ANNALS, *AAPSS*, **523**, September 1992

Incentive Effects of Affirmative Action

By GLENN C. LOURY

ABSTRACT: This article illustrates with a formal economic model a concern often raised by critics of affirmative action—that the policy may discourage its beneficiaries from acquiring work skills. Ironically, this can happen for reasons analogous to those evoked to explain why discrimination may discourage its victims from investing in skills: when skilled workers are less likely to succeed, fewer find it worthwhile to become skilled. Similarly, when unskilled workers are more likely to succeed, fewer deem it necessary to become skilled. Discrimination can lead to the former situation; affirmative action can lead to the latter. The analysis shows how affirmative action can lead employers to patronize minority workers, that is, hold them to a different standard. This patronization can have the effect of making skill acquisition less beneficial for minority workers. The labor market conditions under which this counterproductive effect of affirmative action is most likely are identified.

Glenn C. Loury is professor of economics at Boston University. He holds a B.A. in mathematics from Northwestern University and a Ph.D. in economics from the Massachusetts Institute of Technology. He has published numerous scholarly articles in the area of applied microeconomic theory. He has also written extensively on the issues of racial inequality and social policy toward the poor. A former Guggenheim fellow, Dr. Loury has lectured on his ideas throughout Europe and North America.

NOTE: This article draws on ideas generated in collaboration with Stephen Coate of the Department of Economics, University of Pennsylvania. Our joint paper "Will Affirmative Action Eliminate Negative Stereotypes?" mimeographed (Cambridge, MA: Harvard University, May 1991) develops a more thorough analysis of the issues considered here. Professor Coate, of course, is not responsible for or implicated by any opinions expressed or errors committed in this article.

> I have a dream that my four little children will one day live in a nation where they will not be judged by the color of their skin but by the content of their character.
>
> Martin Luther King, Jr.
> Washington, D.C.
> August 1963

One often encounters the following argument against affirmative action: Ultimately, racial justice requires that people behave toward each other in their economic dealings without regard to skin color—that they obey the color-blind ideal so eloquently expounded by Martin Luther King, Jr. Affirmative action, by encouraging the use of color as a basis for allocating positions, directly violates this color-blind ideal and is thus inconsistent with the attainment of racial justice in the long run. How can we hope to achieve a discrimination-free society while engaging, through public policy, in racial discrimination?

Proponents of affirmative action dismiss this argument as naive and ahistorical. They argue as follows: To remedy the effects of past discrimination, one must direct benefits to those who, because of color, have had their opportunities reduced. Moreover, the ongoing use of color by employers in ways deleterious to minorities requires offsetting color-conscious government action to ensure equal opportunity today, regardless of the effects of past discrimination. The departure from the color-blind ideal that affirmative action represents is a necessary, temporary concession to the realities of race in our society, which will be abandoned in the future, once opportunities have become truly equal.

While this rebuttal makes several valid points,[1] I believe that the concern that affirmative action may be inconsistent with the ultimate achievement of a color-blind society deserves more serious consideration than it currently receives. The reason is that a policy of affirmative action may alter the terms on which employers and workers interact with each other so as to perpetuate, rather than eliminate, existing disparities in productivity between minority and majority populations. In particular, the use of color as a basis for distributing opportunities may have the unintended effect of dulling the incentive to acquire skills for those whom the policy is intended to benefit. The presence of such a counterproductive effect gives greater force to the seemingly naive objection to racial preferences stated previously. This is true even when affirmative action has been introduced in order to counteract the effects of ongoing discrimination by employers.

To illustrate, suppose employers believe that minority workers are, on average, less skillful than majority

1. For an extended discussion of problems with a pure color-blind approach to public policy in the face of racial inequality, see my essay "Why Should We Care about Group Inequality," *Social Philosophy and Policy*, 5(1):249-71 (Autumn 1987). I also provide there an informal discussion of some negative unintended effects of affirmative action other than the one analyzed in the present article. An important theme in that essay, having answered in the affirmative the question "Should 'color' ever be taken into account?" is that preferential treatment is often not the best method of doing so. I make the case that targeting social service benefits to disadvantaged minorities may be a superior means of taking into account the history of racial discrimination.

