.

from Rawls's ideal theory, which act as constraints on the pursuit of ideal conditions. First, the nonideal theory must be consistent with his "general" conception of justice, which I described earlier.[40] This baseline conception of justice applies under all conditions, both ideal and nonideal; the serially ordered principles of justice as fairness are just a "special case of [this] more general conception," one that applies only under ideal conditions. This constraint can probably be met by the stronger varieties of affirmative action because the general conception is both robustly egalitarian and highly tolerant of trade-offs among social primary goods so long as they advance the interests of "everyone," especially the least advantaged.

Second, nonideal theory must reflect the priority relations of ideal theory in its "order of action."[41] In other words, in attempting to realize ideal conditions, the nonideal theory must first focus on those conditions that are required in order for the priority of EL to apply, then on those required for the priority of FEO (whether external or internal), and so on. In what follows, I will simply assume that the conditions for the priority of EL have already been attained, so that the priority relations of FEO are next in line to be "targeted" with social resources and political effort.

Third, and most importantly for our purposes, the nonideal theory must be consistent with the spirit of the ideal theory. In the course of extending Korsgaard's teachings in this area, Tamar Schapiro draws this distinction between "letter" and "spirit": an act of honesty is "honest in letter insofar as it is an act of intentional truth telling. It is honest in spirit insofar as it is undertaken as a way of acknowledging another's right to govern himself"; under ideal conditions, honesty will be a "composite of letter and spirit," but under nonideal conditions, we may have to be honest in spirit alone.[42] Rather than pursuing Schapiro's ethical example any further, I will turn instead to a more pertinent political example, one used by Korsgaard to explain how Rawls's ideal theory constrains the nonideal theory to be consistent with its spirit: "The

40. Rawls, *Theory of Justice*, 54–55 ("injustice, then, is simply inequalities that are not to the benefit of all"), 217–18 ("the common good I think of as certain general conditions that are . . . equally to everyone's advantage").

41. Ibid., 216: "The lexical ranking of the principles specifies which elements of the ideal are relatively more urgent, and the priority rules this ordering suggests are to be applied in nonideal cases as well." As Korsgaard writes by way of example: "If formal equality of opportunity for blacks and women is ineffective, affirmative action measures may be in order. If some people claim that this causes inefficiency at first, it is neither here nor there, since equality of opportunity has priority over efficiency" (*Creating the Kingdom of Ends*, 148).

42. Tamar Schapiro, "Kantian Rigorism and Mitigating Circumstances," *Ethics* 117 (2006): 32–57, here 46–48; also see her remark about "elaborating" on Korsgaard at 45 n. 20.

special conception may also tell us which of our nonideal options is least bad, closest to ideal conduct. For instance, civil disobedience is better than resorting to violence not just because violence is bad in itself, but because of the way in which civil disobedience expresses the democratic principles of the just society it aspires to bring about (Sec. 59 [of *Theory*])."[43] That is, certain features of civil disobedience—its nonviolence, its "fidelity to law" (e.g., willing acceptance of punishment), and especially its public, expressive nature, which addresses itself to the reason of fellow citizens and appeals to liberal-democratic principles that they share—reveal its consistency with the spirit, if not the letter, of the ideal theory.[44] In some cases, of course, the nonideal theory may require us to follow not just the spirit but even the letter of the ideal theory. As Rawls's discussion of "tolerating the intolerant" indicates, justice may insist that we extend basic liberties to citizens who would deny them to others if they could—at least so long as "the constitution itself is secure" and the intolerant can be stopped from violating the rights of their fellow citizens—even though these intolerant citizens would have "no title to complain" if their own liberties were denied; by so modeling mutually tolerant behavior, by practicing what might be called an "aspirational" toleration of the intolerant, we hope eventually to "persuade them to a belief in freedom."[45] Whether or not the nonideal theory demands that we follow the letter of the ideal theory in any given case, though, it always insists that we act in its spirit: only by doing so can we pay due respect to those fundamental values (autonomy, democracy, and freedom in the previous three examples, respectively) that animate it.

Determining the precise meaning of such a constraint will be difficult, of course, and will have to be done on a case-by-case basis, where context will help us to interpret it correctly. One thing is certain, however: if this constraint can be overridden—if it is treated as just one criterion among others, with finite weight attached to its satisfaction—then the nonideal theory may allow and even require deeply troubling policies, ones that are difficult or impossible to reconcile with powerful moral intuitions derived from the ideal theory. In other words, permitting violations of the spirit of the ideal theory in addition to its letter may again lead us to ask whether the nonideal theory can be wedded to its deontological ideal-theory counterpart without fatal tension. To give one example, suppose that the conditions for the priority of FEO

43. Korsgaard, *Creating the Kingdom of Ends*, 148.

44. Rawls, *Theory of Justice*, 319–23 (sec. 55).

45. Ibid., 190–94 (sec. 35); also see Tamar Schapiro's superb discussion of "aspirational" honesty under nonideal conditions in "Kantian Rigorism and Mitigating Circumstances," 48.

