2000

A Liberal Theory of Social Welfare: Fairness, Utility, and the Pareto Principle

Howard F. Chang
University of Pennsylvania Carey Law School

Article

A Liberal Theory of Social Welfare: Fairness, Utility, and the Pareto Principle

Howard F. Chang†

CONTENTS

† Professor of Law, University of Pennsylvania Law School. I wish to thank Anita Allen-Castellitto, Scott Altman, C. Edwin Baker, Richard Craswell, Ronald Dworkin, Peter Huang, Heidi Hurd, Christine Jolls, Michael Knoll, Lewis Kornhauser, Michael Moore, Eric Posner, Hilary Sigman, Lynn Stout, and seminar participants at the University of Pennsylvania, New York University, Georgetown University, and the annual meetings of the American Law and Economics Association and of the National Bureau of Economics Research Summer Institute for helpful comments. I am especially grateful for detailed comments from Louis Kaplow and Steven Shavell.

The economic analysis of law evaluates legal regimes by the criterion of social welfare, which economists usually take to be a function of the utility that individuals enjoy under the laws in question. Economists generally define the utility enjoyed by each individual as the satisfaction of that individual's preferences over various states of the world. Economists normally assume that any reasonable notion of social welfare would conform to the Pareto principle, which holds that if each individual prefers one state of affairs over another, then social welfare must be higher in the first state than in the other state.

Amartya Sen, in his influential article entitled *The Impossibility of a Paretian Liberal*, shows how liberal rights, such as the right to read a book of which others disapprove, can produce outcomes that everyone would prefer to avoid, thereby violating the Pareto principle.[1] Sen infers that one cannot uphold both liberal values and the Pareto principle. Disturbed by the implication that "individual liberty may have to be revoked" under the Pareto principle, Sen concludes that the conflict that he exposes indicates "the unacceptability of the Pareto principle as a universal rule."[2]

In a similar vein, Louis Kaplow and Steven Shavell identify potential conflicts between the Pareto principle and notions of "fairness," which give weight to considerations other than the overall utility level of each individual.[3] Indeed, Kaplow and Shavell claim that "any conceivable notion of social welfare that does not depend solely on individuals' utilities" implies a "social welfare function" that violates the Pareto principle.[4] They infer that "as a matter of logical consistency, a person who embraces a notion of fairness must on some occasions favor adopting legal rules that would make every person worse off," a conclusion that they view as "a serious challenge to proponents of notions of fairness who also care about the well-being of members of society."[5] We might call their claim "The Impossibility of a Fair Paretian." Unlike Sen, however, Kaplow and Shavell view their impossibility claim as a critique of all fairness notions (including liberal rights) rather than of the Pareto principle as a universal rule.

1. Amartya Sen, *The Impossibility of a Paretian Liberal*, 78 J. POL. ECON. 152 (1970). The Royal Swedish Academy of Sciences cited Sen's work on this issue as one of the contributions justifying its award of the 1998 Nobel Prize in Economics to Sen. *See* Royal Swedish Acad. of Scis., http://www.kva.se/eng/pg/prizes/economics/1998/ecoback98.pdf (last visited Sept. 24, 2000).

2. Amartya Sen, *Liberty, Unanimity and Rights*, 43 ECONOMICA 217, 235 (1976).

3. *E.g.*, Louis Kaplow & Steven Shavell, *The Conflict Between Notions of Fairness and the Pareto Principle*, 1 AM. L. & ECON. REV. 63, 65-66 & n.5 (1999).

4. LOUIS KAPLOW & STEVEN SHAVELL, ANY NON-INDIVIDUALISTIC SOCIAL WELFARE FUNCTION VIOLATES THE PARETO PRINCIPLE 1 (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. Kaplow & Shavell, *supra* note 3, at 76.

That is, Kaplow and Shavell use their claim to advocate "welfarism," which makes moral judgments a function only of the utility of individuals.[6] Welfarism includes a broader class of moral theories than utilitarianism, which takes social welfare in a given population to be equal to the sum of individual utilities. A welfarist theory need not assume that social welfare for a given population is equal to the sum of individuals' utilities. Kaplow and Shavell, for example, assume a more general social welfare function that permits the distribution of a given amount of utility among individuals to affect social welfare.[7] Any form of welfarism, however, including utilitarianism, ranks states of affairs entirely on the basis of utility, regardless of other information about those states. Thus, the claims of Sen and of Kaplow and Shavell both have important implications for the fundamental normative question of what criteria we should ideally use to evaluate laws or public policies.

Both Sen's claim and Kaplow and Shavell's claim refer to the "weak" version of the Pareto principle, which holds that if *every* individual prefers any alternative x to another alternative y, then x is socially preferable to y.[8] The "strong" Pareto principle holds that if *at least one* individual prefers x to y, and no one prefers y to x, then x is socially preferable to y.[9] Thus, under the strong Pareto principle, x may be socially preferable to y even if all but one individual is indifferent between x and y. The strong Pareto principle is the stronger condition insofar as it implies the weak Pareto principle. That is, if the strong Pareto principle holds, then the weak Pareto principle must also hold. If the weak Pareto principle holds, however, the strong Pareto principle does not necessarily hold.

The weak Pareto principle is a relatively weak constraint on a social welfare function. It is a "weak form of welfarism," in the sense that while welfarism "asserts that non-utility information is *in general* unnecessary for social welfare judgements," the weak Pareto principle "makes non-utility information unnecessary *in the special case* in which everyone's utility rankings coincide."[10] Thus, if we believe in any welfarism in which social welfare is an increasing function of each individual's utility, including utilitarianism, then this belief will imply that the weak Pareto principle must hold. Kaplow and Shavell, however, make the surprising claim that

6. *E.g.*, Amartya Sen, *Utilitarianism and Welfarism*, 76 J. PHIL. 463, 468 (1979) (defining "welfarism" as the view that "[t]he judgment of the relative goodness of alternative states of affairs must be based exclusively on, and taken as an increasing function of, the respective collections of individual utilities in these states"). Thus, welfarism rejects the relevance of nonutility information. *Infra* note 22 and accompanying text.

7. Kaplow & Shavell, *supra* note 3, at 65-66 n.5.

8. *Id.* at 65 n.3; Sen, *supra* note 1, at 153.

9. Amartya Sen, *Social Choice Theory*, *in* 3 HANDBOOK OF MATHEMATICAL ECONOMICS 1073, 1075 (Kenneth J. Arrow & Michael D. Intriligator eds., 1986).

10. Amartya Sen, *Personal Utilities and Public Judgements: Or What's Wrong with Welfare Economics?*, 89 ECON. J. 537, 549 (1979).

the converse implication is also true: If we believe in the weak Pareto principle, then we must also believe in welfarism.

The weak Pareto principle is also an appealing criterion. Even Sen, who argues that the Pareto principle is unacceptable "as a universal rule," concedes that "there is something very central in the idea that preferences unanimously held by members of a community cannot be rejected by that community."[11] Kaplow and Shavell infer that because "most analysts who accord importance to notions of fairness would not want to contravene the unanimous preferences of the population," they should find the conflict between fairness and the Pareto principle "troubling."[12] After all, why would one ever want to violate the Pareto principle? When we do so, "everyone is worse off, including any person of possible concern under a notion of fairness."[13] Anyone who urges us to follow a rule that requires us to violate the Pareto principle is vulnerable to the charge of "superstitious 'rule worship,'" that is, "the charge of heartlessness, in his apparently preferring abstract conformity to a rule to the prevention of avoidable human suffering."[14] Why should we follow a fairness rule that requires us to comply even when it serves no one's interests? Kaplow and Shavell conclude that "fairness-based analysis stands in opposition to human welfare at the most basic level"[15] and that we should therefore reject all fairness theories and base social choices on individual utility alone.[16]

I argue in this Article that both Sen's critique of the Pareto principle and Kaplow and Shavell's critique of fairness notions are based on questionable assumptions. This Article proposes a middle course, which I argue resolves the supposed conflicts while remaining faithful to both liberal fairness principles and the Pareto principle. Thus, Kaplow and Shavell cannot reject all fairness concerns by relying on the Pareto principle alone, nor can Sen reject the Pareto principle by relying on liberal rights alone. Both positions require stronger and more controversial assumptions to prove their claims.

This Article outlines a moral theory that takes social welfare to be a function of more than just individual utility. In this Article, however, I do not defend a particular theory of fairness against all possible objections. Specifying completely the precise content of a liberal theory of fairness is

11. Sen, *supra* note 2, at 235.

12. Kaplow & Shavell, *supra* note 3, at 64.

13. *Id.* at 73.

14. J.J.C. Smart, *An Outline of a System of Utilitarian Ethics*, *in* J.J.C. SMART & BERNARD WILLIAMS, UTILITARIANISM: FOR AND AGAINST 3, 6 (1973).