workers. As a result, they are less willing to assign them to high-level positions. Such discriminatory beliefs can be self-confirming because, knowing it is more difficult to get the higher positions, minority workers may rationally choose not to invest in the requisite skills, thereby confirming the employers' initial views. Now suppose an affirmative action policy is adopted, requiring employers to assign minority workers to the higher positions at the same rate as the majority. Believing they are on average less skillful, employers may calculate that to comply with this policy they must now make it easier for a minority worker to get a high-level position. But, seeing that they do not have to be as skilled as their majority counterparts in order to achieve the same success, minority workers may have less of an incentive to invest in those skills that enhance a worker's performance. If minorities choose to invest less than the majority, employers' beliefs that they are less skillful will once again be confirmed.

When discriminated against, minorities may invest less in skills than majority workers because it is more difficult for them to achieve high-level positions. When favored by affirmative action they may invest less because, given employers' response to the policy, it has become easier for them to achieve high-level positions. The point is that the incentive to acquire a skill can be lowered by either reducing the likelihood that a skilled worker will succeed or increasing the likelihood that an unskilled worker will succeed. Behavior by employers that is not color-blind can produce the first effect; behavior by the government that is color-conscious—namely, affirmative action—can produce the second effect. In both cases, because minorities have lower incentives to invest than majority workers, there is a systematic difference in the acquisition of skills by workers in the two racial groups.

Under affirmative action, employers may think they have to patronize minorities—that is, not hold them to as high a standard—in order to meet the government hiring requirements. Yet because this patronization can lower incentives for the acquisition of skills by minorities, it can perpetuate the racial skill differential that made the affirmative action policy necessary in the first place. In this sense, the government's departure from the color-blind ideal, by generating the unintended consequence of reduced incentives for the acquisition of skills by minority workers, makes the ultimate attainment of a color-blind outcome impossible. In this article, I illustrate, with the aid of formal economic reasoning, just how and why such an outcome might come about.

A FORMAL MODEL OF DISCRIMINATION

I first consider an idealized model of an employer interacting with a racially diverse population of workers. This model is not a complete or realistic description of any particular setting in which affirmative action is practiced. Rather, it is an abstraction, a thought experiment that, by focusing explicitly on a few key variables of the problem, allows one to gain insight into how these variables

interact with each other. My basic concern is with the standards employers use to decide which workers get desirable positions, the effort workers expend to acquire skills useful in those positions, and the ways in which decisions about these two variables change in the presence of racial hiring standards. These are the factors that figure prominently in the following model.[2]

(1) There is an employer and a population of workers divided into two racial groups, blacks and whites. The employer can distinguish between workers by their color and thus has the option to treat black and white workers differently. The sole action of the employer is to assign each worker to one of two tasks, called task zero and task one. Think of task one as the more demanding and more desirable of these two positions.

(2) All workers can perform satisfactorily at task zero. Workers decide, before the employer assigns them to a task and without the employer's knowledge, whether to invest in the acquisition of a skill essential for effective performance at task one. The investment is costly for a worker to make. The size of this cost varies from worker to worker, though in a manner that is statistically the same for each racial group; imagine, for example, that more able workers find it easier to acquire the skill needed for task one, and that the distribution of ability is the same within each group. The employer cannot observe a particular worker's cost. What he can observe is the group identity of each worker and the outcome of a skills test, to be described momentarily. Although the two groups are characterized by the same distribution of ability, they need not exhibit the same pattern of investment. Workers with the same investment cost but belonging to different groups might make different investment decisions, as will be explained further.

(3) Since task one is more desirable, a worker is assumed to obtain a premium whenever he gains the assignment, whether he has acquired the needed skill or not. But, because an unskilled worker performs inadequately, the employer wants a worker in task one only if he has acquired the requisite skill. Otherwise he wants that worker to go to task zero. The employer maximizes profits when skilled workers are assigned to task one, and unskilled workers to task zero. The size of his gain need not be the same in these two cases. The employer may care more about avoiding the error of putting an unskilled worker in task one than about avoiding the mistake of putting an overqualified, skilled worker in task zero, or he may have the reverse priority.

