Notions of Fairness Versus the Pareto Principle: On the Role of Logical Consistency

Louis Kaplow and Steven Shavell†

In other writing, we advance the thesis that legal policies should be evaluated solely on the basis of their effects on individuals' well-being, meaning that no independent evaluative weight should be accorded to notions of fairness.[1] In that work, we consider a variety of principles of fairness, justice, and corollary concepts that are conventionally employed in the assessment of legal rules.[2] In the course of our research, we discovered that each of the leading notions of fairness that we examined could be shown to conflict with the Pareto principle; that is, consistent adherence to any of the notions of fairness would, in some circumstances, make everyone worse off. This observation led us to inquire about the generality of the conflict, and we explored it in two articles. In the first, we demonstrated that, in symmetric settings (in which every person is similarly situated), every individual will necessarily be made worse off whenever a welfare-independent notion of fairness is decisive.[3] In the second, a short, technical article intended for economists, we presented a formal proof of the proposition that, in all cases (symmetric or not), if a welfare-

† Harvard Law School and National Bureau of Economic Research. We thank the John M. Olin Center for Law, Economics, and Business at Harvard Law School for financial support. We have benefited from numerous exchanges with Howard Chang concerning the substance and many of the details presented in this comment during the year prior to our receiving the final version of his article. Our comment does not, however, reflect any particular points that may be raised in his rejoinder, which we have not seen as of this writing.

1. LOUIS KAPLOW & STEVEN SHAVELL, PRINCIPLES OF FAIRNESS VERSUS HUMAN WELFARE: ON THE EVALUATION OF LEGAL POLICY (John M. Olin Ctr. for Law, Econ., & Bus., Harvard Law Sch., Discussion Paper No. 277, 2000), *forthcoming in* 114 HARV. L. REV. (2001).

2. We note, however, that many notions of fairness concerned solely with income distribution are *not* included in the notions of fairness that we examine and criticize. This is because many distributive concerns involve individuals' well-being and are thus, by definition, accommodated in the welfare economic framework that we defend. *Id.* secs. II.A-.B. Indeed, we do not formally limit the distributive principles that may be employed (the favored distribution could be utilitarian or egalitarian, or one could give more to Joe because he is tall and less to Jill because her preferences are objectionable), although, of course, the spirit of some of our arguments and others that could be adduced would narrow the range of plausible distributive principles substantially.

3. Louis Kaplow & Steven Shavell, *The Conflict Between Notions of Fairness and the Pareto Principle*, 1 AM. L. & ECON. REV. 63 (1999).

independent notion of fairness is given any weight in making social decisions, there will exist circumstances in which everyone is made worse off.[4]

Howard Chang has written an article in this journal that addresses aspects of our second article.[5] He accepts the validity of our formal proof, but he challenges the appeal of our assumptions. Furthermore, he suggests that certain notions of fairness that do not satisfy these assumptions, including his own conception of liberal welfarism, could be sufficiently modified so that they would not conflict with the Pareto principle. (Chang's precise claim is unclear, however. The overall thrust of his article may give the reader the impression that many familiar notions of fairness might be susceptible to modification so as to avoid conflict with the Pareto principle, yet the analysis itself suggests only the logical possibility that the conflict can be circumvented with respect to modifications of certain types of notions of fairness, which are not formally specified.[6])

We begin this Reply by summarizing the demonstration in our first paper of the conflict between notions of fairness and the Pareto principle. This demonstration, which Chang does not contest, is easier to understand than our second, and it is independently sufficient to establish our conclusion. Then we consider our second, more technical demonstration of our conclusion. We emphasize that the two assumptions that Chang challenges are really minimal in character; in essence, they amount to requirements that normative theories be logically consistent. (Indeed, the relevance of centuries of moral philosophers' normative discourse depends on one of these assumptions.) We also explain how, even when one does not require logical consistency in the respects to be articulated, Chang's effort to show that certain notions of fairness can be altered so as to avoid

4. LOUIS KAPLOW & STEVEN SHAVELL, ANY NON-INDIVIDUALISTIC SOCIAL WELFARE FUNCTION VIOLATES THE PARETO PRINCIPLE (Nat'l Bureau of Econ. Research, Working Paper No. 7051, 1999), *forthcoming as* Louis Kaplow & Steven Shavell, *Any Non-Welfarist Method of Policy Assessment Violates the Pareto Principle*, 109 J. POL. ECON. (2001).

5. Howard Chang, *A Liberal Theory of Social Welfare: Fairness, Utility, and the Pareto Principle*, 110 YALE L.J. 173 (2000). Chang also considers other matters (notably, he criticizes an argument of Amartya Sen) that we do not take up here.