15. LOUIS KAPLOW & STEVEN SHAVELL, PRINCIPLES OF FAIRNESS VERSUS HUMAN WELFARE: ON THE EVALUATION OF LEGAL POLICY 47-48 (John M. Olin Ctr. for Law, Econ., and Bus., Harvard Law Sch., Discussion Paper No. 277, 2000), *forthcoming in* 114 HARV. L. REV. (2001).

16. *Id.* at 1, 454.

an ambition well beyond the scope of this Article. This Article does not, for example, present a defense of the Pareto principle as a universal rule. Instead, I assume for the sake of argument that we believe in the Pareto principle, because I am interested in exploring the logical implications of that belief. In particular, I analyze whether this belief must invariably produce a conflict with liberal notions of fairness.[17] My goal is simply to demonstrate that plausible theories of fairness, especially a liberal theory, need not violate the Pareto principle. Thus, although Kaplow and Shavell may present an effective critique of some theories of fairness, their critique cannot reasonably be viewed as effective against all theories of fairness.

Part I of this Article presents a critique of welfarism from a liberal perspective and proposes a liberal alternative to the utilitarian theory of social welfare. I first discuss the conflict between classical utilitarianism and liberal values and then describe how various economists and philosophers, including Ronald Dworkin, have proposed modifying the utilitarian notion of social welfare to incorporate liberal notions of fairness. Drawing upon this literature, I outline the basis for a liberal notion of social welfare. In particular, I take issue with welfarism's indiscriminate inclusion of all forms of satisfaction, including the satisfaction of racist or malicious preferences, in the calculation of social welfare. I argue that a liberal theory would exclude the satisfaction of such objectionable preferences from our notion of social welfare.

Part II describes how these liberal fairness notions raise the possibility of conflicts with the Pareto principle and sets forth the claims of Sen and of Kaplow and Shavell in more detail. Section III.A defends a resolution of the conflict between the principle of liberal rights and the Pareto principle that, I argue, is faithful to both principles, contrary to the claims of Sen. Indeed, I argue that a solution based on alienable rights flows naturally from familiar liberal principles.

Section III.B then extends this solution to notions of fairness generally, arguing that a plausible theory of fairness can conform to the Pareto principle, contrary to the claims of Kaplow and Shavell. I argue that any plausible fairness theory includes multiple principles and that any such pluralistic theory must specify rules for when some principles take priority over others. There is nothing to prevent a pluralistic fairness theory from including the Pareto principle among its principles and giving the Pareto principle priority over other principles. Such a theory would never violate

17. This analysis should be of interest even to those who do not believe that the Pareto principle should apply to all social choices, as long as we can agree that there are at least some choices to which we should apply the Pareto principle. This Article reveals that we can apply fairness principles to those matters to which the Pareto principle is also applicable; that is, a belief in the Pareto principle with respect to those questions does not commit us to a welfarist theory regarding those issues.

the Pareto principle but would still apply fairness principles when doing so would not create a conflict with the Pareto principle. This fairness theory would not stand in opposition to human welfare, because it would never make everyone worse off. Fairness principles would only come into play when they protect the interests of some person willing to invoke those principles when preferences conflict. To invoke fairness in these contexts does not reflect any lack of concern with human welfare. Rather, under such a theory, fairness notions would go to the question of how we trade off the welfare of some against the welfare of others. Fairness notions would help determine the appropriate distribution of benefits and burdens in society.

Part IV sets forth and evaluates two examples of social welfare functions that incorporate notions of fairness while remaining faithful to the Pareto principle. I explore the features of these social welfare functions, which retain the most important virtues of utilitarian social welfare functions, such as transitivity and a complete ranking of alternatives, but do not neglect fairness concerns. They also feature some drawbacks, but I suggest that the advantages of such a solution would outweigh these disadvantages. In particular, the proposed social welfare functions violate certain conditions that social choice theorists sometimes impose on such functions. I argue that these conditions are unduly restrictive because they rule out plausible moral theories, including fairness theories that feature alienable rights. Finally, Part V summarizes my conclusions.

I. THE PROBLEM OF OBJECTIONABLE PREFERENCES

Critics of classical utilitarianism have relied on the existence of illiberal or antisocial preferences to generate some of their most cogent objections to utilitarianism. For example, Bernard Williams poses a hypothetical that features a minority group despised by an intolerant majority of citizens, who have such prejudices against this group that they find "even the knowledge of its presence" to be "very disagreeable."[18] If some propose to remove this minority, "a utilitarian calculation might well end up favouring this step, especially if the minority were a rather small minority and the majority were very severely prejudiced."[19] This implication of utilitarianism runs counter to ordinary moral intuitions, which would not consider the "benefits" of such a step to be a justification, regardless of the relative sizes of the persecuted minority and the intolerant majority and of the intensity of the prejudice of the majority. Thus, Williams asks "whether

18. Bernard Williams, *A Critique of Utilitarianism*, *in* SMART & WILLIAMS, *supra* note 14, at 75, 105.

19. *Id.*

the unpleasant experiences of the prejudiced people should be allowed . . . to count."[20]

Similarly, Sen criticizes any version of welfarism, which deems information regarding "different *sources* of utility and the motivation underlying it" to be morally irrelevant.[21] Sen argues that this "uncompromising rejection of the relevance of non-utility information" makes welfarism "a very limiting approach."[22] To illustrate the point, Sen asks us to compare two social states, *y* and *z*, in a hypothetical society with only two individuals.[23] In each of these states, person 1 has gained at the expense of person 2, relative to a third state *x*, producing a net gain in total utility in either state *y* or *z* over state *x*. Each individual enjoys the same level of utility in state *y* as in state *z*. In state *y*, person 1 enjoys higher utility because an egalitarian government has redistributed some of the wealth enjoyed by person 2 in state *x* to person 1, who is poorer than person 2. In state *z*, however, person 1 enjoys higher utility because the government allows him to torture person 2, and person 1 derives sadistic pleasure from the suffering of person 2. Under welfarism, we must ignore the nonutility differences between states *y* and *z*, including the malicious or sadistic nature of the preferences satisfied in state *z*. Thus, we draw a moral distinction between states *y* and *z* only by rejecting welfarism. For Sen, welfarism's claim that we must "attach the same weight to a person's enjoyment of food or shelter or personal liberty as to his joy from torturing is surely subject to critical questioning."[24]

A. *The Utilitarian Response*

R.M. Hare rises to the defense of utilitarianism by questioning the relevance of "fantastic cases" that critics of utilitarianism use as a "trick" to make the utilitarian "look like a moral monster."[25] He concedes that under utilitarianism, "if the Nazi's desire not to have Jews around is intense enough to outweigh all the sufferings caused to Jews by arranging not to have them around, then . . . it ought to be satisfied,"[26] but he argues that such hypotheticals are too fanciful to provide a legitimate test for a moral

20. *Id.*

21. Sen, *supra* note 6, at 478.

22. Sen, *supra* note 10, at 548. As Robert Goodin has noted, Sen's use of the term "non-utility information" here may be misleading, because "information about *why* individuals want what they want" is "still very much information about individual utilities." Robert E. Goodin, *Laundering Preferences*, *in* FOUNDATIONS OF SOCIAL CHOICE THEORY 75, 76-77 (Jon Elster & Aanund Hylland eds., 1986). Nevertheless, I use this term as Sen uses it.

23. Sen, *supra* note 10, at 547-48.

24. Sen, *supra* note 6, at 476.

25. R.M. Hare, *Ethical Theory and Utilitarianism*, *in* UTILITARIANISM AND BEYOND 23, 31 (Amartya Sen & Bernard Williams eds., 1982).

26. *Id.* at 30.

theory. After all, Hare responds, it is "obvious" that "no *actual* Nazis had such intense desires."[27]

It is not "obvious," however, that the hypothetical requires a "fantastic" intensity of desire among the Nazis. We can simply increase the number of Nazis and reduce the number of Jews in the hypothetical until the benefits of "ethnic cleansing" (or perhaps even genocide) exceed the costs. Under utilitarianism, for any given intensity of satisfaction for each Nazi and any given amount of suffering for each Jew, there must be some ratio of Nazis to Jews that would be large enough to justify the policy in question.[28] Hare apparently believes that the necessary ratio would border on "fantasy,"[29] but given the intensity of violent ethnic hatreds we observe in the world, it is not "obvious" that such a ratio is necessarily "fantastic," especially if we assume a very small number of victims.

Suppose Hare is correct, however, that we are at least unlikely to encounter such cases in real life, and "cases that are never likely to be actually encountered do not have to be squared with the thinking of the ordinary man, whose principles are not designed to cope with such cases."[30] His principles "are for use in practical moral thinking, especially under conditions of stress," not for use in "highly unusual situations," and "have to be general enough to be impartable by education."[31] These ordinary principles do not necessarily correspond to those that we would arrive at "by leisured moral thought in completely adequate knowledge of the facts, as the right answer in a specific case."[32] Thus, Hare argues, our ordinary

27. *Id.* Similarly, Hare observes that "we are most *un*likely, even if we give sadistic desires weight in accordance with their intensity, to encounter a case in which utility will be maximised by letting the sadist have his way." *Id.* Kaplow and Shavell give a similar response to the problem of objectionable preferences. KAPLOW & SHAVELL, *supra* note 15, at 416 (suggesting that "the gain to the sadist will be exceeded by the aggregate utility loss to others").