(4) The employer wants to match workers to their most productive tasks. Lacking any prior information, the employer tests a worker's qualifi-

2. The argument set out in the model is largely expressed verbally and is, therefore, less rigorous than the mathematical model that it approximates. Due to space limitations, mathematical proofs of the propositions have been omitted. They are available from the author on request, or, for a more complete treatment, see Stephen Coate and Glenn C. Loury, "Will Affirmative Action Eliminate Negative Stereotypes?" mimeographed (Cambridge, MA: Harvard University, May 1991).

cation for doing task one. That is, he gathers what information he can—from an interview, analysis of previous work history, written exam, and so on—in order to assess the worker's capabilities. I assume that this test has three possible outcomes: (1) it shows clearly that the worker can do task one; (2) it shows clearly he cannot; and (3) its outcome is ambiguous, so the employer remains uncertain of the worker's skill. The worker passes the test in case (1); in case (2) he fails it; and in case (3) his result is unclear. Only investors pass the test and only noninvestors fail it, but each has some chance of getting an unclear result. I assume the test is better at revealing noninvestors than investors in this sense: an investor has a lower chance of passing the test than does a noninvestor of failing it.[3]

(5) The behavior of workers and the employer in this model may be described as follows. Each worker, knowing his color and his investment cost, decides whether to acquire the skill needed for task one. The employer then encounters the worker, gives him the test, and, on the basis of the test result and a worker's color, assigns the worker to a task. I assume that all of these decisions are made in a way that maximizes the decision maker's anticipated net reward, given the available information. An equilibrium for this model is defined as a joint specification of behavior for the employer and the workers in each racial group that is optimal for all parties, given the behavior specified for the others. I will show in the following that, despite the absence of any racially invidious motive on the part of the employer, discrimination against blacks can arise in an equilibrium of this model.

(6) To find the equilibria, I begin by considering the employer's decision in each contingency. Clearly, he assigns anyone passing the test to task one, and anyone failing it to task zero, regardless of color. If the test result is unclear, however, he needs to estimate the likelihood that the worker has invested to determine which assignment is best. If that likelihood is great enough, he puts the worker in task one; otherwise he puts the worker in task zero. Given an unclear test result, the odds that the worker producing it has invested depend on the relative number of investors in the population from which the worker comes and on the respective probabilities that investors and noninvestors get unclear results. For a given worker population, if the employer believes the fraction of investors is large, he will think that anyone with an unclear result is probably an investor. Conversely, if he thinks the fraction of investors is small, he will take an unclear result as a probable indicator of a noninvestor. So his assignment decision for a worker whose test is unclear ultimately rests on his belief about the fraction of investors in the subpopulation from which that worker has been drawn. If he thinks the fraction of investors is large enough, he will give the benefit of the doubt to a worker with an unclear test and assign him to task one; otherwise he will assign that worker to task zero.

3. Specifically, let p_1 (p_0) be the probability that an investing (noninvesting) worker gets an unclear test result. Then $1 - p_1$ is the probability that an investing worker passes the test, and $1 - p_0$ is the probability that a noninvesting worker fails it. I assume $p_0 < p_1$.

(7) I call the employer optimistic about a group of workers if he believes enough of them to have invested that when he sees one with an unclear result he nevertheless assigns him to task one. Otherwise I say he is pessimistic. I can express this by using the symbol π to denote the employer's belief about the fraction of investors in a group and by saying there is a critical belief π^* such that if $\pi \geq \pi^*$, then he is optimistic about the group, while if $\pi < \pi^*$ then he is pessimistic. I call the employer liberal toward a group if he gives them the benefit of the doubt, and conservative if he does not. So the employer is liberal toward groups about which he is optimistic, and conservative toward groups about which he is pessimistic. Because the employer observes a worker's color, he can distinguish between those drawn from the subpopulations of blacks and whites. Therefore, if his beliefs about the fractions of investors in these groups are not the same, it is possible that he treats black and white workers with unclear tests differently, based on this difference of belief. I say that the employer discriminates against blacks—and in favor of whites—if he is pessimistic about and conservative toward blacks while being optimistic about and liberal toward whites. To see how the employer might end up discriminating in an equilibrium of this model, we must consider the workers' behavior.