6. We should note at the outset that we have experienced general difficulty in interpreting Chang's article. Despite his technical economic training, Chang does not translate his argument and examples into unambiguous, formal statements or offer proofs of any of his claims, although our article to which he is responding consists almost entirely of formal statements and presents a proof of its main claim. The problem of Chang's imprecision includes the definition of fairness itself. We formally defined how we were using the term in each of our writings, KAPLOW & SHAVELL, *supra* note 4, at 2; Kaplow & Shavell, *supra* note 3, at 65-66 n.5; KAPLOW & SHAVELL, *supra* note 1, at 39 n.69, whereas we are not always sure whether Chang uses the term fairness in the same manner (for example, we cannot be sure he is excluding purely distributive notions, as we do, *see supra* note 2). Chang neither mentions our precise, formal definition at any point nor offers any definition of his own. (And, despite Chang's apparent displeasure with our definition, we note that it does encompass all of the many leading notions of fairness that we examine in KAPLOW & SHAVELL, *supra* note 1, pts. III-VI.)

conflicts with the Pareto principle is unsuccessful. Finally, we offer brief remarks on Chang's liberal welfarism, drawing upon arguments from our larger work.

Before proceeding, we should note that we are well aware that many readers will be reluctant to accept our claims. First, the idea that all plausible notions of fairness conflict with the Pareto principle may seem surprising; indeed, it came as a surprise to us during the course of our research. Yet, analysis has revealed it to be true to a general extent. Second, some of our analysis—particularly the second demonstration with which Chang takes issue—is of a technical nature. We endeavor to explain the relevant points in accessible terms so the reader can see what is really involved in considering them. Third, the fact that notions of fairness have broad intuitive appeal to everyone—including to us—seems in tension with our critique. In our conclusion, however, we briefly draw upon our other writing to sketch some of the ways that this appeal can be reconciled with our overall thesis that notions of fairness should not be employed as independent evaluative principles in the assessment of legal policy.

I. OUR FIRST DEMONSTRATION: NOTIONS OF FAIRNESS ALWAYS MAKE EVERYONE WORSE OFF IN SYMMETRIC CASES

A basic, natural setting to consider is the symmetric case: that in which all persons are similarly situated with respect to the policies under consideration. For example, in analyzing tort rules and the principle of corrective justice in the context of automobile accidents, one might examine cases in which every person is equally likely to be an injurer or victim and faces the same costs of accident avoidance, the same harm if an accident occurs, and so forth.

In such symmetric settings, we have demonstrated that, whenever a notion of fairness leads one to choose a policy different from that which would be chosen were the social goal concerned exclusively with effects on individuals' well-being, everyone will necessarily be made worse off.[7] The reasoning is straightforward. Suppose that a notion of fairness leads to the choice of legal rule *A* when overall well-being would be greater under rule *B*. Since overall well-being is greater under rule *B*, and since each person is identically situated, it must be that each person's well-being is greater under rule *B* (and by the same amount). Hence, everyone is made worse off by choosing rule *A*.

This simple demonstration, which Chang does not question, is by itself sufficient to establish that there is a conflict between any notion of fairness and the Pareto principle. For if one endorses a notion of fairness,

7. Kaplow & Shavell, *supra* note 3, at 68-70.

consistency requires that one cannot ignore symmetric cases, in which the conflict arises.[8]

Moreover, we suggest in our writing that the symmetric case has special significance for a range of systems of morality, including those associated with the Golden Rule, the categorical imperative, and choice behind a "veil of ignorance."[9] The reason is that such frameworks, which are designed to capture a notion of impartiality among individuals, in essence require that proposed moral principles pass muster in symmetric settings. (Consider, for example, the categorical imperative. If one did not demand that individuals be viewed as if they were symmetrically situated, a person who is mighty could advance the principle "might makes right," for such a person would happily generalize that principle, as that would simply enlarge his or her personal benefit. Only if one is required to assume symmetry among individuals—that no person is more mighty than another, or that a person is not more mighty than others any more often than anyone else is—would the test of the categorical imperative rule out such a self-interested principle.) We suspect that many readers, as well as most commentators upon whom Chang relies for various normative principles, do endorse the fundamental requirement of impartiality that is embodied in these frameworks. Thus, our demonstration that notions of fairness can only make individuals worse off in the symmetric case should be viewed as particularly important.