28. John C. Harsanyi, *Problems with Act-Utilitarianism and with Malevolent Preferences*, *in* HARE AND CRITICS 89, 96 (Douglas Seanor & N. Fotion eds., 1988) (noting that if the ratio is sufficiently large, then "we will have to conclude that the social-utility maximizing policy will be to kill all Jews").

29. R.M. Hare, *Comments*, *in* HARE AND CRITICS, *supra* note 28, at 199, 245. Hare also suggests that "in all actual cases" it is better "to push our institutions in the direction of the abandonment of harmful pleasures and desires, and hope that those who now indulge in them will soon change their ways." *Id.* at 246. Similarly, Kaplow and Shavell argue that "adopting policies that are contrary to the current preferences of some could produce a *change* in their preferences, and, over the long run, social welfare may rise as a result." KAPLOW & SHAVELL, *supra* note 15, at 416. Making the immorality of genocide contingent on the mutability of the intolerant preferences of the majority, however, seems no more plausible than making it contingent on the ratio of proponents to victims. Even if we were persuaded that these preferences are immutable, such that liberal policies would not change the intolerant preferences of the majority, we would still be inclined to oppose genocidal policies that would satisfy these preferences.

30. Hare, *supra* note 25, at 30.

31. *Id.* at 31.

32. *Id.* Similarly, J.J.C. Smart defends the utilitarian position that "there are no pleasures which are intrinsically bad," including sadistic pleasures. Smart, *supra* note 14, at 26. Smart explains that "[o]ur repugnance to the sadist arises, naturally enough, because in our universe

moral intuitions may lead us astray in "highly unusual" cases. There are two problems with Hare's response.

First, while Hare might persuade us that our ordinary moral intuitions may be mistaken in "highly unusual" cases, he does not demonstrate that in fact they are mistaken in the particular hypotheticals in question. Thus, if after "leisured moral thought," we still find that we cannot attach the same urgency to utilities arising from different sources, this intuition need not be put down as the result of the inculcation of simple principles during our "moral education."[33] That is, we still must address the question: What answer does "leisured moral thought" produce? For example, does Hare's premise that we should give "equal weight, impartially, to the equal interests of everybody"[34] necessarily imply that we must be sensitive only to the intensity of an individual's preferences and blind to the nature of those preferences? Why not take into account not only the level but also the source of utility for each individual?

Second, we can also produce less unusual cases that cannot be so easily dismissed. For example, consider the following questions. Should the satisfaction of racist preferences count as a legitimate reason to adopt public policies that discriminate or segregate on the basis of race?[35] Should the satisfaction of xenophobic preferences count as a legitimate reason to adopt immigration restrictions, to raise trade barriers, or to use military force abroad against foreign nationals?[36] Should the satisfaction of homophobic preferences count as a legitimate reason to adopt laws that discriminate against homosexuals? It does not seem unrealistic to imagine marginal cases in which a decision to count the satisfaction of intolerant preferences as a social benefit might tip the balance in favor of illiberal policies. When we consider whether to adopt these policies, must we indiscriminately include *all* types of satisfaction in the calculation of social welfare?

sadists invariably do harm." *Id.* at 25. Thus, our distaste for sadism merely reflects "our distaste for the consequences of sadism," which the utilitarian has already taken into account. *Id.* at 26.

33. Sen, *supra* note 6, at 476.

34. Hare, *supra* note 25, at 26.

35. In *Palmore v. Sidoti*, 466 U.S. 429 (1984), the Supreme Court held that a state cannot deny a divorced mother custody of her child on account of her interracial remarriage. The Court rejected the legal relevance of racial prejudice in society: "Private biases may be outside the reach of the law, but the law cannot, directly or indirectly, give them effect." *Id.* at 433; *see* Loving v. Virginia, 388 U.S. 1 (1967) (holding that the Fourteenth Amendment prohibits antimiscegenation laws); Shelley v. Kraemer, 334 U.S. 1 (1948) (holding that the Fourteenth Amendment prohibits states from enforcing racially restrictive covenants); Buchanan v. Warley, 245 U.S. 60 (1917) (holding that the Fourteenth Amendment prohibits laws that forbid blacks to reside in white neighborhoods).

36. *See* Howard F. Chang, *Liberalized Immigration as Free Trade: Economic Welfare and the Optimal Immigration Policy*, 145 U. PA. L. REV. 1147, 1210-21 (1997) (arguing that racist and xenophobic preferences should not count as reasons to adopt immigration restrictions).

B. *Laundering Preferences*

Robert Goodin proposes that we respond to the problem of objectionable preferences by using "input filters," which "work by refusing to count certain classes of desires and preferences when aggregating individual utilities."[37] This response raises the question of which preferences we should disregard and the related question of how we justify this "laundering of preferences."[38] In particular, can we justify this "laundering" in terms consistent with a "want-regarding" morality that still calculates social welfare by "respecting preferences" in some broader sense?[39]

1. *External Preferences*

Ronald Dworkin proposes a framework for distinguishing between two types of preferences: "[T]he preferences of an individual for the consequences of a particular policy may be seen to reflect . . . either a *personal* preference for his own enjoyment of some goods or opportunities, or an *external* preference for the assignment of goods and opportunities to others, or both."[40] He illustrates this distinction using an example of a state law school that excludes blacks: "A white law school candidate might have a personal preference for the consequences of segregation, for example, because the policy improves his own chances of success, or an external preference for those consequences because he has contempt for blacks and disapproves social situations in which the races mix."[41] Thus, external preferences may derive from ill will directed at members of particular racial groups or from racist political theories. External preferences may also be "moralistic," as when "members of the community disapprove on moral grounds of homosexuality, or contraception, or pornography, or expressions of adherence to the Communist party" and want to prohibit these activities as immoral.[42]

External preferences, however, need not be malicious, moral, or political. Suppose a white man derives no utility from the plight of blacks. Perhaps he even takes some pleasure from observing their success, but he

37. Goodin, *supra* note 22, at 78.

38. *Id.* at 81.

39. *Id.* at 77.

40. RONALD DWORKIN, TAKING RIGHTS SERIOUSLY 234 (1977). Dworkin does not, however, specify what counts as a "good." If we define "good" broadly enough to include the satisfaction of external preferences, then any external preference can also be described as a personal preference. Therefore, to maintain the distinction between external and personal preferences, Dworkin must have a narrower definition of "good" in mind, perhaps a definition restricted to the sensory or physical experiences of the holder of the preference.

41. *Id.* at 234-35.

42. *Id.* at 235, 275-76.

takes still more pleasure from the success of other whites, with whom he sympathizes or identifies more readily. These "altruistic" preferences favor some individuals over others in the allocation of scarce social resources or opportunities, and Dworkin would also count these as external preferences.[43]

Dworkin argues that "the only defensible form of utilitarianism" would "fix only on personal and ignore external preferences."[44] He explains that "the principal source of the great appeal that utilitarianism has had" is its egalitarian treatment of "the wishes of each member of the community on a par with the wishes of any other."[45] This treatment embodies the "right to equal concern and respect" that each individual enjoys under the "liberal conception of equality."[46] To count external preferences, Dworkin argues, would "be a form of double counting" inconsistent with the liberal egalitarianism supposedly embodied in utilitarianism:

> If a utilitarian argument counts external preferences along with personal preferences, then the egalitarian character of that argument is corrupted, because the chance that anyone's preferences have to succeed will then depend, not only on the demands that the personal preferences of others make on scarce resources, but on the respect or affection they have for him or for his way of life. If external preferences tip the balance, then the fact that a policy makes the community better off in a utilitarian sense would *not* provide a justification compatible with the right of those it disadvantages to be treated as equals.[47]

The victim of these external preferences will suffer on account of not only the personal preferences of those competing for scarce resources but also the external preferences of those who have no personal interest at stake.

In what sense, however, is the counting of external preferences a form of "double counting" inconsistent with equal concern and respect? Dworkin's characterization is question-begging: The counting of external preferences strikes us as "double counting" only if we have already

43. *Id.* at 235.

44. *Id.* at 276. To give a satisfactory result in Hare's genocide hypothetical, a liberal theory of social utility would exclude the satisfaction of external preferences entirely rather than merely discount this satisfaction by some factor. Otherwise, there would still exist some ratio of Nazis to Jews that would imply that genocide would increase social utility. *See supra* note 28 and accompanying text.

45. DWORKIN, *supra* note 40, at 275.

46. *Id.* at 273.