(8) A worker decides to invest only if he expects to gain more by doing so than it costs him. His gain from investing is the difference between the reward he expects if he invests and the reward he expects if he does not. Investing is beneficial because it raises the chance that a worker will be assigned to task one and thus enjoy the reward associated with that assignment. But the amount by which investing raises a worker's chance of getting this reward depends on whether the employer is liberal or conservative toward members of his group. If the employer is liberal, an investor is guaranteed to get task one, while a noninvestor gets it only if he does not fail the test. Thus investing raises the chance of getting the reward by an amount just equal to the probability that a noninvestor fails. On the other hand, if the employer is conservative, an investor gets task one only if he passes the test, and a noninvestor has no chance to get it. So in this case investing raises the chance of getting the reward by an amount just equal to the probability that an investor passes. Since I assumed the test is better at revealing noninvestors than investors, it follows that the gain from investing is greater if the employer is liberal than if he is conservative. Hence the fraction of a group of workers who would choose to invest is greater if they expect the employer to be liberal than if they think he will be conservative.

(9) I now identify the equilibria in this model. Denote by π_l (π_c) the fraction of workers in a group who would invest if they expected the employer to be liberal (conservative). If $\pi_l \geq \pi^*$, then, when a group of workers expects the employer to be liberal, sufficiently many invest as to make him optimistic. If $\pi_c < \pi^*$, then, when a group of workers expects the em-

ployer to be conservative, sufficiently few invest as to make him pessimistic. But an optimistic employer wants to be liberal and a pessimistic one wants to be conservative. So when $\pi_l \geq \pi^*$, it can be an equilibrium for the employer to be optimistic about and liberal toward any group and for that group to invest at rate π_l. And if $\pi_c < \pi^*$, it can be an equilibrium for the employer to be pessimistic about and conservative toward any group and for that group to invest at rate π_c. At least one of these conditions always holds. I will assume the parameters of the model to be such that they both hold, that is, $\pi_c < \pi^* \leq \pi_l$. Then there can be equilibria in which the employer is either optimistic or pessimistic about any group of workers, and in every case his belief turns out to be self-confirming.

(10) When the parameters of this model are such that $\pi_c < \pi^* \leq \pi_l$, it is possible for a discriminatory equilibrium to exist. In such an equilibrium the employer is, at the same time, pessimistic about one group—blacks, say—and optimistic about the other. Being pessimistic about blacks, he is conservative toward them when their test result is unclear. Being optimistic about whites, he is liberal toward them in the same situation. By behaving in this discriminatory way, he creates different incentives for workers in the two groups to become skilled at doing task one. But this difference in incentives is precisely what induces black and white workers to invest at different rates in the first place. That is, in a discriminatory equilibrium, the belief that blacks are on average less skillful than whites is a self-fulfilling prophecy. Given such beliefs, blacks do not enjoy equality of opportunity.

THE PROBLEM WITH AFFIRMATIVE ACTION AS A REMEDY IN THIS SITUATION

Of course, the foregoing model is highly stylized. It does not reflect many considerations that are important in real-world employment relationships. Nevertheless, it captures the essence of the problem I described in the introduction. It shows how an employer can come to rely on color as an indicator of the character of a worker, when other means of assessing the worker's merit—the test—fail. Moreover, it illustrates that the racial generalizations on which the employer relies need have nothing to do with the intrinsic qualities of the groups but instead may be the result of the fact that discrimination reduces the incentives of workers in the disadvantaged group to acquire skills.

In this discriminatory equilibrium, the employer is obviously not color-blind. A natural way for a policymaker to try to correct this discrimination would be to force the employer to assign workers from each group to each task at the same rate.[4]

4. A more direct way to eliminate discrimination would be to forbid the employer to treat whites and blacks with unclear tests any differently. That is, the government could merely insist on color-blind behavior from the employer, without regard to results. This would be difficult to enforce in practice. The government would have to observe all information upon which an employer might base his assignment—interviews, work history, and so on—to determine if he is really treating blacks and whites the same. In most employment situations this is not possible. The analysis offered

This policy, which I refer to as "affirmative action," is itself a departure from color-blind practice. It involves the government in monitoring the racial composition of the employer's work force in each task, insisting on equal proportionate representation. I will now examine in the context of the model set out previously whether this intervention eliminates the black-white difference in investment incentives that prevails in the discriminatory equilibrium. Imagine then that the employer, when faced with a worker whose test is unclear, assigns that worker to task one if he is white and to task zero if he is black. The fractions π_l of whites and π_c of blacks acquire the skill needed to do task one ($\pi_c < \pi^* \leq \pi_l$). Let the government enact a policy requiring that each racial group be assigned to each task at the same rate. Initially the employer is violating this policy. All whites who invest plus those who do not but whose test is unclear end up in task one, while only those blacks who invest and who pass the test do so. Since proportionately more whites than blacks are investing in this initial situation, a larger fraction of whites is being assigned to task one.