(internal and external) could be achieved most rapidly by publicly executing anyone convicted of racial or gender discrimination. Such a punishment would surely violate both the letter and the spirit of the ideal theory: as noted earlier, even the ideal theory needs "an account of penal sanctions," and any reasonable theory of punishment would have to include some principle of proportionality.[46] If this third constraint can be overridden under some circumstances, though, then we cannot categorically rule such policies out of bounds, as they may be so effective (e.g., allowing the priority of FEO to be established in a few years rather than many decades) that an override is justified on seemingly consequentialist grounds. To offer another, less radical example, suppose that the quickest way to bring about the conditions for the priority of FEO is to impose a public and nationwide policy of hiring no white males in any given organization until racial and gender parity has been achieved there, regardless of the burdens thereby imposed—a policy which is at least questionably consistent with the spirit of the ideal theory, as I shall argue below. In short, unless this third constraint has "bite" to it, we might be driven on instrumentalist grounds to endorse morally intolerable policies, a sin that we normally associate with consequentialist theories like utilitarianism.[47]

Are the stronger forms of affirmative action (namely, categories 3–5) consistent, then, with the spirit of the ideal theory? First consider category 3 interventions. The central point of FEO is to liberate citizens' natural abilities and ambitions so that they are able to compete effectively for offices and positions in the basic structure, a social space where those abilities/ambitions can best be developed and exercised and self-realization thereby attained.[48] Such liberation can only occur if the social contingencies of family, class, race, gender, and so on are effectively neutralized.[49] Category 3 interventions all serve this purpose and are

46. Rawls offers some brief thoughts on punishment in *Theory of Justice*, 211–12, 276–77, 504–5.

47. Korsgaard does not say whether she thinks the third constraint can be overridden. As I have just indicated, however, the failure to give it appropriate "bite" opens Rawls up to the charge of crypto-consequentialism. Tamar Schapiro has similar worries about Korsgaard's interpretation of both Rawlsian and Kantian nonideal theory: see her article "Compliance, Complicity, and the Nature of Nonideal Conditions," *Journal of Philosophy* 100 (2003): 329–55, here 331 n. 4.

48. Rawls states that citizens who were denied opportunities in violation of the priority of FEO would be "debarred from experiencing the realization of self which comes from a skillful and devoted exercise of social duties. They would be deprived of one of the main forms of human good" (*Theory of Justice*, 73). For a detailed explication of this claim, see my article "Self-Realization and the Priority of Fair Equality of Opportunity," *Journal of Moral Philosophy* 1 (2004): 333–47. (I provide a summary of this article in the conclusion.)

49. Some interpreters of Rawls think that FEO requires only the mitigation, not the neutralization, of social contingencies: e.g., see Freeman, *Rawls*, 98.

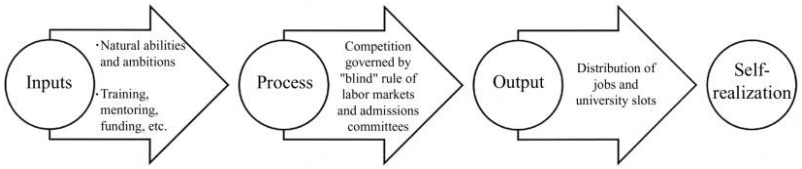

FIG. 1

thus consistent with the spirit though not the letter (due to their inconsistency with color-/gender-blindness) of FEO, including inter alia:

1. Training: to counterbalance the effects of poor schools through SAT preparation classes, co-op programs, and so forth.
2. Mentoring: to counteract the results of unsupportive or ill-informed parents, neighbors, and peers through Big Brother/Big Sister-style programs, vocational counseling, and so forth.
3. Funding: to compensate for financial disabilities through scholarships and fellowships, grants for professional wardrobes, and so forth.

Notice that none of these interventions to level the playing field of competition for employment and college admissions bends the rules of the subsequent game, so to speak. To adapt the imagery of Lyndon Johnson, category 3 interventions remove the weights from the legs of participants in a race rather than rigging its rules (e.g., giving some runners advanced starting positions or even guaranteed "placing") to produce a desired outcome—in fact, they imply nothing regarding what a desirable outcome would look like, consistent with pure procedural justice, which FEO is asked to bring about in the distribution of opportunities for offices and positions in the basic structure.[50] This distributive process is depicted in figure 1.

Category 3's neutralization of social contingencies focuses exclusively on equalizing the second kind of Input (by providing supplementary training, mentoring, funding, etc. to some citizens) so that their natural abilities and ambitions can come into undistorted competition with each other, whatever form such distortion might take (e.g., unfair advantages in the second kind of Input, rigged rules of competition, racist or sexist biases in the selection Process, etc.).

Before examining category 4 and 5 interventions, we should review what Rawls has to say about pure procedural justice, to which I just alluded above. Pure procedural justice assumes:

50. "The role of the principle of fair opportunity is to insure that the system of cooperation is one of pure procedural justice" (Rawls, *Theory of Justice*, 76).

1. "There is *no* independent criterion for the right result."
2. "There is a correct or fair procedure such that the outcome is likewise correct or fair, *whatever it is*, provided only that the procedure has been properly followed."
3. "A fair procedure translates its fairness to the outcome *only when it is actually carried out.*"[51]

Regarding the third point, part of Rawls's concern here is that if we do not require the procedure to be actually implemented, then "almost any distribution of goods is just, or fair, since it *could* have come about as a result of *fair gambles*."[52] This "looseness," as Rawls points out, is caused by the stochastic nature of the procedure itself, which does not mechanically or deterministically yield a unique outcome. As a result of this feature, we cannot even know what a just distribution looks like unless we have actually carried out a just procedure, because a stochastic process may generate different results at different times. What these criteria suggest in the FEO context is that the only way we can determine what a just distribution of offices and positions would look like is to make requisite compensating interventions in the second kind of Input (category 3), guarantee a "blind" competitive Process (category 2), and then see what follows from it—which should be an equitable distribution of offices and positions because it would arise from a fair procedure that neutralizes social contingencies.