II. OUR SECOND DEMONSTRATION: IN GENERAL, NOTIONS OF FAIRNESS SOMETIMES MAKE EVERYONE WORSE OFF

In our second demonstration that any notion of fairness conflicts with the Pareto principle, we do not restrict analysis to symmetric settings. We believe it will be helpful to sketch our proof briefly. (The proof is somewhat abstract, and we ask the reader to bear with us.) Let us begin by considering any notion of fairness. (For example, consider the principle of corrective justice, under which wrongdoers are required to compensate their victims.) Now, as a theoretical matter, if this notion of fairness is given any weight (that is, if, other things equal, it sometimes affects the social

8. To an extent, Chang can be seen as challenging this demonstration, because, as we explain in Part II, he does not seem to accept the view that normative principles have to be applied consistently. We do note, however, that the continuity assumption that he challenges regarding our second demonstration is not employed in our first demonstration.

The reader may wonder about notions of fairness concerned with the distribution of income or well-being, which would be moot in symmetric cases. As we indicate above in note 2, however, such principles are not among the notions of fairness that we criticize, and we are explicit about this in our prior writing. *E.g.*, Kaplow & Shavell, *supra* note 3, at 67; KAPLOW & SHAVELL, *supra* note 1, sec. II.A.

9. Kaplow & Shavell, *supra* note 3, at 73-74.

decision), we must be able to imagine two regimes—call them *Fair* and *Unfair*—that have the following characteristics: First, every individual is equally well off in *Fair* and in *Unfair*; and second, one regime—*Fair*—is definitely more fair than the other, *Unfair*, and hence it is deemed normatively superior.[10] (To be concrete, suppose that *Unfair*, unlike *Fair*, does not follow the requirement of corrective justice that wrongdoers compensate victims in a class of cases; nevertheless, injurers are not better off than in *Fair* because they pay higher fines in *Unfair*, and victims are not worse off in *Unfair* because social insurance is provided to them.)

Next, consider a slightly modified unfair regime, *Unfair-II*, that is identical to *Unfair* except in one respect: There is a tiny savings of administrative costs in *Unfair-II*, which is distributed uniformly per capita. Now, if fairness has any real weight, it must be true that *Fair* is deemed superior overall to *Unfair-II*: After all, *Fair* was definitely superior to *Unfair*, *Unfair-II* is every bit as unfair as *Unfair*, and the cost advantage of *Unfair-II* over *Unfair* was stated to be tiny. (For example, compared to *Fair*, suppose that *Unfair-II* involves a major sacrifice of corrective justice and a trivial administrative cost savings, perhaps a penny per person.) However, observe that everyone is worse off in *Fair* than in *Unfair-II*. (This is because everyone is equally well off in *Fair* and in *Unfair*, while everyone is worse off in *Unfair* than in *Unfair-II*.) Hence, the notion of fairness has been shown to conflict with the Pareto principle.[11]

10. Chang, *supra* note 5, at 222 n.193, questions our observation to this effect and refers to our making strong and demanding assumptions, but we see his comments as red herrings. We simply ask the reader to contemplate the converse: Suppose that *no matter how much the degree of fairness differed between two regimes*, a notion of fairness *never* implied that one regime was superior to another when all else was equal, namely, when everyone had the same level of well-being. Clearly, there is no sense in which the notion of fairness is receiving any independent weight. (Put technically, our formal definition asks whether an evaluative principle can be implemented if the analyst knows how regimes affect individuals' well-being but has no information about any other characteristics of the regimes, including how fair they are. If the supposition in the text is false, then the analyst can make the evaluation knowing only well-being (for whenever the information on well-being is the same, so is the assessment); hence, as we define it, there is no independent notion of fairness involved.)

We are also puzzled by Chang's statements in the same note that our language usage (concerning the meaning of an "individualistic" social welfare function) is ambiguous and inconsistent. First, we offer a formal definition for the term in question. KAPLOW & SHAVELL, *supra* note 4, at 2. Second, our other use of the term to which Chang refers is also formally stated. *Id.* Third, although we do not offer a formal proof of stated equivalence between our two usages, this is only because the equivalence in our framework is (we thought) obvious. (When Chang questioned the equivalence, we supplied him with the proof, tracking the above-described intuition, which he has not questioned.)