Therefore, in order to comply with the affirmative action mandate, the employer must either assign more blacks or fewer whites to task one. Since he is maximizing his profits in the initial equilibrium, both alternatives lower his net payoff. Which course is least undesirable to him, however, depends on the relative numbers of black and white workers in the population. In general the employer will try to minimize the number of instances where, in order to comply with the affirmative action policy, he has to assign a worker of either race to a task that he believes will not be most profitable for him. If blacks are comparatively few, then, by assigning more of them than he might desire to task one, he could meet the affirmative action mandate with a relatively small number of unprofitable assignments. On the other hand, if blacks are numerous in comparison to whites, then, by reassigning a relatively small number of whites to task zero instead of task one, he could meet the government's hiring requirement at least cost to himself.

I will assume here that blacks are a relatively small proportion of the total work force. If whites are sufficiently numerous relative to blacks, then the employer's best response to the government's mandate is to increase the number of blacks assigned to task one, while continuing to be liberal toward whites. Notice, however, that initially he will not think it adequate simply to engage in equal treatment of black and white workers in order to achieve this goal. Because a smaller fraction of blacks than of whites are investing initially, the employer anticipates that even if he becomes liberal toward blacks, he still will be assigning them to task one less frequently than whites. To achieve equal racial representation in the face of unequal racial investment rates, the employer will need to assign some of the blacks who fail the

here applies to those situations where affirmative action takes mainly a results-oriented rather than a process-oriented form, with the government's focus being on the numbers hired, not the hiring procedures.

test, and who he therefore knows have not invested, to task one as well. When he does this, I say that he is patronizing these black workers. The probability that a black worker who fails the test will nevertheless be assigned to task one is what I call the employer's degree of patronization. The precise degree of patronization the employer thinks he will need depends on his beliefs about the rates of investment by members of the two racial groups. The less skilled he thinks blacks are relative to whites, the more he anticipates a need to patronize them so as to comply with the government's mandate.

On the other hand, if blacks anticipate that they will be patronized, then they will want to reassess their decisions about skill acquisition. Any positive degree of patronization makes a worker's expected gain from investing less than it would have been if his group were merely treated liberally, but not patronized. Compared with liberal treatment, a positive degree of patronization raises the chance for a noninvestor to get into task one without affecting the fact that an investor is guaranteed to gain that assignment. Hence, compared with merely liberal treatment, a positive degree of patronization reduces the amount by which investing improves a worker's chances to get task one, and so lowers the fraction of workers who calculate that the benefit of investing exceeds its cost.

Consider now what happens when, starting from a discriminatory equilibrium, an affirmative action mandate is imposed. Because blacks are a relatively small fraction of the worker population, the employer's best response to the government's policy is to continue being liberal toward whites. Initially, he thinks the fractions π_e of blacks and π_l of whites are investing. He therefore anticipates the need for some patronization. By patronizing blacks, however, he alters their investment incentives and hence changes the rate at which they acquire the skill needed for task one. This change in black workers' behavior in turn implies that the employer must alter the degree of patronization required for compliance. Define an "equilibrium under affirmative action" to be a degree of patronization toward blacks together with a fraction of black investors such that (1) if the employer expects this fraction of blacks to invest, he would select the indicated degree of patronization in order to comply with the government's mandate; and (2) if the workers expect this degree of patronization, they would choose to invest at the indicated rate.

One equilibrium under affirmative action is obvious: if the employer should come to believe that blacks are investing at rate π_l, the same as whites, he would want to be liberal but not patronizing toward them and would comply with the government's mandate by doing so. If blacks expect liberal but not patronizing treatment they, like whites, would invest at rate π_l. When this equilibrium arises, the employer's initial discriminatory beliefs have been eliminated by the use of affirmative action. This is the ideal outcome predicted by proponents of the policy. The government's insistence on equal representation for each racial group creates a situation in which the opportunities, and so the

distribution of skills, for each group of workers are equalized. Having achieved this result, the policy of affirmative action can wither away, because the employer's discriminatory beliefs that warranted the initial unequal treatment of blacks have been dispelled.