Under nonideal conditions, we may violate the letter of FEO's internal priority, which is what category 3 interventions do, but we may not violate the spirit of FEO, which is captured by its pure proceduralism: a fair distribution here is simply whatever emerges from a fair procedure, defined as one that neutralizes social contingencies (by way of particular interventions in Inputs and Process). As I shall now argue, category 4 and 5 interventions violate this spirit. Assuming that category 2 and 3 interventions are already under way, the only reason to use the strongest forms of affirmative action is to compensate in terms of outcome (be it weakly through selection-index bonus points [category 4] or strongly through hard quotas [category 5]) for the remaining inequities in Inputs and Process, that is, ones that have not yet been eliminated by category 2 and 3 interventions. As Thomas Nagel contends, the need for the strongest forms of affirmative action "comes when it is acknowledged that some unjustly caused disadvantages, which create difficulties of access to positions formally open to all, cannot be overcome by special programs of preparatory or remedial training [i.e., Category 3]. One is then faced with the alternative of either allowing the effects of social

51. Ibid., 75 (emphasis added).
52. Ibid. (emphasis added).

injustice to confer a disadvantage in the access to desirable positions that are filled simply on the basis of qualifications relevant to performance in those positions, or else instituting a system of compensatory discrimination."[53] The fatal flaw in this approach is that, as we have seen, we lack the knowledge to use category 4 and 5 interventions to make the necessary compensations in outcome, because we simply cannot know what the counterfactual results of a "clean" competition would look like unless we run one, but we have supposed that the remaining inequities in Inputs and Process make that impossible at present. Nonetheless, we would need precisely this knowledge to carry out the requisite outcome compensations—specifically, the outcome compensations would need to equal the counterfactual outcome minus the existing one, which is presumptively unjust due to the remaining inequities in Inputs and Process. Notice that this is not an effectiveness problem but rather a conceptual, even an epistemic one: the pure procedural quality of FEO deprives us of an independent criterion for judging outcomes, forcing us to suspend judgment until a fair procedure has been achieved—but at that point, there would be no further need for category 4 and 5 interventions, as the outcome would already be just on pure procedural grounds. Thus, rejigging competitive results on justice grounds is inevitably arbitrary and inconsistent with the spirit of FEO, at least if one accepts the interpretation of FEO as an application of pure procedural justice to the distributive domain of offices and positions, as Rawls very clearly does.

I will elaborate upon this critique of category 4 and 5 interventions, which needs more development, over the course of responding to four extremely important objections to it:

> *Objection 1*: The critique is overinclusive, as it does not apply to category 4. Such interventions focus on Process, like category 2, rather than Output, like category 5. In fact, they merely offer a "head start" that counteracts residual inequities in Inputs and Process. They therefore avoid the force of the above critique.

In order to make things more concrete, I will use the "plus factor" version of category 4 to reply to this objection: applicants for university slots or jobs will have "bonus points" added to their selection indices if they belong to a socially disadvantaged race, gender, and so on and those who exceed a selection-index threshold will be admitted or hired.

53. Nagel, "Equal Treatment and Compensatory Discrimination," 350–51. Later in the article I will take issue with Nagel's implicit claim that category 2 and 3 interventions cannot eliminate most "unjustly caused disadvantages" no matter how long and how aggressively they are carried out. What does seem true is that such disadvantages cannot be eliminated solely with category 2 and 3 instruments in the short to medium term, regardless of aggressiveness.

Picking the number of bonus points to award here, however, is necessarily parasitic on concern over Output, on both conceptual and practical grounds. On a 100-point selection-index scale, how many bonus points should a woman or a black receive? 4? 7? 15? How would a particular choice here be defended against claims that it should be twice or half as big? Category 4 interventions are presumably being entertained only because category 2 and 3 interventions have not yet worked, so there remain Input and Process inequities. These inequities are highly heterogeneous, though: they include (1) residual forms of discrimination in a variety of areas (employment, education, housing, etc.), be they conscious or subconscious; (2) toxic economic and cultural legacies for historically oppressed groups, such as poverty, isolation, poor self-esteem, and dysfunctional identities; (3) a lack of information about available employment and educational opportunities; (4) an absence of positive role models; and so forth. If category 4 interventions are meant to compensate for such disparate inequities, precisely how should these inequities be "converted" into bonus points?

We need a metric, some common measure into which we can convert both for the sake of comparison. The only candidate metric that I can identify is Output, given that we must act in the spirit of pure procedural justice. The world has already converted the residual inequities in Inputs and Process into Output, in the form of the unjust employment and educational outcomes that are all around us. Bonus points granted in selection processes modify these outcomes to mimic those that would occur in a fully fair world, which according to pure procedural justice is just whatever world would result from a fair Process and equitable Inputs. In other words, selection committees compare the counterfactual results of a fair competition with that of the inequitable existing one, then choose bonus-point totals for the different disadvantaged groups to bring enough applicants over the threshold to make the disparity in outcomes vanish. Such a procedure would be subject to the same epistemic objections, however, that I raised earlier with respect to both categories 4 and 5: we cannot know what this counterfactual outcome would be without actually organizing a fair competition, but if we could run such a competition, we would do so and have no remaining need for category 4 and 5 interventions, because the result of a fair competition is itself fair on pure procedural grounds.[54]

54. To return to an earlier example: how much of a head start should we give to some runners in a race to compensate for inequitable athletic training, mentoring, facilities, officiating, and so on? In order to figure this out, we would need to know what the results of a clean (i.e., equitable) competition would look like; head starts in our dirty world could then be designed to approximate these clean results. Such knowledge is unavailable to us, however.