47. *Id.* at 235. Bruce Ackerman takes a similar concept of neutrality to be fundamental in the liberal state. This neutrality principle precludes anyone from justifying a legal regime by claiming that "his conception of the good is better than that asserted by any of his fellow citizens" or that "he is intrinsically superior to one or more of his fellow citizens." BRUCE A. ACKERMAN, SOCIAL JUSTICE IN THE LIBERAL STATE 11 (1980).

decided on other grounds that they should not count.[48] Otherwise, as H.L.A. Hart suggests, it would seem that the exclusion of external preferences is a form of "undercounting" that denies equal concern and respect for those who hold those external preferences.[49] How are we to choose between these two different notions of what "equal concern and respect" would entail?

Although I ultimately adopt Dworkin's distinction between personal and external preferences, I offer a critique of the reasoning that Dworkin sometimes uses to justify the exclusion of certain preferences. In analyzing the rationales for the exclusion of some preferences, it is useful to distinguish more carefully between the various species of external preferences, because we might exclude each type of external preference on somewhat different grounds. I propose a modified version of Dworkin's framework that I argue is more faithful to the liberal ideals underlying utilitarianism.

I do not present this framework as a complete moral theory, specifying the appropriate treatment of every imaginable type of preference. My objectives here are more modest. I intend merely to suggest that the laundering of preferences is in general a plausible response to the problems raised by welfarism, and to outline how one might justify this laundering using liberal principles.

2. *Political and Moral Preferences*

The fundamental objection to external preferences goes to the substantive content of those preferences, which themselves deny the equal concern and respect that utilitarianism owes all individuals. Thus, Dworkin argues that the utilitarian must reject such preferences because by their very nature they contradict the egalitarian premise of utilitarianism:

> Suppose the community contains a Nazi, for example, whose set of preferences includes the preference that Aryans have more and Jews less of their preferences fulfilled just because of who they are. A neutral utilitarian cannot say that there is no reason in political morality for rejecting or dishonoring that preference, for not dismissing it as just wrong . . . . For utilitarianism itself supplies such a reason: its most fundamental tenet is that peoples' preferences should be weighed on an equal basis in the same scales,

48. Dworkin has explained that his claim of "double-counting" was "meant to summarize the argument, not to make it." RONALD DWORKIN, A MATTER OF PRINCIPLE 366 (1985).

49. H.L.A. Hart, *Between Utility and Rights*, 79 COLUM. L. REV. 828, 842 (1979); *see also* C. Edwin Baker, *Counting Preferences in Collective Choice Situations*, 25 UCLA L. REV. 381, 386 (1978) (suggesting that "the egalitarian nature of utilitarian analysis would seem to be corrupted not by counting all of people's preferences but by favoring those with only a certain type of preferences, *i.e.*, personal preferences" and that "ignoring [an] 'external preference' would appear to be 'half counting'").

> that the Nazi theory of justice is profoundly wrong, and that officials should oppose the Nazi theory and strive to defeat rather than fulfill it. A neutral utilitarian is barred, for reasons of consistency, from taking the same politically neutral attitude to the Nazi's political preference that he takes toward other sorts of preferences.[50]

Dworkin's argument here turns on the nature of the Nazi's preference as based on a political theory other than utilitarianism: "Political preferences, like the Nazi's preference, are on the same level—purport to occupy the same space—as the utilitarian theory itself."[51] Thus, the utilitarian cannot be neutral between Nazism and utilitarianism: "Utilitarianism must claim . . . truth for itself, and therefore must claim the falsity of any theory that contradicts it. It must itself occupy, that is, all the logical space that its content requires."[52] This logic sweeps quite broadly, because it requires the utilitarian to ignore all preferences that flow from political and moral theories that are not utilitarian, including many that appear to be egalitarian, such as the theory of justice proposed by John Rawls,[53] and are thus far more appealing than Nazism.

50. DWORKIN, *supra* note 48, at 362. In a similar spirit, Paul Brest declares that "decisions are irrational insofar as they reflect the assumption that members of one race are less worthy than other people." Paul Brest, *The Supreme Court, 1975 Term—Foreword: In Defense of the Antidiscrimination Principle*, 90 HARV. L. REV. 1, 6 (1976). This assumption would represent a defect in "the *process* by which race-dependent decisions are made" providing a rationale for the "antidiscrimination principle" that is distinct from "harmful *results* of race-dependent decisions." *Id.* Similarly, Larry Alexander classifies biases that "rest on erroneous judgments of others' inferior moral worth" as "intrinsically immoral" preferences. Larry Alexander, *What Makes Wrongful Discrimination Wrong? Biases, Preferences, Stereotypes, and Proxies*, 141 U. PA. L. REV. 149, 192 (1992). He distinguishes "intrinsically immoral" preferences from those that are "extrinsically—contingently—immoral because of the effects on others of acting on those preferences." *Id.* at 194. Thus, like Dworkin, both Brest and Alexander would reject such intrinsically immoral preferences as justifications for decisions even before we weigh the costs that the satisfaction of these preferences imposes upon others.

51. DWORKIN, *supra* note 48, at 363.

52. *Id.* at 361.

53. JOHN RAWLS, A THEORY OF JUSTICE (1971). Some have proposed that we exclude a narrower class of preferences. For example, Baker suggests that "taking into account" external preferences "is not always inappropriate" but that "*counting these preferences when they deny the inherent equality or worth of persons is improper*." Baker, *supra* note 49, at 403. Eric Rakowski considers the suggestion that we exclude "preferences attributable to the belief that some persons are not entitled to equal moral consideration" but still count "desires that are not antiegalitarian," but he concludes that this solution is "unavailing" because other external preferences are also problematic. ERIC RAKOWSKI, EQUAL JUSTICE 29 n.13 (1991). For example, Alexander notes that other "preferences based on moral ideals," such as the belief that heterosexuality is "the only morally proper form of sexuality," that "the role of sex partner is gender-specific as a matter of morality," or that "it is immoral for women to perform certain roles and for men to perform certain roles," need not entail a belief that "men and women have differential moral worth." Alexander, *supra* note 50, at 164. Nevertheless, Alexander considers these preferences to be "intrinsically immoral." *Id.* at 192.

It may be an overstatement, however, for Dworkin to suggest that it is a "contradiction" for a utilitarian to count political preferences.[54] After all, a utilitarian may deny the truth of rival political theories and may even strive to persuade others to give up these theories, while still regarding it to be the moral obligation of a utilitarian to count the political preferences of others, even as they persist in believing rival theories. Even as Kaplow and Shavell criticize notions of fairness, for example, they advocate a welfarist theory that credits the "tastes for fairness" that people may have, just as this theory would credit any other tastes.[55] Dworkin observes that if utilitarianism counts political preferences, "then it will be, from the standpoint of personal preferences, self-defeating," because the outcome "will then not be, from that standpoint, utilitarian at all."[56] This utilitarianism, however, is "self-defeating" only if it adopts "the standpoint of personal preferences." If a utilitarian does not adopt this standpoint, then in what sense is the outcome self-defeating?

The stance of welfarists like Kaplow and Shavell does not entail a logical contradiction, but it can put the welfarist in the awkward position of making some people worse off in terms of their personal preferences in order to satisfy the external preferences of others who believe in a political theory that the welfarist regards as false and strives to displace. This approach makes optimal policies depend on the prevalence and intensity of competing political beliefs, including those that might be unsound or prove illogical upon reflection. The utilitarian becomes, in part, a political pollster, asking what others believe to be true rather than identifying a truth that is independent of the political opinions of others.[57] This utilitarian would seek to implement the views of others that are baseless or ill-considered from a utilitarian perspective, even if this implementation satisfied their external preferences by sacrificing the personal interests of those who do not share these views, simply because these views are popular.

This stance may not be illogical, strictly speaking, but it certainly makes for a utilitarianism that is curiously vulnerable to capture by profoundly illiberal views. In this sense, Williams suggests that a utilitarian who counts political preferences would be "defeating his own view of things" when he credits a moral or political theory that is "from the

54. DWORKIN, *supra* note 48, at 361, 363.

55. Kaplow & Shavell, *supra* note 3, at 75.

56. DWORKIN, *supra* note 40, at 235.

57. Brian Barry suggests that such a utilitarian "would seem rather eccentric to take account of *my* judgment as an ingredient in *his* judgment about social welfare." Brian Barry, *Lady Chatterley's Lover and Doctor Fischer's Bomb Party: Liberalism, Pareto Optimality, and the Problem of Objectionable Preferences*, *in* FOUNDATIONS OF SOCIAL CHOICE THEORY, *supra* note 22, at 11, 37.

utilitarian point of view itself irrational."[58] Thus, Williams suggests that the utilitarian exclude such preferences "on the principle that no pains or discomforts are to count in the utilitarian sum which their subjects have just because they hold views which are by utilitarian standards irrational."[59] If we adopt the principle suggested by Williams and exclude all political and moral preferences as suggested by Dworkin, however, these exclusions seem to sweep both too narrowly in some dimensions and too broadly in others.