Another equilibrium under affirmative action is less obvious: the employer continues to think blacks invest less frequently than whites. He therefore persists in patronizing them to some degree; but because blacks, when patronized, have less of an incentive to invest than whites, the employer's belief that patronization is needed becomes a self-fulfilling prophecy. This is not the outcome forecast by proponents of affirmative action. Rather than creating equality of opportunity, the policy in this case leads to a situation in which, in order to meet the government's requirement of equal representation, the employer favors unskilled blacks. Because noninvesting blacks have superior opportunities, the return from acquiring a skill is lower for blacks than whites, and relatively fewer blacks invest. The employer, therefore, has to continually favor black workers in order to comply with the government's mandate. In this equilibrium, affirmative action, far from withering away, sets in motion a sequence of events that guarantee that it will have to be maintained indefinitely. The incentives for the employer, and hence for black and white workers, are altered by the government's use of color-conscious strategy in such a way that a racial difference in workers' acquisition of skills is sustained. This is precisely the unintended negative consequence of racial preferences to which I alluded in the introduction.

It is therefore of some interest to determine which of these two equilibria under affirmative action will actually obtain. At the initial discriminatory equilibrium, the employer thinks he needs some patronization, but his use of it alters blacks' investment incentives. As black workers change their behavior, the degree of patronization that the employer thinks he needs also changes. Imagine a process in which the employer and black workers alternately adjust their behavior over a sequence of stages, each party reacting to the behavior observed from the other at the previous stage of adjustment. It is plausible to postulate that the equilibrium reached under affirmative action is the one that eventually emerges from this iterative process.

Using simple mathematics one can show that when $\pi_l < \frac{1}{2}$, this process culminates at the first—obvious—equilibrium described previously, and when $\pi_l > \frac{1}{2}$, it culminates at the second—less obvious—one. Another way of saying this is that the undesirable outcome obtains under affirmative action if, when facing a liberal employer, the average worker would strictly prefer to invest in the skill needed for task one. Recall that the average worker will want to invest when facing a liberal employer only if the expected return from doing so exceeds his investment cost. This expected return is greater, the greater the gain is to a worker from being assigned to task one and the lower the probability is that a worker who does not invest gets an unclear test

result. Thus the higher the value of assignment to task one is, relative to the average worker's investment cost, and the more powerful the test at identifying noninvestors is, the more likely it is that a patronizing equilibrium will arise under affirmative action. The patronizing outcome is also more likely when the disadvantaged group is a relatively small fraction of the total population.

CONCLUSION

The point of this exercise has been to illustrate, with the aid of formal economic reasoning, that the concerns expressed by some critics of affirmative action should be taken seriously. I have shown, in the context of a simple, stylized model of worker-employer interaction under racial hiring guidelines, that requiring equal representation of minority and majority groups in high-level positions may produce a situation in which the incentives provided minorities to acquire the skills needed to perform adequately in such positions are maintained permanently below the incentives provided majority workers. Whether this outcome occurs depends upon such factors as the proportion of the total work force belonging to the minority group, the advantage to a worker of obtaining a high-level position relative to the average cost in the population of acquiring the skill needed to perform in that position, the relative importance to the employer of assigning skilled and unskilled workers to their most productive positions, and the extent to which the employer can accurately gauge a worker's productivity in a given task before actually employing him there.

This article is not an attack on the practice of using preferential treatment as a tool to enhance opportunity for minority workers. Indeed, I have shown that sometimes the use of racial preference can have the desired results that its advocates predict. Departure from color-blind practice by the government, however, need not have these desirable consequences. It is important that we try to understand, in the many concrete circumstances in which preferences are now employed, just when the risks of generating negative unintended consequences of the sort I identify here are worth taking. Thus I am urging that more empirical research be done on the actual effects of affirmative action. Too often, both advocates and critics are content to base their arguments entirely on first principles, without reference to the direct or indirect consequences of this contentious policy. The analysis offered here is meant to graphically illustrate a possibility. Further study is required to identify practically significant cases exemplifying the effects uncovered here.