This conceptual connection between categories 4 and 5 has practical consequences that can be observed in the real-world operation of such systems. Bonus points (or related qualitative plus factors, like those used in "holistic" admissions procedures) should be set to compensate for many disparate racial and sexual disadvantages, but selection committees lack the information to do so effectively, as I have just suggested. They may not be flying blind, but they are flying with severe visual impairment. For example, in the aftermath of Proposition 209 and the University of California Board of Regents' decision to end affirmative action, an initial precipitous drop in the number of disadvantaged minorities admitted to Berkeley was followed by a move to an opaque holistic admissions procedure that takes into account the various "obstacles" faced by applicants over their lives. Disadvantaged-minority admissions began to rise again, a result trumpeted by its administrators. How can these administrators know, however, whether to celebrate this outcome? After all, they may have overshot the mark, yielding a result even more unjust than the one with which they started. Regardless of whether we use category 4 or category 5 interventions, we are severely hampered by a lack of information about what a just world would look like. I will return to this point in my responses to objections 3 and 4, which challenge my epistemic assumptions.

> *Objection 2*: The critique is underinclusive, as it also applies to category 3. Such interventions cannot avoid the use of Outputs as measures of whether compensation has been adequate; they need them to "meter" their effectiveness. They are therefore vulnerable to the above critique.

Insofar as category 3 instruments are designed to equalize the Inputs of training, funding, mentoring, and so on among citizens, it is not clear why we would have to rely upon Outputs to judge or "meter" their effectiveness in this respect, at least as a conceptual or epistemic matter. To give an example: if we notice that students at a predominantly black school have poorer facilities than students at a predominantly white school (e.g., fewer or lower-quality computers, fewer volumes in the school library, older and more decrepit plumbing, inferior audiovisual aids, etc.), then we can surely equalize these facilities without knowing anything about ultimate college-admissions figures. The same would apply to teacher quality, presumably, as we could use measures such as years of experience, educational attainment, standardized test scores, student evaluations, and so forth to ascertain quality and try to equalize it across schools, all without reliance upon Outputs. Granted, in some cases Outputs may be one useful measure inter alia in figuring

out whether interventions are successfully moving us toward the equalization of Inputs, especially where comparability of Inputs is an issue: for example, if we try to counteract a deficit in one Input (e.g., parenting) with a supplement in another (e.g., mentoring through Big Brothers/Big Sisters), we may find that we have to use indirect measures—such as criminal records, psychological and IQ tests, and Outputs like college admissions and job placements—to tell whether equalization of Inputs is really being achieved. Even in these cases, however, equalization could be carried out without knowing what Outputs looked like, though such ignorance might make it more expensive or difficult to realize. In short, unlike categories 4 and 5, category 3 does not require (counterfactual) Output results for its equalizing interventions, though such information might be helpful were it available.

> *Objection 3*: We already know what a fair Output would look like, because with compensating interventions in Inputs and a genuinely "blind" Process we would expect group proportionality to emerge. Thus, any existing disproportionalities are necessarily a sign of residual inequities in Inputs and Process.

Despite this objection's prima facie inconsistency with the pure procedural interpretation of FEO—it appears to assert that we have an "independent criterion for the right result"—it does not have to be read this way. To pose this objection, we need not deny that a fair Output can only be the product of a fair procedure; rather, we only have to deny the claim that we are barred from knowing what the result of such a procedure would be. So understood, the objection has a certain plausibility: if race and gender are as irrelevant to college or job performance as, say, being redheaded, then once discrimination and its legacies had been eliminated through category 2 and 3 interventions, one might expect group representation in industries and occupations to track their representation in the national population, at least approximately.[55]

Surely, however, there are disproportionate group outcomes that are (at least in part) of innocent origin, that is, not due to clear inequities in Inputs and Process. That is, the mere fact that a group is overrepre-

55. By keeping the analysis at the level of industries and occupations, I make it more likely that convergence occurs. Think of an industry or occupation as randomly drawing applicant samples from a national population. The Central Limit Theorem tells us that the larger that sample, the more likely it is that its proportion of women, blacks, etc., will correspond to that of the larger population. This being the case, the smaller samples that are drawn by individual businesses or universities are much more likely to deviate from national proportions—for entirely innocent reasons—than the larger samples drawn by whole industries or occupations. I will therefore focus on these larger entities, where the objection's claim seems more plausible.

sented in a specific occupation or industry does not necessarily imply that they had unfair advantages in Inputs or Process; similarly, underrepresentation does not always imply unfair disadvantages.[56] To give but two examples:

1. *Jewish overrepresentation among academics:* Given Jews' long history of suffering from discrimination, marginalization, expulsion, pogroms, and genocide, it would be absurd to claim that their overrepresentation within the professoriate was a consequence of unfair advantages in Inputs or Process.[57]
2. *White underrepresentation among professional athletes:* Given that whites usually have better access to facilities, funding, and so on for athletic training than blacks and Hispanics, it would again be absurd to claim that their underrepresentation in professional sports was a consequence of unfair disadvantages in Inputs or Process.[58]

Thus, even existing disproportionalities are not fully attributable to residual inequities in Inputs and Process. Moreover, if and when such inequities are eventually eliminated, there is still every reason to think that continuing cultural differences across groups will lead to disproportionalities in many, perhaps even most occupations and industries.