3. *Altruistic and Antisocial Preferences*

The exclusion of all political and moral preferences may still leave us with other preferences that we may wish to exclude. The discriminatory altruistic preferences that Dworkin classifies as external, for example, need not reflect any moral or political beliefs at all. Such preferences may simply flow from "our tendency to sympathize most readily with those who seem most like ourselves."[60] Sen, for example, distinguishes between preferences that derive from some effect on our own psychological welfare, such as those based on our "sympathy" for other people, and those that depend on our political or moral "commitments."[61] The individuals who feel sympathy for some and not for others need not be making any moral statement but simply experiencing pleasure when the objects of their sympathy experience pleasure. They may merely present their preferences as tastes, as entitled to satisfaction under utilitarianism as any other. These tastes would be no more political or moral than the tastes underlying the personal preferences that Dworkin would have the utilitarian satisfy.[62]

For example, John Harsanyi also distinguishes between "personal preferences" and "moral preferences" and uses only personal preferences to arrive at the "social utility" he seeks to maximize under his utilitarian theory, but he includes altruistic preferences in the concept of "personal preferences."[63] Most individuals' personal preferences, he explains, "will not be completely selfish, but they will assign higher weights to their own interests and to the interests of their family, their friends, and other personal associates than they will assign to the interests of complete strangers."[64]

58. Williams, *supra* note 18, at 106.

59. *Id.*

60. Brest, *supra* note 50, at 8.

61. Amartya K. Sen, *Rational Fools: A Critique of the Behavioral Foundations of Economic Theory*, 6 PHIL. & PUB. AFF. 317, 326-29 (1977). Williams uses the term "commitments" in a similar sense. Williams, *supra* note 18, at 112-13.

62. The preference for public policies that satisfy one's own altruistic preferences might be deemed political, but then personal preferences would be political in precisely the same sense.

63. John C. Harsanyi, *Morality and the Theory of Rational Behaviour*, *in* UTILITARIANISM AND BEYOND, *supra* note 25, at 39, 47, 54.

64. *Id.* at 47.

These preferences fail to extend equal concern and respect to all, however, and Dworkin would exclude them as external preferences from the utilitarian calculation of social welfare.[65]

Thus, to avoid the sort of inconsistency that Dworkin and Williams identify, the utilitarian would have to exclude not only preferences derived from moral or political views, but also any altruistic preference that discriminates among other individuals.[66] This exclusion would prevent sympathies that reflect nepotism or tribalism from influencing public policy. The basis for excluding these preferences, however, cannot be that they occupy the same "logical space" or are "on the same level" as utilitarianism, because they are not political preferences. They do not conflict with utilitarianism, because they do not assert the truth of any rival moral or political theory.

Instead, we would exclude such preferences because they are, in a more general sense, inconsistent with the perspective of the utilitarian. These preferences are, by their very nature, incompatible with the egalitarian premises of utilitarianism, so that counting such preferences would allow partiality to infect the outcome.[67] If the motivation underlying utilitarianism is, as Hare puts it, to give "equal weight, impartially, to the equal interests of everybody," then perhaps an "ideal observer" committed to this egalitarian ideal would refuse to consider any interest based on unequal concern for others.[68] Such an interest would not be the "equal" of any other personal preference, because like political or moral preferences, this interest differs in kind from purely self-regarding preferences. Thus, the exclusion

65. More recently, Harsanyi has endorsed Dworkin's claim that "the very nature of utilitarian ethics suggests the exclusion of all external preferences," including altruistic preferences. Harsanyi, *supra* note 28, at 98.

66. Thus, Yew-Kwang Ng argues that in order to avoid "double or rather multiple counting," we should ignore "concern for the welfare of others" in "social evaluation." Yew-Kwang Ng, *Utility, Informed Preference, or Happiness: Following Harsanyi's Argument to Its Logical Conclusion*, 16 SOC. CHOICE & WELFARE 197, 199 (1999). Other economists have argued that cost-benefit analysis should not include ethical or altruistic values. *E.g.*, Paul Milgrom, *Is Sympathy an Economic Value? Philosophy, Economics, and the Contingent Valuation Method*, *in* CONTINGENT VALUATION: A CRITICAL ASSESSMENT 417, 418-22 (Jerry A. Hausman ed., 1993).

67. Brest, for example, concludes that decisions that reflect "racially selective sympathy and indifference," that is, "the unconscious failure to extend to a minority the same recognition of humanity, and hence the same sympathy and care, given as a matter of course to one's own group," are "unfair," like "those reflecting overt racial hostility." Brest, *supra* note 50, at 7-8. Brest explains that "such treatment violates the cardinal rule of fairness—the Golden Rule." *Id.* at 8. We can distinguish these *inherently* discriminatory preferences from purely self-regarding preferences, which may advantage some individuals over others merely as an incidental consequence of satisfying these preferences in a market economy. For example, "a preference for music advantages individuals born with musical talent," and "a preference for reading novels advantages individuals born with literary talent." KAPLOW & SHAVELL, *supra* note 15, at 414 n.865 (noting that "virtually any taste may have third-party effects that result in relative advantages and disadvantages to other people"). We can give weight to these purely self-regarding preferences, which are not "intrinsically immoral." Alexander, *supra* note 50, at 194.

68. Hare, *supra* note 25, at 26.

of these external preferences would be faithful to the ideal underlying utilitarianism.[69]

Using similar reasoning, Harsanyi endorses the exclusion of other preferences from his definition of social welfare. He would exclude "antisocial preferences, such as sadism, envy, resentment, and malice" from his concept of "social utility."[70] Like altruistic preferences, antisocial preferences are other-regarding preferences that Harsanyi classifies as personal preferences rather than moral preferences, but which Dworkin would classify as external preferences rather than personal preferences.[71] Unlike altruistic preferences, which reflect the utility some derive from the happiness of others, these antisocial preferences reflect the disutility some derive from the happiness of others, or the utility derived from the unhappiness of others.

69. Rakowski has suggested that a utilitarianism that excludes external preferences "forsakes its motivating idea." RAKOWSKI, *supra* note 53, at 26. The truth of this claim depends on what one takes to be the "motivating idea" of utilitarianism. The claim would be correct if this idea is that "whatever moral principles would be chosen by self-interested, risk-neutral individuals ignorant of their own desires and abilities are justified." *Id.* at 24; *see* John C. Harsanyi, *Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking*, 61 J. POL. ECON. 434 (1953) (deriving utilitarianism from such a premise). Kaplow and Shavell similarly "consider contexts in which all individuals are symmetrically situated before events arise that call for the application of legal rules" to justify welfarism. Kaplow & Shavell, *supra* note 3, at 68. This idea, however, "begs the question" of whether we should consider this particular hypothetical to be the appropriate test for a moral theory. Barbara Fried, Can We Really Deduce Utilitarianism from the Pareto Principle? 41 (Dec. 6, 1999) (unpublished manuscript, on file with *The Yale Law Journal*). We can reject this idea as the basis for moral principles and substitute a different motivating idea. If the motivating idea is instead an ideal observer committed to equal concern and respect for all persons, then the exclusion of external preferences would be a means to implement, not abandon, that idea.

70. Harsanyi, *supra* note 63, at 56; *see* RICHARD A. POSNER, THE ECONOMICS OF JUSTICE 83 (1981) (criticizing utilitarians who "ascribe value" to "envy and cruelty, because these are common sources of personal satisfaction and hence of utility"). Shavell also suggests in his earlier work that "illicit utility (such as would arise when a person batters another for pleasure)" may not enter "social welfare." STEVEN SHAVELL, ECONOMIC ANALYSIS OF ACCIDENT LAW 160-61 (1987). Other economists make similar suggestions. *E.g.*, Robert D. Cooter, *Economic Analysis of Punitive Damages*, 56 S. CAL. L. REV. 79, 87, app. at 99 (1982) (suggesting that an economic analysis may exclude an "illicit benefit" that an injurer enjoys from wrongful behavior); Alvin K. Klevorick, *On the Economic Theory of Crime*, *in* NOMOS XXVII: CRIMINAL JUSTICE 289, 299 (1985) (suggesting that an economic analysis of the social value of criminal acts may discount "the criminal's gain" from the crime); George J. Stigler, *The Optimum Enforcement of Laws*, 78 J. POL. ECON. 526, 527 (1970) (suggesting that when "society has branded the utility derived from [crimes] as illicit," the "gain to the offender" is not "a gain to society").

71. Like Dworkin, others have suggested the exclusion of all other-regarding preferences from consideration. *E.g.*, Jean Hampton, *Rational Choice and the Law*, 15 HARV. J.L. & PUB. POL'Y 649, 671 (1992) (suggesting that to avoid "unfairness," the parties to an ideal social contract would consider only "self-concerned" preferences and exclude "other-regarding interests that could be either malicious (thus skewing the contract 'unfairly' against those who were hated) or beneficent (thus skewing the contract 'unfairly' in favor of those who where loved)"); Peter J. Hammond, The Economics of Justice *and the Criterion of Wealth Maximization*, 91 YALE L.J. 1493, 1499, 1501 (1982) (reviewing POSNER, *supra* note 70) (arguing that counting antisocial preferences is "ethically unacceptable" and suggesting that only "good" and "self-interested" preferences should count, where "good" is defined using "ethical criteria" to determine "moral worthiness").