11. Consider another way to see the intuition behind the general argument (which, to prove rigorously, requires stronger assumptions). (1) Suppose that there exist two regimes, *F* and *W*, such that *F* is more fair, welfare is higher in *W*, and *F* is viewed as superior overall to *W*. (This must be true of at least some such choices of *F* and *W* if fairness ever determines social choice.) (2) Construct *W'* from regime *W* as follows: Maintain the same degree of (un)fairness and total welfare, but redistribute income such that the resulting distribution of well-being in *W'* is the same as in *F*. (If, for example, the redistribution reduces inequality and one's welfare assessment favors a more egalitarian outcome, the overall adjustment in generating *W'* from *W* would reduce total

We now can consider the two assumptions that Chang challenges. The first, referred to in formal literature as "continuity," involves our formulation of the idea that a notion of fairness has "real" weight. The justification for our continuity assumption will, we believe, be apparent once we make its meaning and implications clear (which, we believe, Chang does not do).[12] Continuity, which we use in its standard mathematical sense[13] in our proof, turns out to have a fairly simple meaning here, namely that the weight given to a notion of fairness is not infinitesimal. In other words, we assumed that a given, perhaps unboundedly large amount of unfairness was more important than some, however small, savings in administrative costs (shared per capita). The contrary assumption—which is really what Chang, upon examination, is asking readers to embrace[14]—is that, no matter how much unfairness is involved, it can be outweighed by the tiniest amount of administrative cost savings, whether it be one cent or even one billionth of a cent. (We emphasize that if one caps the fairness-cost tradeoff at, say, a trillion to one, then continuity formally obtains and our proof holds; it truly is the case that one must embrace the most extreme view of the triviality of the weight to be given to fairness to avoid our conclusion.) We grant that our rejection of

income sufficiently to keep total welfare constant.) (3) Since *F* is more fair than *W'* by the same amount that it was more fair than *W*, and since welfare is no higher in *W'* than it was in *W*, it must be that *F* is deemed overall superior to *W'*. (4) However, *W'* has higher total welfare than *F* and also the same distribution of welfare as *F* (that is how we constructed *W'*). Hence, everyone must be worse off in *F* than in *W'*, even though *F* is judged to be superior overall under the notion of fairness.

12. Chang's most direct statement about continuity is, we believe, misleading. He states that "[t]here is no apparent reason why a slight increase in consumption of some good might not have a discontinuous impact on social welfare, especially if it is so widespread as to be shared by all individuals in society." Chang, *supra* note 5, at 224. First, he does not explain, as we do in the text to follow, that discontinuity means that the ratio of the impact on social welfare to small changes in consumption becomes infinite (as one considers smaller and smaller changes in consumption). Thus, any slight increase in consumption (say, a peanut per capita) is posited by Chang, in finding discontinuity plausible, to be vastly more important than any degree of unfairness, no matter how large. (This explains why Chang's other statements, *see id.* at 210-11 & n.165, are also incorrect with regard to the concept of continuity.) Second, the idea that the number of individuals might be relevant is fallacious, for the amount of increase in the consumption good per person can be arbitrarily small—indeed, whatever unit one imagines (say, one peanut) can be divided by the number of people in determining how much of a per capita increase to consider (so, if the population is a million, we can imagine each individual gaining by only a millionth of a peanut). In addition, Chang states that the literature on social choice theory generally does not assume continuity. *Id.* at 224 n.197. Chang does not mention, however, that continuity is often moot in much of that literature, because it is concerned only with discrete possibilities (that is, it does not allow anything, such as the amount of a good an individual has, to vary gradually) and because it uses only rankings (which, for example, allows no statements about degrees of preference or well-being, which renders consideration of most notions of distributive justice impossible). In our more encompassing and realistic framework, the assumption of continuity is relevant and, as we explain in the text, compelling.

13. Chang repeatedly refers to our "particular" continuity property, *e.g.*, *id.* at 223, although our usage is entirely standard, straight from any textbook or dictionary that includes mathematical terms.

14. *See id*. at 222-26.

Chang's view does amount to an assumption, but it is one that we imagined would be endorsed by anyone who believed that a notion of fairness was worth taking seriously.[15] (And, not surprisingly, virtually all notions of fairness we have studied, including all of the major ones we examine in detail in our main project, do not violate this continuity assumption.)

We offer some brief, additional remarks about our continuity assumption. (1) Formally, our argument only requires that the principle of fairness be continuous *in something*. (Hence, corrective justice should not be given infinitesimal weight with respect to administrative cost savings, trivial aesthetic pleasures, or the consumption of some good—in other words, to some factor that is unrelated to the notion of fairness.) (2) For our proof not to apply to a notion of fairness, the discontinuities that Chang suggests be allowed must exist not merely here or there, but in every conceivable setting in which a notion of fairness might be given any weight (for our proof that a Pareto conflict can be shown to exist applies at every point where fairness matters). (3) The direction of the involved discontinuity is opposite to that which is sometimes posited. That is, the required discontinuity gives absolute weight to, say, administrative costs, so that consideration of such costs would trump any amount of unfairness, rather than vice versa.