Of course, some of these cultural differences might not be wholly innocent, that is, unrelated to inequities in Inputs and Process. One example drawn from personal experience: Appalachians are sometimes suspicious of higher (tertiary) education and even have a phrase—"getting above your raising"—to criticize compatriots who receive "too much" education, which sets them apart from their community. Needless to say, these attitudes have the feel of adaptive preferences and are probably the result of years of economic and educational deprivation

56. See the exchange on this point between Cohen and Sterba in *Affirmative Action and Racial Preference: A Debate*, 254, 296.

57. Jews constitute 2 percent of the U.S. population but 5 percent of U.S. university faculty—see Gary A. Tobin and Aryeh K. Weinberg, *Profiles of the American University*, vol. 2, *Religious Beliefs and Behavior of College Faculty* (San Francisco: Institute for Jewish and Community Research, 2007), 20. The overrepresentation grows substantially as one ascends the academic food chain: e.g., 39 percent of Nobel Laureates in economics (twenty-four of sixty-two; 1969–2008) have been Jewish (http://www.science.co.il/Nobel.asp; http://nobelprize.org/nobel_prizes/economics/laureates).

58. Whites were 66.4 percent of the U.S. population in 2006 but only 59.8 percent of major league baseball players in 2007, 31 percent of National Football League players in 2006, and 21 percent of National Basketball Association players in 2006–7—see the Census Bureau Web page for the white-population numbers (http://quickfacts.census.gov/qfd/states/00000.html) and the recent *Racial and Gender Report Cards* put out by the Institute for Diversity and Ethics in Sports for the white-athlete numbers (http://www.tidesport.org/racialgenderreportcard.html).

and isolation combined with certain conservative religious beliefs.[59] All that I need, however, for my point to hold is for some of these Output-related cultural differences to be innocent, which would be difficult though not impossible to deny. Consequently, in order to sustain objection 3's claim, all Output-related cultural differences would have to be ascribable to Input and Process inequities, and this strikes me, at least, as highly implausible.

The only apparent way to sustain the claim that proportional group outcomes are the right ones is to do what objection 3 has tried to avoid: abandon pure procedural justice and find some "independent criterion" to justify group proportionality, which will likely be one that emphasizes the claims of groups as moral agents to particular outcomes. This would be a worrying departure, though, from the individualism and proceduralism of both Rawls and liberalism more generally, as Freeman notes in this paper's initial quotation. Perhaps more importantly for my purposes, it would fail utterly as a reading of Rawls.

> *Objection 4*: Even if we do not know precisely what a fair Output would look like, we may have a rough idea (e.g., there are almost surely too few black physicists), and such admittedly imperfect knowledge is enough to justify category 4 and 5 interventions, at least insofar as category 1–3 interventions are failing or not succeeding quickly enough.

To fix ideas, let us assume that we have normally distributed expectations regarding what a fair outcome would look like in any given occupation or industry, in terms of the percentage of jobs in it held by a specific disadvantaged minority. Let x_I be the ideal percentage of jobs held by this minority, where x_I is normally distributed with a mean μ and standard deviation σ, and let x_C be the current percentage of jobs thus held. Let us also assume that $\mu > x_C$, that is, we expect that in an ideal world the percentage of jobs held by this minority will expand from its current level. We may be wrong, however; in fact, given the strict limits of our knowledge here, we cannot rule out the possibility that the ideal will be closer to the current percentage than the predicted percentage $(= \mu = \mathrm{E}(x_I))$. The probability of this happening is just

59. For a discussion of this phenomenon in a British context, see Kristen Voigt, "Individual Choice and Unequal Participation in Higher Education," *Theory and Research in Education* 5 (2007): 87–112, here 97–99; on adaptive preferences more generally, see Jon Elster, *Sour Grapes* (Cambridge: Cambridge University Press, 1983).

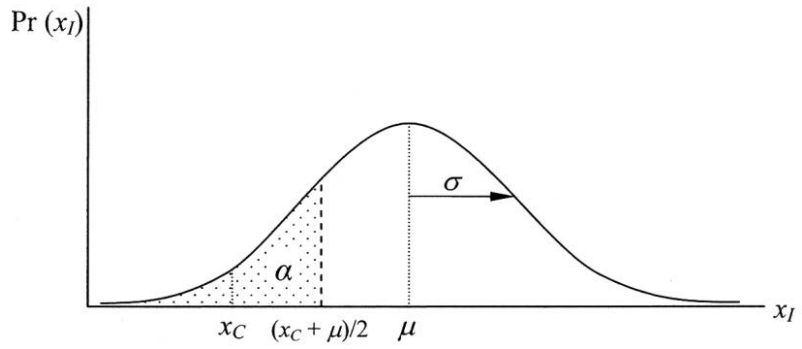

FIG. 2

the probability that the realization of x_I will be to the left of the midpoint between x_C and μ, that is, Pr $(x_I < (x_C + \mu)/2) = \alpha$ (see fig. 2).[60]

Therefore, whether we use soft or hard quotas (categories 4 and 5, respectively) to reach μ, we run the risk of further separating actual and ideal. The probability of doing so, α, grows as (1) we become less confident about where x_I lies (i.e., σ rises) and/or (2) we expect a value for x_I that is closer to the current percentage (i.e., $\mu \rightarrow x_C$), ceteris paribus.[61] So long as categories 1–3 will eventually bring about a just world, categories 4 and 5 cannot be justified even if they might speed the approach of that day: to risk increasing injustice only in order to attain ideal conditions more quickly is to fall prey to instrumentalist reasoning and, as I described it earlier in the article, to put nonideal theory into fatal tension with its deontological ideal-theory counterpart. For ideal theory to play its assigned role in guiding and constraining