How does Harsanyi justify this exclusion? Like Dworkin, Harsanyi appeals to the ideals underlying utilitarianism: "According to utilitarian theory, the fundamental basis of all our moral commitments to other people is a general goodwill and human sympathy."[72] A preference that is itself incompatible with this general benevolence deserves no weight:

> [N]o amount of goodwill to individual *X* can impose the moral obligation on me to help him in hurting a third person, individual *Y*, out of sheer sadism, ill will, or malice. Utilitarian ethics makes all of us members of the same moral community. A person displaying ill will toward others does remain a remember [*sic*] of this community, but not with his whole personality. That part of his personality that harbours these hostile antisocial feelings must be excluded from membership, and has no claim for a hearing when it comes to defining our concept of social utility.[73]

Thus, only a preference that is consistent with the principle of universal beneficence can give rise to a moral claim that obliges others to satisfy it.

In this sense, Goodin suggests that we justify the laundering of preferences using "reasons which are located in the logic of the social decision process."[74] From this perspective, the exclusion of preferences follows from this logic, because "our very choice of aggregating preferences as a way of making social decisions carries consequences for the kind of preferences that we can count."[75] This type of reasoning allows us to exclude external preferences while remaining "within the want-regarding framework," which takes the satisfaction of preferences to be the basis for social choices.[76]

4. *Personal Preferences Based on External Preferences*

On the other hand, Dworkin seems to go too far in excluding all moral, political, and altruistic preferences, for he would exclude even personal preferences that derive from moral, political, or altruistic preferences. Consider, for example, the preference of an ascetic who pursues austere consumption patterns so as to conserve social resources for the use of others. Dworkin argues that a preference "for less of a certain good on the assumption, or rather proviso, that other people will have more" would be

72. Harsanyi, *supra* note 63, at 56.

73. *Id.* Thus, Harsanyi endorses the use of a "censored utility function" that excludes antisocial preferences, because an ideal observer would have no "moral sympathy" for such preferences. JOHN C. HARSANYI, RATIONAL BEHAVIOR AND BARGAINING EQUILIBRIUM IN GAMES AND SOCIAL SITUATIONS 62 (1977).

74. Goodin, *supra* note 22, at 85.

75. *Id.*

76. *Id.* at 86.

"parasitic upon external preferences, in the shape of political and moral theories, and . . . may no more be counted in a defensible utilitarian argument than less attractive preferences."[77] This suggestion implies that a utilitarian social planner in theory should try to induce people holding this preference to consume more of the good in question, in spite of their preferences not to do so. From this standpoint, their consumption patterns would be "distorted" by their moral beliefs, and the social planner would seek to "correct" this distortion, so as to eliminate the "deadweight loss" flowing from their moral beliefs.

This "moral distortion" in an individual's behavior would present the social planner with a prima facie reason to change the individual's behavior. That is, the planner would deem any reduction in this "distortion" to be a social benefit militating in favor of any policy that produces such a reduction. For example, Dworkin's suggestion implies that if environmentalists derive satisfaction from devoting their time and effort to recycling activities, due to their moral or political convictions, a utilitarian should ignore this benefit that the environmentalists perceive as the result of their labors and implement policies to reduce their recycling efforts.

In fact, Dworkin's suggestion supports the use of coercion to correct these supposed distortions in people's activities if the application of coercive policies would eliminate these distortions at a sufficiently low cost. Although we can presume that Dworkin does not intend to endorse such policies, the notion of social welfare that he suggests in theory would justify such policies.[78] These attempts to correct "moral distortions" in people's behavior, however, would violate the liberal view that "decisions about what people value should be left up to them."[79]

Harsanyi suggests that an attractive feature of a utilitarianism based on personal preferences is its fidelity to "the important philosophical principle of *preference autonomy*," which holds that "in deciding what is good and what is bad for a given individual, the ultimate criterion can only be his own wants and his own preferences."[80] Similarly, Rawls argues that in a

77. DWORKIN, *supra* note 40, at 277.

78. Kaplow and Shavell criticize Dworkin's theory using similar examples, pointing to "the difference between an opera singer who performs for the money and one who would not be induced to perform by the money alone but does choose to perform because of the pleasure of pleasing an audience" and noting that "under the posited theory, the latter singer should not be permitted to perform." KAPLOW & SHAVELL, *supra* note 15, at 414 n.864.

79. W. Michael Hanemann, *Valuing the Environment Through Contingent Valuation*, J. ECON. PERSP., Fall 1994, at 19, 33 ("When estimating demand functions for fish prior to Vatican II, no economist ever proposed removing Catholics because they were eating fish out of a sense of duty.").

80. Harsanyi, *supra* note 63, at 55. Harsanyi also calls this idea "[t]he principle of individual self-determination," HARSANYI, *supra* note 73, at 61, or "the *principle of acceptance*, because it requires us to accept each individual's own personal preferences as the basic criterion for assessing the utility . . . that he will derive from any given situation," *id.* at 52.

"well-ordered society, . . . persons are left free to determine their good."[81] We find the classic expression of this liberal principle in John Stuart Mill's *On Liberty*:

> [T]he sole end for which mankind are warranted, individually or collectively, in interfering with the liberty of action of any of their number is self-protection. That the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others. His own good, either physical or moral, is not a sufficient warrant. He cannot rightfully be compelled to do or forbear because it will be better for him to do so, because it will make him happier, because, in the opinions of others, to do so would be wise, or even right. . . . The only part of the conduct of any one, for which he is amenable to society, is that which concerns others. In the part which merely concerns himself, his independence is, of right, absolute.[82]

Harsanyi, however, does not endorse this liberal principle in the absolute form expressed by Mill. Instead, Harsanyi allows for the possibility of paternalistic intervention to promote a person's own good as long as we "use his own preferences in some suitable way as our final criterion in judging what his real interests are and what is really good for him."[83] Harsanyi explains that "a person may irrationally want something which is very 'bad for him'" when "his own preferences at some deeper level are inconsistent with what he is now trying to achieve."[84] Thus, Harsanyi distinguishes between a person's "manifest preferences" and his "true preferences."[85] Manifest preferences are "actual preferences as manifested by his observed behaviour, including preferences possibly based on erroneous factual beliefs, or on careless logical analysis, or on strong emotions that at the moment greatly hinder rational choice," whereas true preferences are "the preferences he *would* have if he had all the relevant factual information, always reasoned with the greatest possible care, and were in a state of mind most conducive to rational choice."[86] Harsanyi argues that "social utility must be defined in terms of people's true preferences rather than in terms of their manifest preferences."[87] An analysis of the merits and proper scope of this qualification of the principle

81. RAWLS, *supra* note 53, at 448.
82. JOHN STUART MILL, ON LIBERTY 6 (Longmans, Green & Co. 1921) (1859).
83. Harsanyi, *supra* note 63, at 56.
84. *Id.* at 55.
85. *Id.*
86. *Id.*
87. *Id.* For example, Harsanyi suggests that we are justified in using a "corrected utility function" for an individual if we think that the individual "would *approve* of this" correction if "made aware of the possibility that his actual preferences were based on factual mistakes and of the likely costs of these mistakes to him." HARSANYI, *supra* note 73, at 61-62.

of preference autonomy is beyond the scope of this Article. To maintain a focus on other issues, I assume that the preferences under discussion are "true" preferences in the sense described by Harsanyi, not "manifest" preferences based on mistaken beliefs regarding one's "true" preferences.[88]

Subject to this proviso, the principle of preference autonomy prohibits intervention based on our disapproval of the choices others make between satisfaction of their purely personal preferences and satisfaction of their moral, political, and altruistic preferences. Thus, under "classical liberal doctrine," it is "when assessing the conduct of *others*" that we agree "not to take into account certain feelings," such as "an abhorrence for certain religious or sexual practices," which then "have no weight" from the standpoint of justice.[89] To take these feelings into account in one's own personal preferences does not violate the liberal principle of equal concern and respect for others. To impose these moral views on others against their will, however, would violate that principle.

Thus, for example, if some prefer for moral reasons not to encounter pornography, then this preference would be a legitimate reason to restrict the public display of pornography, as long as these reasons "emphasize not how others should lead their lives, but rather the character of the sexual experience people want for themselves."[90] To the extent that we count such moral preferences only as they affect personal preferences and not as external preferences, we can remain faithful to the liberal principle of equal concern and respect. Thus, to uphold the liberal ideals underlying utilitarianism, we must exclude external preferences, but we need not exclude all personal preferences that derive from moral or political views. It is only when moral and political preferences intrude into the external realm that utilitarianism risks undermining its own ideals by seeking to satisfy these preferences.

88. Similarly, Kaplow and Shavell assume that "individuals understand fully how various situations affect their well-being," so that their arguments apply to "individuals' actual well-being rather than to individuals' well-being as reflected in their mistaken preferences." Kaplow & Shavell, *supra* note 3, at 65 & n.2. I discuss the implications of the distinction between manifest and true preferences when this issue becomes relevant. *Infra* note 153.