The second assumption implicit in our demonstration that Chang challenges is "independence."[16] At its core, the notion of independence that we employ (which differs from the one that is the subject of part of Chang's discussion)[17] is an aspect of logical consistency, and, as such, we are quite surprised that Chang questions it. One way to express our independence assumption is as follows. Suppose that a person believes that regime *X* is clearly morally superior to regime *Y*, and also that regime *Y* is clearly morally superior to regime *Z*. Then, our independence assumption holds that that person is not free to assert that regime *Z* is clearly morally superior

15. For example, we would not take seriously a notion of fairness that was used to break literal ties (where even a fraction of a cent would swing the decision one way or the other) but that otherwise never mattered.

16. Chang, *supra* note 5, at 226-32.

17. Chang suggests that our independence assumption is analogous to that used in Arrow's Theorem and thus may be subject to criticism. *Id.* at 226-29. But our independence assumption is, in important respects, weaker than Arrow's. Arrow assumed that the social ranking of *X* and *Z* can depend only on different individuals' *rankings* of *X* and *Z*—and thus not on the intensity of their preferences or on any difference between one individual's gain and another's loss (interpersonal comparisons). This limitation implies, among other things, that ordinary distributive judgments are ruled out in social decisionmaking. This aspect of Arrow's assumption is specifically identified by those Chang quotes, *id.* at 228-29 & n.206, in support of the contention that independence is a controversial assumption. *See* ALFRED F. MACKAY, ARROW'S THEOREM: THE PARADOX OF SOCIAL CHOICE 48 (1980); DENNIS C. MUELLER, PUBLIC CHOICE II 393-95 (1989). Our independence assumption, by contrast, does not rule out preference intensities or interpersonal comparisons (and thus allows distributive judgments). Rather, it only requires logical consistency in the sense discussed in the text. Chang does not mention this crucial difference.

to regime *X* on the ground that regime *Y* is not "on the table" in some sense.[18] (For example, only regimes *X* and *Z* may currently be before the decisionmaker, or regime *Y* may not be politically or practically feasible.) Chang rejects this implication. *In rejecting our independence assumption, he is asserting that the morally correct choice between* X *and* Z *generally should, as a matter of principle, depend on the presence or absence of* Y*—even when it is acknowledged that* Y *should never be chosen in any event*. Quite frankly, we find his position absurd.[19] (We also note that philosophers have always emphasized the importance of logical consistency in the sense at issue; indeed, their frequent use of hypothetical examples to illuminate moral questions, including hypothetical constructs such as the categorical imperative, means that they consider them relevant for normative choices even though, being hypothetical, they are not "on the table" at the moment of a particular, actual decision.)

Though we may be belaboring the obvious, we note some implications of violating independence. (1) One would need a criterion to determine which regimes were and were not to count as on the table (which seems hard to come by since we are considering regimes that are not going to be chosen in any event). (2) Chang's proposed procedure (described below) deems feasible regimes to be those on the table,[20] but without costly inquiry (which is hardly worthwhile if the regimes are not to be chosen), one may not know which these are. (3) One's normative choice would change as

18. It may not be apparent how our proof, sketched above, uses this assumption. One way to express the point is that, when we hypothetically contemplated regime *Unfair-II*, we implicitly assumed that the statement that regime *Fair* was superior to regime *Unfair* was not thereby rendered invalid. To be sure, as we note in the text to follow, it is common and appropriate in moral argument to change one's view in some cases through the contemplation of other cases. Our point is that, once one believes one has arrived at a complete and correct moral view, after considering all imaginable cases, one is not free to proclaim both that *X* is morally superior to *Z* and that *Z* is morally superior to *X*, solely depending on what else is stated to be "on the table" at the moment of an actual decision. (We acknowledge that there may be some instances in which a choice might be said to depend upon other options, such as when a person might be upset at the very fact that an available, preferred option has been rejected, but such possibilities are irrelevant in our framework because the regimes that are to be assessed by a normative criterion are explicitly defined as *complete* descriptions, *see* KAPLOW & SHAVELL, *supra* note 4, at 1, which would already include any such feelings. The moral question is: Given all such information, which is morally superior, *X* or *Z*?)

19. In some of our private correspondence, Chang seemed to accept our view that rejecting independence was morally indefensible, indicating that he merely wished to establish as a logical matter that our position depended on this assumption. Now it is surely correct that logic itself does not require that one accept logical consistency in normative analysis, but if this is the extent of his challenge, then there is no dispute about whether the assumption is ultimately an appropriate—indeed, a compelling—one. In fact, elsewhere in his article, he takes Sen to task for deeming a hypothetical regime irrelevant to moral analysis on grounds of infeasibility: Sen's "objections go to the 'pragmatic' question of whether the solution is *feasible*, not to the 'ethical' question of whether the trade would be socially *desirable* if it were available to us as a social choice, and it is the ethical question that is relevant for our purposes here." Chang, *supra* note 5, at 202; *see also id*. at 202 n.135 (citing Hammond's statement that arguments such as Sen's involve "misconceptions" due to "confusion").