60. We can think of α (the shaded area in the graph in fig. 2) as having two distinct parts. First, to the left of x_C, i.e., Pr $(x_I < x_C)$, is the probability that the ideal percentage is actually less than the current percentage. Given the strict limits on our knowledge here, we cannot definitively rule out such a possibility, as we cannot know with any confidence what decisions members of different groups would make in a wholly just world. For example, it may seem obvious that in an ideal world the number of black lawyers would rise, and that would indeed be a reasonable expectation $(\mu > x_C)$; however, once other professions became fully accessible via affirmative action, and the prestige of lawyering among blacks fell (owing to a lesser "defensive" need for it in civil rights activism and politics more generally), the number of black lawyers might fall—an unlikely occurrence, to be sure, but still possible. Second, between x_C and $(x_C + \mu)/2$, i.e., Pr $(x_C < x_I < (x_C + \mu)/2)$, is the probability that the ideal percentage is more than the current percentage but closer to it than to the expected percentage (μ). This possibility is much easier to imagine.

61. Ceteris paribus: as σ rises, more of the probability mass slides into the tails, so α increases; as $\mu \rightarrow x_C$, it does so at precisely twice the rate of the midpoint $(x_C + \mu)/2$, so again α increases.

the nonideal theory, it has to prohibit tempting trade-offs of precisely this kind.

What if we have good reason to believe, however, that categories 1–3 will not eventually bring about a just world, either because the Input and Process inequities are too deep-seated to be corrected by such interventions or the political will to implement them is (permanently) lacking? First, I think such cases will be rare: patiently and systematically applied, categories 1–3 should in time erase both discrimination and its legacies, and if the political will is truly lacking, then it seems implausible that even more radical interventions would be entertained. Moreover, even if such interventions would be entertained, we would face a policy dilemma, with the possibility of moral error on either side: we could use category 4 and 5 interventions, but this would run the risk of exacerbating injustice; we could alternatively do nothing (assuming that categories 1–3 are impotent or politically infeasible), but this would likely leave certain injustices uncorrected. In criminal justice settings, at least, liberal intuitions usually lean toward the latter: better to let the guilty go free (leave injustice uncorrected) than to punish the innocent (exacerbate injustice). I would argue that the same intuition should apply here. The high risk of having the state author injustice by means of "positive" discrimination will generally outweigh the risk of leaving unjust inequalities uncorrected. At least for nonconsequentialist liberals, sins of commission should be of much greater concern than sins of omission—especially when the sinner is the state.[62]

Having said this, I can nonetheless envision situations where categories 4 and 5 might be justifiable. Suppose that (i) we are highly confident about where x_I lies (i.e., σ is low) and (ii) we expect a value for x_I that is much higher than the current percentage (i.e., $\mu \gg x_C$); therefore, the probability α of further separating real and ideal is fairly low. Also suppose that (iii) category 1–3 interventions are now obstructed but (iv) could be made effective with a comparatively small quota that "primed the pump," so to speak; keeping this quota as small as possible, consistent with effectiveness, would further reduce α. If all these conditions were met, then the case for category 4 and 5 interventions would be relatively strong—but only as temporary enablers for category 1–3 interventions.

Are there any real-world cases that meet these conditions, at least roughly? Consider the overwhelmingly male-dominated profession of firefighting. Conditions i and ii are doubtless met here: female representation is significantly lower than it would be in an ideal world, and

62. For a discussion of the asymmetry (within a nonconsequentialist framework) between doing and allowing harm, see Samuel Scheffler, "Doing and Allowing," *Ethics* 114 (2004): 215–39.

our confidence in this assessment is relatively high. Moreover, one can make the case that conditions iii and iv hold too. Two decades of aggressive category 1–3 interventions have not dampened discrimination and harassment, causing qualified, interested women understandable reluctance to pioneer the hostile territory of "boys' clubs." Were a small quota for women imposed here—one large enough to start changing the sexist culture of firehouses and create a cohort of like-minded, mutually supportive female firefighters, yet small enough to keep α low—category 1–3 policies could become effective and self-sustaining, allowing category 4 and 5 policies to be set aside.[63] Thus as a way to get past certain "tipping points," after which categories 1–3 would be effective, soft or even hard quotas of modest size and short-term use might be justifiable. Only under these rare, restrictive conditions, however, are category 4 and 5 policies consistent with FEO's spirit.

CONCLUSION

The most important conclusion of this study—and probably the most surprising one—is that although Rawls's theory can endorse category 2 and 3 interventions under those nonideal conditions in which we find ourselves, it seldom supports category 4 and 5 interventions (soft and hard quotas, respectively), as these are ordinarily inconsistent with the spirit of FEO, whose pure proceduralism insists that we focus our political attention on establishing fair conditions of competition rather than on guaranteeing ostensibly fair outcomes. By no means is this conclusion an indication that justice as fairness is "soft" on racism, sexism, or their atrocious legacies. Quite the contrary: given the second, "order of action" constraint in his nonideal theory, political effort and social resources must be aggressively devoted to all the permissible interventions until these stains on our society are wiped clean, no matter how long it takes and even if it means that other important goals (e.g., general poverty reduction, a thriving artistic and musical culture, etc.) must be neglected for the time being.