89. RAWLS, *supra* note 53, at 450. Similarly, Richard Epstein argues that "no one person is entitled to elevate his or her beliefs about how *others* should act above those of anyone else." RICHARD A. EPSTEIN, FORBIDDEN GROUNDS: THE CASE AGAINST EMPLOYMENT DISCRIMINATION LAWS 415 (1992) (emphasis added). Yet he otherwise seems to favor a calculation of costs and benefits that does not require us "to ask whose preferences are legitimate and whose are not." *Id.* at 43.

90. DWORKIN, *supra* note 48, at 364-65. Dworkin would allow such preferences to justify regulations on public displays of pornography, but it seems he does so only because it is impossible to separate those preferences "that express moral condemnation or would not exist but for it" from "the remaining strands" of personal "taste[s]." *Id.* at 356.

C. *Liberal Consequentialism*

Once we exclude external preferences, we are left with a form of utilitarianism that resists corruption by illiberal preferences and remains faithful to the motivating ideals that give this philosophy much of its appeal. Dworkin's liberal principle of "equal concern and respect" embodies these ideals, which include the principles of universal benevolence and of preference autonomy as described by Harsanyi and Mill. Thus, we can use these liberal principles as the basis for a theory of preference laundering.

Our theory is, however, no longer "utilitarian" or even "welfarist" as Sen uses these terms, because it takes the *source* of utility, and not just its quantity, to be morally relevant. Our philosophy remains a form of "consequentialism," because it claims that policies "are to be chosen on the basis of the states of affairs which are their consequences."[91] It is also "teleological," insofar as it defines what makes states of the world "good," then seeks to maximize "the good."[92] We have defined the good as a particular notion of social welfare, one that uses nonutility information about preferences as well as information about each individual's level of utility.

Under this theory, liberal rights, such as the right to read books of which others disapprove, are derived from our theory of the good.[93] We design these rights as institutions that are likely to maximize our chosen definition of social welfare. For example, Dworkin defends such rights as a response to the problem of excluding external preferences in a democracy that "cannot discriminate, within the overall preferences imperfectly revealed by voting, distinct personal and external components."[94] We may invoke "rights to certain liberties like the liberty of free expression and of free choice in personal and sexual relations" as trumps over the preferences of a political majority, because this arrangement enables us "to enjoy the institutions of political democracy, which enforce overall or unrefined utilitarianism, and yet protect the fundamental right of citizens to equal concern and respect by prohibiting decisions that seem, antecedently, likely to have been reached by virtue of the external components of the preferences democracy reveals."[95]

91. *Introduction* to UTILITARIANISM AND BEYOND, *supra* note 25, at 1, 3-4 (emphasis omitted).

92. RAWLS, *supra* note 53, at 24 (defining a teleological theory as a theory that holds that "those institutions and acts are right which of the available alternatives produce the most good").

93. Other theories, like the theory of justice advanced by Rawls, are "deontological" theories, which do not design rights to maximize some good specified in advance. *Id.* at 30.

94. DWORKIN, *supra* note 40, at 276.

95. *Id.* at 277.

This liberal theory of social welfare provides a firmer foundation for such rights than does the welfarism of Kaplow and Shavell, who suggest that "rights of individuals against the government" may "be embodied in rules in order to constrain the behavior of agents who cannot be trusted to use their discretion to maximize social welfare."[96] For Kaplow and Shavell, the satisfaction of external preferences would count as part of social welfare.[97] From a liberal perspective, even government agents who faithfully maximize that notion of social welfare are a source of concern. Thus, liberal rights are valuable precisely because they are effective against government intrusions that *do* maximize Kaplow and Shavell's concept of social welfare.

II. THE CONFLICT WITH THE PARETO PRINCIPLE

The liberal consequentialism that we have outlined introduces "fairness," as Kaplow and Shavell use the term, into the analysis. Kaplow and Shavell define "fairness" to include any "method of policy assessment" that gives any weight to "characteristics of the situation that will prevail under a legal regime" other than their effects on each individual's level of overall utility.[98] Thus, "fairness" includes "principles that are based on factors that matter to individuals but under which the factors are weighed differently from the manner in which individuals themselves weigh them."[99] A liberal measure of social welfare that gave no weight to external preferences, for example, would give less weight to such preferences than the individuals themselves would give them. This liberal measure of social welfare would thereby introduce some nonutility information (that is, the "external" nature of the preference) as a morally relevant consideration. Kaplow and Shavell make the strong claim that "any conceivable notion of social welfare" that is a function of anything but each individual's utility level is a "social welfare function" that "will sometimes require adoption of a policy that makes every person worse off," in violation of the Pareto principle.[100]

A. *Sen's Example*

Would the liberal consequentialism sketched in Part I violate the weak Pareto principle? We can show how it would using a simple example made

96. KAPLOW & SHAVELL, *supra* note 4, at 5.

97. KAPLOW & SHAVELL, *supra* note 15, at 411-17 (arguing in favor of including all preferences, including objectionable preferences, in the calculation of social welfare).

98. Kaplow & Shavell, *supra* note 3, at 66 n.5.

99. *Id.* at 65 n.4.

100. KAPLOW & SHAVELL, *supra* note 4, at 1.

famous by Sen.[101] Imagine a society with two individuals: The first is a prude (*P*), and the second is lewd (*L*). There is one copy of a book, say *Lady Chatterley's Lover*, that each may read. The prude, *P*, has a personal preference not to read the book, while lewd *L*'s personal preference is to read the book. If we count only personal preferences, then the social optimum is for *L* to read the book and *P* not to read the book. Let *l* denote this state of affairs.

Suppose, however, that each person also has external preferences. The prude, *P*, would prefer that *L* not read the book, because the thought of *L* enjoying the book distresses *P*, and *L* prefers that *P* read the book, because he would enjoy the thought of *P*'s discomfort and shock in reading the book. In fact, each person's external preference is so intense that each would rather satisfy this external preference than his own personal preference. Therefore, each would prefer an alternative in which *P* read the book and *L* did not rather than the supposed social optimum. Let *p* denote this alternative, which each person would prefer over *l*. Although *l* ranks higher than *p* if we count only personal preferences, *l* is also "Pareto inferior" to *p* and thus not Pareto optimal.[102] Thus, a social welfare function that counts only personal preferences would violate even the weak Pareto principle.[103]

Sen uses a different notion of liberalism and thus states the problem in a somewhat different form. He assumes that we believe in "[l]ibertarian values," which require that society respect each person's desires in that person's "personal domain" or "protected sphere."[104] Thus, in Sen's example, each person enjoys a libertarian right to decide whether to read the book in question. Sen asks whether we can respect both these libertarian rights and the Pareto principle over an "[u]nrestricted [d]omain," that is, if we allow for every "logically possible" set of preferences that individuals could have.[105] He shows that if we grant individuals libertarian rights, then the result conflicts with even the weak Pareto principle.

Consider three possible alternatives in Sen's example: In addition to alternatives *l* and *p*, there is alternative *o*, in which nobody reads the book. On libertarian grounds, we must prefer *l* over *o*, because *L* prefers to read

101. Sen, *supra* note 1, at 155.

102. *Id.*

103. Note that we define the Pareto principle based on what individuals actually prefer, including not only their personal preferences but also their external preferences. If we were to define the Pareto principle based on personal preferences only, then the conflict with liberal consequentialism would disappear. I argue, however, that a liberal theory of social welfare should respect a Pareto principle based not only on personal preferences but also external preferences. *Infra* Part III.

104. Sen, *supra* note 6, at 479-80. Sen uses the term "liberalism" for this concept in his original article. Sen, *supra* note 1, at 153-54. In subsequent writings, Sen replaces this "more ambiguous" term with "libertarianism." Sen, *supra* note 2, at 218.

105. Sen, *supra* note 1, at 153.

the book, and we must also prefer *o* over *p*, because *P* prefers not to read the book. Yet the Pareto principle requires us to prefer *p* over *l*, in accord with the unanimous preferences of *P* and *L*, which make *p* "Pareto superior" to *l*, thus completing a strict "preference cycle."[106] In this "Pareto-libertarian cycle," *p* is strictly better than *l*, which is strictly better than *o*, which is strictly better than *p*.[107] These strict preference cycles imply that we cannot identify the optimal social choice, and thus a theory of social welfare that produces such a cycle does not satisfy what theorists normally consider a basic requirement for social welfare functions.[108]

B. *The Horns of the Dilemma*

Sen infers that "in a very basic sense liberal values conflict with the Pareto principle," which "can have consequences that are, in fact, deeply illiberal."[109] He states the dilemma as follows: "If someone takes the Pareto principle seriously, as economists seem to do, then he has to face problems of consistency in cherishing liberal values . . . ."[110] For Sen, the "impossibility" of a Paretian liberal "points towards a serious questioning of the Pareto principle" or at least of the "mechanical use of the Pareto rule irrespective of context."[111] He concludes that we should not follow the Pareto principle in all cases. Instead, he insists that in some cases, "the optimal solution involves waiving the Pareto principle."[112] At least sometimes, Sen suggests, we should refuse to implement a Pareto improvement if it is based on illiberal preferences.