20. *Id*. at 214.

views on the feasibility of such other regimes change (and feasibility might even change if, say, self-interested individuals spend resources to make regimes feasible, just to change the social choice—again, even though the newly feasible regime would not be chosen). In all, we are truly at a loss in understanding why Chang insists on taking such an approach.[21]

Suppose, however, that one follows Chang and is willing to consider notions of fairness that do not have the basic elements of consistency that we have described. One might imagine from the tenor of Chang's reply that it then would be straightforward to construct notions of fairness that did not violate the Pareto principle. But this is not the case. Our second demonstration shows that our assumptions are *sufficient* to imply that all notions of fairness conflict with the Pareto principle; it is not suggested that they are necessary, and they are not. Hence, it is possible that notions of fairness that do not obey our assumptions will still conflict with the Pareto principle. Now Chang does not offer direct, precisely stated examples of existing, modified, or novel notions of fairness under which the conflict with the Pareto principle is avoided; nor, correspondingly, does he indicate the nature of such principles (an important question, given how strange they may have to be once our consistency assumptions are violated).[22] Instead, he describes procedures under which a conventional notion of fairness is the input and, he asserts, a modified notion of fairness that is free of Pareto conflicts is yielded as the output.[23] Chang's procedures do not, however, seem to advance his position.

21. To illustrate some of the problems with both of his arguments about our assumptions, we offer an example that is in the spirit of much of Chang's discussion of the idea that individuals should be allowed to waive (alienate) their rights if it is to their advantage. In regime *A*, Bill, the richest, most well-off (and most unpleasant) person in the society does very well, and every other person is in misery, on the brink of starvation; no one's "rights" are violated. In regime *B*, Bill is a cent worse off than he is in *A*, because a trivial violation of one of his rights occurs, and everyone else is markedly better off. Chang's approach would require adopting regime *A*, because Bill does not have to waive his right, which can be given infinite weight even though Bill hardly cares about it and the rest of society suffers greatly. Now, if we imagine a regime *C*, in which no rights are violated, Bill is much worse off than in *A*, and everyone else is better off than in *A* but not as well off as in *B*, Chang's preference for higher overall welfare would lead him to choose *C*. This time Bill has no veto because his right is not violated (even though Bill loses far more moving to *C* than to *B*, where we gave him a veto). Of course, if one compares *B* and *C*, one can see that everyone is better off in *B*—yet *B* was rejected (versus *A*) and *C* was adopted (versus *A*). This, then, raises the question of which regime should really be chosen. (*A* is chosen over *B*, because Bill has a right; *B* over *C*, because of adherence to the Pareto principle; and *C* over *A*, because of the welfarism half of Chang's liberal welfarism.) If the choice is purely between *A* and *B*, Chang says *A* is morally superior. If among all three, he says *B* is morally superior.

22. Relatedly, since the modified notion of fairness that emerges from Chang's procedure may differ markedly from the original notion with which one began, it is also important to inquire whether the original rationale for the notion will continue to be applicable (an inquiry that obviously will be difficult if one cannot readily determine what the modified notion will look like).

23. Chang, *supra* note 5, at 215-19.

First, assume for the sake of argument that his procedures do work. They only guarantee that the principle that emerges will not have Pareto conflicts; nothing assures us that the principle will retain *any* substantial fairness (that is, welfare-independent) content. Indeed, if the modified principle satisfied our consistency assumptions, we know from our proof that it would not retain any substantial fairness content. And, even if the assumptions are violated, they were only sufficient conditions, so for all we know there will be no fairness left after the procedures do their work. (Moreover, none of the examples of modified fairness principles that Chang has informally described to us over the past year, in his article or otherwise, has in fact been a Pareto-conflict-free notion that is not solely based on individuals' well-being.[24])

Second, in the domain our proof addresses, Chang's procedure cannot work. The reasons, which are somewhat technical, are roughly as follows: Even after an infinite number of iterations of his procedures, one would be zero percent of the way done in generating the modified notion of fairness, and the moves required in most of the iterations are impossible to make, even in principle.[25] In addition, it turns out that one could not even use a

24. In prior exchanges with Chang, we have shown that his example, *id.* at 220-21, like others he has suggested to us previously, fails. In the manner he has verbally interpreted the present example to us, we have shown that his modified fairness function is not a notion of fairness as we define it. (In essence, the only "fairness" left concerns distribution, which, as we have said explicitly in all of our writings, is not included in our definition and thus not subject to our claim. *Supra* note 2. Indeed, Chang identifies two aspects of social decision in his example. Chang, *supra* note 5, at 220. The first ends up being governed by welfare, and not fairness, to avoid a Pareto conflict, *id.* at 221, and the second involves the need to determine the distribution of wealth, *id*. We further showed that other interpretations of his example (which may be what many readers would infer Chang means from the first half of his article) would involve a violation of the Pareto principle. Chang's statement that he has given an example of a notion of fairness (as we have consistently defined it, *see supra* notes 2, 6) that does not conflict with the Pareto principle is no more than a pure assertion that we have previously demonstrated to be false.