As indicated by the earlier quotation from Nagel, some scholars believe that category 2 and 3 interventions are incapable by themselves

63. For a discussion of progress—or lack thereof—in integrating the firefighting profession, see Denise M. Hulett et al., *A National Report Card on Women in Firefighting* (Ithaca, NY: Cornell University's Institute for Women and Work, 2008; go to http://www.i−women.org/images/pdf−files/35827WSP.pdf for a copy of this report). Wayne Sumner has also argued that quotas might be used to break open "traditional bastions of male privilege" and challenge "sexist attitudes" by requiring men to attend to female qualifications, at least for those positions set aside for women; unlike me, however, he doubts that category 1–3 policies will ever be effective, sets quotas at group-proportionality levels, and operates within an explicitly consequentialist framework. See Wayne Sumner, "Positive Sexism," *Social Philosophy and Policy* 5 (1987): 204–22, especially 209–14.

of eliminating these stains, even in the long run. I do not see why this would necessarily be so, except in those rare cases discussed in my reply to objection 4. Patient and comprehensive political effort to eliminate systematic discrimination by way of sensitivity training, external monitoring and enforcement, and outreach efforts, combined with the devotion of substantial social resources over time to supplementary training, mentoring, and funding for disadvantaged groups, should eventually level the competitive playing field and allow the internal priority of formal EO to be (re)established. What justice as fairness does imply is that even if soft and hard quotas would permit a color- and gender-blind society to be founded more quickly, they are almost always ruled out as inconsistent with the spirit of the ideal theory. To pursue a just society by unjust means is a corruption of both deontological justice and those who would practice it.

How persuasive should this Rawlsian argument against the stronger forms of affirmative action be to non-Rawlsian liberals? Much hinges, of course, on their reactions to its fundamental assumptions, which I have largely taken for granted over the course of my essay. Key among the assumptions of justice as fairness are its individualism and proceduralism, which find expression in FEO: its main focus, as we have seen, is on securing fair competitive conditions for individual citizens, not on guaranteeing certain outcomes for the groups to which they belong. Some liberal multiculturalists like Will Kymlicka have called these assumptions into question, though, saying that groups matter because their cultural traditions serve as conditions for meaningful individual choice by their members—a central concern of liberalism—and that the survival, coherence, and influence of these groups should consequently be promoted, whether by temporary policies such as affirmative action or permanent ones like corporate political rights.[64] Resolving this dispute is obviously beyond the scope of my study, but as Samuel Freeman suggests in the article's opening quotation, many if not most liberals are staunchly committed to individualism and proceduralism and should therefore be sympathetic, at least initially, to my affirmative-action argument.

This argument depends upon other, more controversial assumptions, however, especially FEO's lexical priority. Even if my argument is sound and FEO is typically inconsistent with soft and hard quotas, why should it take priority over income equality and social utility, which might be advanced by such measures? For example, Nagel has argued that because jobs come attached to various "economic and social rewards," we may have good reason to override the meritocratic imperative

64. Will Kymlicka, *Liberalism, Community, and Culture* (Oxford: Oxford University Press, 1989), chap. 8 and 190–91.

of FEO in order to raise the incomes and status of members of historically oppressed groups; also, such overrides might greatly advance social welfare by, say, increasing the number of black physicians, who would be considerably more likely than nonblack ones to serve inner-city communities with desperate health-care needs.[65] Richard Arneson states the criticism more generally: "Enabling all individuals to have real opportunities for job satisfaction, educational achievement, and responsibility fulfillment is not plausibly regarded as a justice goal that trumps all other justice values and should be pursued no matter what the social cost."[66]

Rawls barely sketches a defense of FEO's lexical priority, however, and I have therefore reconstructed it in another article, which I'll summarize here.[67] FEO's priority must be grounded, as I hinted above, in our highest-order interest in self-realization through work. Various elements of Rawls's theory, including the Aristotelian principle (which motivates our perfectionist pursuit of ever deeper and wider skill sets) and the Humboldtian idea of social union (which provides the social context for such pursuits, especially in occupational settings), explain why vocational self-realization takes priority over income equality and social utility; such pursuit must be consistent, of course, with the interests protected by higher principles (namely, the basic-needs principle and first principle of justice).[68] The modest perfectionism of this reconstructed defense, though appearing to violate the priority of right, can be shown to follow from Rawls's own Kantian commitment to autonomy: just as reasonableness and rationality are facets of a Kantian conception of autonomy, so is self-realization, whose product is not a moral law or plan of life but instead a freely chosen plan of self-development and an associated ideal of personal excellence. Rawls is keenly aware that there are some who hold that "all human interests are commensurable, and

65. Nagel, "Equal Treatment and Compensatory Discrimination," 355–59, 361; also see Ronald Dworkin, *Taking Rights Seriously* (Cambridge, MA: Harvard University Press, 1977), 223–39. George Sher provides insightful criticisms of Nagel and Dworkin as "utilitarian defenders of affirmative action" in "Reverse Discrimination, the Future, and the Past," *Ethics* 90 (1979): 81–87, here 83–84. Robert L. Simon offers a much more focused critique of what he calls Dworkin's "utilitarian" argument in "Individual Rights and 'Benign' Discrimination," *Ethics* 90 (1979): 88–97, especially 91–93.

66. Arneson, "Against Rawlsian Equality of Opportunity," 99; also see Larry Alexander, "Fair Equality of Opportunity: John Rawls' (Best) Forgotten Principle," *Philosophy Research Archives* 11 (1986): 197–208, here 205–6.

67. See Rawls, *Theory of Justice*, 73, where Rawls characterizes FEO's lexical priority as underwriting "the realization of self which comes from a skillful and devoted exercise of social duties"; for my reconstruction, see Taylor, "Self-Realization and the Priority of Fair Equality of Opportunity."