25. Chang's procedures make various comparisons, choose "best" points, and change "ranks" in particular ways. However, on the domain that we considered, which realistically allows for continuous variation (for example, in how much funds may be spent on courts or in how much of a good people may have), there are what is formally referred to as an uncountably infinite number of comparisons to make. And, the concept of "best point" (used in his first procedure) is not defined, except at boundaries, as Chang acknowledges. *Id.* at 218 n.183. (By analogy, if one is asked to state the second highest real number in the interval from 0 to 10.0, inclusive, after the highest number (10.0) is removed, there is, as some readers will recall, no such thing as the next highest real number. Yet Chang's procedure cannot move forward until the second highest number is identified.) Moreover, the changes in "ranks" required in his second procedure—which Chang explicitly offers to remedy the preceding problem—are undefined for essentially the same reason. (Chang states that one should give an alternative a rank that is "higher than the rank in question but lower than any higher rank." Chang, *supra* note 5, at 219. Suppose that the "rank" in question is 10.0 and all the real numbers between 10.0 and 11.0 have been assigned to other alternatives. Obviously, there does not exist a real number higher than 10.0 but no higher than any other real number between 10.0 and 11.0.) Chang's failure to write his procedures in a more formal manner, actually to show their operation on the sort of domain in question (which allows continuous variation of parameters), and to prove that any particular output of the procedures has the alleged properties—despite our suggestions for over a year that he do so—is suggestive of their underlying deficiencies.

streamlined version of his procedure in an attempt to provide rough justice. The reasons have to do with the very assumptions he eschews: Because he rejects continuity, it is literally true that a very close approximation could be wrong to an extent that is deemed infinitely weighty; and because he rejects independence, it is always possible that skipping over some possibilities—even ones that will not be chosen in any event—will be fatal, because which possibilities one considers or skips can have a morally decisive effect on other (seemingly unrelated) assessments.[26]

III. CHANG'S LIBERAL WELFARISM

Most of Chang's discussion of his particular blend of liberalism and welfarism has little if any bearing on our claim that notions of fairness conflict with the Pareto principle. Nevertheless, we suspect that the aspects of Chang's argument that suggest that individuals' "external" preferences should be censored have intuitive resonance with many readers, for most of us probably cringe at the thought of crediting, say, sadistic preferences. Accordingly, we offer some comments about such external preferences drawing upon our other writing, which considers a number of general arguments that bear on external preferences and also contains a section that addresses the specific topic.[27] (We will be brief here, because the question of how to address external preferences is not our focus in this Reply or in the article that Chang criticizes.)

On its face, the idea that external preferences should be ignored raises a number of questions, including the following: What is an "external" preference? (If it means any preference regarding other individuals, as suggested in most discussions, then one would have to censor preferences for watching opera and sports events, for conversation and companionship, indeed, for virtually all human interaction that was not purely instrumental to some other end.[28]) Why should people have preferences regarding only

26. Thus, his statements that we could use "a reasonable rule of thumb" and eschew "investing much of our scarce resources in the search for . . . Pareto improvements," *id.* at 230-31, are wrong in his own framework. (Relatedly, he suggests that one may focus primarily on the fairness optimum, *id*. at 231, but because he rejects our assumptions, this works only if that optimum is identified in a literally perfect, precise manner, which of course is impossible.) It should also be clear, contrary to Chang's suggestion, *id*. at 219 n.186, that these defects are unique to his particular approach. For example, under welfare economics (and most notions of fairness that people actually believe in), only the policies under consideration would have to be evaluated in order to choose among them, and the use of approximations would usually yield approximately good results. (Indeed, in the latter part of his footnote, he acknowledges differences between welfare economics and his approach but does not reveal how they render his, and only his, procedure inoperable.)

27. *E.g.*, KAPLOW & SHAVELL, *supra* note 1, secs. II.D, VIII.B.

28. Chang himself is difficult to interpret. He does not offer a definition of external preferences and criticizes the one definition in the literature that he mentions explicitly. Chang, *supra* note 5, at 183 n.40. In response to one of the definitional problems he mentions briefly, he

other things and not other people? (Is a particularly narrow sort of materialism meant to be endorsed?) Is it really acceptable to trump all other-regarding preferences, including love for one's children, concern for individuals in distress, and so forth? And what of negative external preferences that have socially desirable effects (such as the feelings of disapprobation that we have when others act wrongfully, the anticipation of which often deters individuals from improper behavior)? Since, we presume, only some preferences—and, from the foregoing, it would appear, only some external preferences—are to be censored, who gets to choose which ones? Using what criteria? On what grounds?[29]

We suggest, in our other writing, that the welfare economic framework offers answers to these sorts of questions, allowing one to make reasoned arguments about which external preferences really are socially problematic and how and when it may make sense to adjust social policy in the light thereof. In addition, that analysis suggests that the appeal of the common view, which Chang exemplifies, in fact is related to these underlying welfarist arguments. For example, we explain that it is no accident that negative and socially counterproductive external preferences are most often invoked in support of the preference-censoring view. Upon reflection, the widely held antipathy (itself an external preference of sorts) toward such other-regarding preferences as sadistic ones can contribute to the proper socialization of individuals and also serve to discourage undesirable behavior that may be undertaken to satisfy these types of preferences.

IV. CONCLUSION

In this Reply, we have reviewed our two main arguments, each of which independently demonstrates that consistent adherence to any notion

asserts that we "can distinguish" the problematic preferences from others, by reference to which preferences "are not 'intrinsically immoral,'" *id.* at 189 n.67, a patently question-begging response, since the whole point of the distinction was to identify which preferences should be deemed immoral. Finally, Chang concludes his discussion of censoring preferences by stating that, to avoid what he regards as decisive objections to the preference-censoring position as formulated by others, he would admit (that is, not censor) "personal preferences based on external preferences." *Id.* at 191-94. This seems to imply, however, that although he would exclude the preferences of, say, sadists that were expressed in terms of the supposed political inferiority of their victims, he would admit sadistic preferences reflecting the pure pleasure of seeing the victims suffer—hardly a distinction that succeeds in capturing the intuitions that Chang claims motivate the preference-censoring view in the first place.

29. There exist additional difficulties as well. It is not clear, for example, how to define what it means to censor a preference since, in general, the intensity of an individual's other preferences may well depend on the extent to which the preference to be censored is in fact satisfied. Also, paradoxically, censoring a preference could lead to results under which more of the undesirable activity occurs. (Ignoring a sadist's preferences may lead us to deem him worse off—because a source of his satisfaction is not counted—which under some distributive theories would entitle him to a greater share of resources, part of which he may devote to satisfying his sadistic preferences.)

of fairness will make everyone worse off in some circumstances. A focus of our discussion has been on the assumptions underlying the second of our demonstrations of this point. We emphasized that the assumptions embody basic elements of logical consistency that we would think anyone advancing a normative principle for evaluating legal policy would embrace. Although logic alone cannot demonstrate a normative conclusion, we do believe that our result poses a challenge to those who would employ notions of fairness in policy analysis, particularly since many such notions seem motivated by concerns for individuals who might be unfairly treated, whereas consistent pursuit of notions of fairness, we have shown, may be to the detriment of everyone, including those same individuals.

We have not rehearsed a wide range of additional arguments that we advance in our larger work in support of our general thesis that legal policy analysis should depend exclusively on effects on individuals' well-being, with no independent weight being given to notions of fairness. We do, however, wish to elaborate briefly on one of the themes of this writing, namely, how our thesis can be reconciled with the intuitions and instincts that most individuals (including us) hold regarding various notions of fairness.

An important part of the reconciliation concerns the respects in which notions of fairness may be relevant under welfare economics, even though they are not considered to be independent evaluative principles. First, individuals may have tastes for notions of fairness, which is to say that their well-being may depend on whether what they view to be fair treatment is in fact provided. (Punishment that is substantially disproportionate to an offense may be upsetting to many individuals.) Second, notions of fairness may serve as proxy principles that may be useful in identifying policies that advance welfare. (The notion of corrective justice holds that wrongdoers should compensate their victims, the prospect of which tends to enhance deterrence.) Third, notions of fairness can be important as rules of common morality, which are valuable to teach and reinforce because they lead individuals to be less opportunistic in their interactions in their everyday lives. In all, we believe that the broad appeal of notions of fairness can in many respects be reconciled with what we have shown to be serious defects in the use of notions of fairness as independent evaluative principles—including, importantly, that consistent adherence to such notions entails the view that it may be desirable to choose policies that make everyone worse off.