68. Rawls, *Theory of Justice*, secs. 65, 79, *Justice as Fairness*, 44 n. 7, and *Political Liberalism*, 7.

that between any two there always exists some rate of exchange"; his defense of the priority of liberty challenges this view, as does my reconstructed defense of FEO's priority and whatever may follow from it, including opposition to the strongest forms of affirmative action.[69]

Although some contemporary liberals have been critical of affirmative action—ranging from David Miller, who offers qualified criticisms of certain justifications for affirmative action, to George Sher, who condemns nearly all of them—most have been supportive, even of soft and hard quotas.[70] If my arguments in this essay are sound, then I have shown that Rawls's justice as fairness, arguably the seminal theory of contemporary analytic political philosophy, usually rules out the strongest forms of affirmative action. For those liberals who share Rawls's commitments to individualism, proceduralism, and autonomy and who are skeptical of liberal-multiculturalist and consequentialist defenses of affirmative action, this essay may provide further reasons to rethink and perhaps temper their support for race- or gender-based quotas.

Finally, what implications does this article have for constitutional law and public policy? First, insofar as Supreme Court Justice Lewis Powell's opinion in *Bakke* is read to permit those category 4 interventions that avoid hard quotas but still use race as a "plus factor" in university admissions, it would be very tough to defend on my reading of justice as fairness.[71] More recent Supreme Court decisions have slightly narrowed but essentially affirmed Powell's original point: *Gratz* found unconstitutional an undergraduate-admissions policy of giving twenty bonus points to all underrepresented-minority applicants on their selection indices, even though it steered clear of hard quotas, but *Grutter* allowed

69. Rawls, *Political Liberalism*, 312. Other defenses of the lexical priority of FEO may be possible, but my reconstructed defense has important advantages over them, including fidelity to Rawls's text and reliance upon the same commitment to autonomy that underwrites the priorities of both right and liberty in his theory. (See Taylor, "Self-Realization and the Priority of Fair Equality of Opportunity," 346.) This being said, my defense is probably inconsistent with justice as fairness in its later, politically liberal incarnation—but so, I would contend, is his defense of the priority of liberty. I criticize his "political turn" for just this reason in "Rawls's Defense of the Priority of Liberty: A Kantian Reconstruction," *Philosophy & Public Affairs* 31 (2003): 246–71, here 267–71.

70. David Miller, *Principles of Social Justice* (Cambridge, MA: Harvard University Press, 1999), 172–76; Sher, *Approximate Justice*, especially chaps. 3, 4, and 6.

71. *Regents of the University of California v. Bakke*, 438 U.S. 265 (1978). My claims here might be rebutted by showing that the systematic, aggressive application of category 2 and 3 policies would have been either ineffective or politically infeasible in these three cases, even in the long run. Also recall that at the beginning of the essay I set aside defenses of affirmative action based on reparations or diversity. Diversity-based defenses are commonly invoked in higher-education contexts (especially by university defendants in court cases), but for such a defense to rebut my claims successfully here, it would need to show not only why diversity should trump the morally weighty concerns of FEO but also how the powerful objections lodged against it by George Sher in "Diversity" could be overcome.

a law-school-admissions policy that made race one selection criterion among others, because its evaluation procedure was qualitative and "individualized."[72] As I indicated earlier, category 4 interventions include any biased (i.e., color-/gender-sensitive) selection rule, no matter how qualitative or individualized, so even the *Grutter* decision would be difficult to justify using my interpretation of Rawls. Interestingly, Nagel reports that in the wake of the *Bakke* decision "Rawls expressed in conversation his view of the importance of defending the constitutionality of affirmative action"; assuming that he had soft or hard quotas in mind, his own theory would have offered him little support, at least on my reading of it here.[73]

Second, as I noted in the introduction, ballot measures in California, Washington, and Michigan have (where consistent with applicable federal laws, consent decrees, etc.) eliminated affirmative action in state hiring and contracting, college admissions, financial aid, and so forth.[74] What implications do my conclusions have for these voter initiatives? In brief, these initiatives have moved their states in the right direction but overshot the mark, because they appear to rule out not only category 4 and 5 interventions but category 3 interventions (supplementary gender- and race-based training, mentoring, and funding) as well.[75] On my reading of Rawls, justice as fairness looks critically at rigged admissions/hiring procedures and quotas but permits and even requires special scholarship funds, co-op programs, vocational-counseling offices, and so forth in order that the social contingencies of race and gender may be neutralized over time.[76] These initiatives effectively raise "Mission Accomplished" banners over their respective states, whereas the battle against racism, sexism, and their legacies is a continuing one. Abandoning legitimate legal and policy weapons prematurely is both a dereliction of duty and an admission of defeat.

72. *Gratz v. Bollinger*, 539 U.S. 244 (2003); *Grutter v. Bollinger*, 539 U.S. 306 (2003).

73. Nagel, "John Rawls and Affirmative Action," 82. Nagel also believes that "racial preferences" are "a natural consequence of [Rawls's] ideal of justice" (ibid., 84).

74. The initiatives were California's Proposition 209 (1996), Washington's I-200 (1998), and Michigan's Proposition 2 (2006).

75. For a discussion of California Proposition 209's potential implications, see the report of California's Legislative Analyst at http://vote96.sos.ca.gov/Vote96/html/BP/209analysis.htm.

76. The Department of Education's Ronald E. McNair Postbaccalaureate Achievement Program, which "prepares participants for doctoral studies through involvement in research and other scholarly activities," combines many of these features in a single program, though only one-third of slots are specifically set aside for students from "groups that are underrepresented in graduate education" (http://www.ed.gov/programs/triomcnair/index.html).