Others who have considered Sen's problem have gone further than Sen in the direction of libertarianism. Robert Nozick, for example, concludes that we should always give libertarian rights priority over the Pareto principle. He argues that individual rights should be viewed as constraints on all social choices:

> [E]ach person may exercise his rights as he chooses. The exercise of these rights fixes some features of the world. Within the constraints of these fixed features, a choice may be made by a social choice mechanism based upon a social ordering; if there are any choices left to make! . . .

106. Sen, *supra* note 2, at 218.
107. Sen, *supra* note 10, at 550.
108. *See, e.g.*, Sen, *supra* note 9, at 1079 (discussing the requirement of "acyclicity").
109. Sen, *supra* note 1, at 157.
110. *Id.*
111. Sen, *supra* note 2, at 219.
112. *Id.* at 226.

> . . . How else can one cope with Sen's result?[113]

For Nozick, social choices may be guided by a "social ordering," which may well comply with the Pareto principle, but our social choices must never violate individual rights.

Others are inclined to solve Sen's problem by giving priority to the Pareto principle rather than liberalism when the two come into conflict.[114] Kaplow and Shavell, however, draw far more sweeping implications. They address the more general conflict between any fairness notion (including libertarian rights or the use of any other nonutility information) and the Pareto principle, which they are inclined to resolve in favor of the Pareto principle. They view this conflict as "a serious problem for proponents of notions of fairness."[115] They claim that to follow any fairness principle will make everyone worse off in at least some cases, so that "as a matter of logical consistency," the advocate of fairness must sometimes "choose legal rules that hurt everyone."[116] This result, in their view, makes it problematic to "give *any* weight in normative analysis to notions of fairness."[117] That is, they infer that fairness principles should never enter our normative analysis of legal regimes, even in cases where the Pareto principle is not at stake: "Belief in the Pareto principle not only rules out . . . Pareto-dominated policies; it also renders inadmissible certain criteria for assessing policy."[118] Kaplow and Shavell conclude that we should reject all fairness principles. Thus, they would include any utility valued by the individual, including external preferences, in their social welfare function, much as a utilitarian like Hare would.[119]

III. A LIBERAL SOLUTION TO THE CONFLICT

Do we really confront such a stark choice between the Pareto principle and liberal rights? Must we in fact reject all fairness principles, including principles of liberal toleration, in order to avoid violations of the Pareto principle? Is there a solution to Sen's problem? Such a solution seems

113. ROBERT NOZICK, ANARCHY, STATE, AND UTOPIA 166 (1974) (emphasis omitted).

114. *E.g.*, Peter Bernholz, *Is a Paretian Liberal Really Impossible?*, 20 PUB. CHOICE 99, 106 (1974) ("The only . . . conclusion which can . . . be drawn from Sen's proof is that there are situations in which other decision rules than liberalism should be applied.").

115. Kaplow & Shavell, *supra* note 3, at 64.

116. *Id.* at 76.

117. *Id.* at 72 (emphasis added).

118. KAPLOW & SHAVELL, *supra* note 4, at 4; *see* Kaplow & Shavell, *supra* note 3, at 72 (arguing that the Pareto principle "has powerful implications for what *criteria* for making policy choices one can consistently employ" even if it "may not directly determine policy *choices* in most real situations").

119. KAPLOW & SHAVELL, *supra* note 15, at 411-17 (arguing in favor of including all preferences, including objectionable preferences, in the calculation of social welfare).

imperative if we are to uphold liberal values without coming into conflict with the unanimous preferences of the population. In the discussion that follows, I describe a solution proposed in the prior literature and then defend it against Sen's responses.

A. *The Possibility of a Paretian Liberal*

Allan Gibbard proposes a solution to Sen's problem of the Paretian liberal.[120] Rather than qualifying the Pareto principle, Gibbard modifies Sen's libertarian rights so as to avoid a conflict with that principle. He argues that Sen has adopted an unreasonable notion of libertarian rights.[121] His solution is to make these rights alienable, so that when it would make everyone (including all those who hold these rights) better off to waive these rights, then the Pareto principle would prevail.[122]

Consider Sen's example, in which the supposedly liberal outcome (*l*) is for *L* to read the book and for *P* not to read the book. If both *L* and *P* would prefer *P* rather than *L* to read the book, and both can agree to waive their rights, then *L* and *P* would prefer to enter a contract in which *P* agreed to read the book and *L* agreed not to read the book. Once *L* and *P* are free to trade their rights, then Pareto inferior outcomes will not be an equilibrium. If there is a Pareto superior outcome, then the parties can bargain to move to the Pareto optimum. It is in this sense that "the Pareto criterion has been thought to be an expression of individual liberty."[123]

To ensure the supposedly liberal outcome *l*, a rule would need to prevent *L* and *P* from trading their rights as they see fit. Thus, Russell Hardin declares that Sen is "transparently wrong" to regard *p* as an illiberal outcome, because "it is obvious that among the most important of all rights in the liberal canon are the right of exchange and its correlative right of contract."[124] Similarly, Brian Barry suggests that Sen's notion of liberalism is "antithetical to a conception of liberalism that emphasizes the freedom of individuals to make their own choices with as few constraints as possible."[125]

Given the "strong libertarian tradition of free contract," Gibbard argues that it may seem "too paternalistic" for a rule to keep parties from striking "bargains to which everyone would agree."[126] Why should a libertarian

120. Allan Gibbard, *A Pareto-Consistent Libertarian Claim*, 7 J. ECON. THEORY 388, 401-06 (1974).

121. *Id.* at 397.

122. *Id.* at 400-01.

123. Sen, *supra* note 1, at 157.

124. RUSSELL HARDIN, MORALITY WITHIN THE LIMITS OF REASON 109 (1988).

125. Barry, *supra* note 57, at 19.

126. Gibbard, *supra* note 120, at 397. We may, however, believe in a version of liberalism that can justify such paternalistic interventions when parties would make choices that would not

interfere with such a contract between consenting adults? One may disapprove of their motives, but if no one is harmed by the contract, then it would seem consistent with liberal principles to deem their motives to be their own business. Thus, Gibbard suggests, "a libertarian may want to allow the parties to bargain."[127] To do otherwise smacks of illiberal paternalism, because trying to protect the parties from their own preferences would violate the liberal principle of preference autonomy. If liberal principles require that social choices respect decisions not only to exercise libertarian rights but also to waive them, then we can escape from Sen's dilemma. Gibbard concludes: "A libertarian can consistently hold the Pareto principle and still claim that in a strong sense, everyone has rights."[128]

1. *Sen's Responses*

Sen objects to Gibbard's solution, because it allows external preferences to outweigh personal preferences and would always give the Pareto principle priority over liberal rights:

> When meddling in each other's affairs causes a cycle involving the Pareto principle and personal rights, the axe in the Gibbard system falls invariably on personal rights (based on the "self-regarding" part of a person's preference), leaving intact the effectiveness of the Pareto rule (based on the "non-self-regarding" parts of people's preferences).[129]

"To axe invariably personal rights . . . and never the Pareto principle, when they conflict," seems to Sen "hard to justify."[130] Gibbard's approach, however, would give external preferences priority over personal preferences only insofar as these preferences are held by the same person, who gives these external preferences such priority in his own hierarchy of values. If we show a liberal respect for preference autonomy, then it is exactly in such cases that the Pareto principle *should* prevail over Sen's libertarian rights.

Sen also observes that *P* and *L* may not trade, because such a contract would be difficult to enforce. Specifically, *L* "may not be able to ensure that the prude will, in fact, read the book once it has been handed over to

satisfy their "true" preferences. For a discussion of such constraints on individual choice, see *infra* note 153.

127. Gibbard, *supra* note 120, at 398.

128. *Id.* at 406.

129. Sen, *supra* note 2, at 224.

130. *Id.* at 226.

him."[131] In a similar vein, others have noted practical problems in implementing Gibbard's solution.[132] All of these problems, however, merely suggest that the solution would be difficult to achieve as a matter of public policy. These 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.[133] As Sen concedes, the failure of a trade to take place due to these types of difficulties does not imply that it would not have been good for social welfare if it had taken place.[134] These pragmatic objections point to possible "constraints" on the outcomes that the social planner can achieve, but they do not tell us how to choose the "objectives" that the social planner should pursue.[135]

Finally, Sen introduces political commitments into his example and suggests that *P* or *L* may not want to trade, because they may be committed libertarians who "see no moral gain in the 'trade.'"[136] If one of them has libertarian convictions, these political beliefs may prevent one party from satisfying those other-regarding preferences that militate in favor of the trade. If one party refuses to trade on political grounds, however, then *P* reading the book (*p*) is no longer Pareto superior to *L* reading the book (*l*). If we introduce these political commitments and take account of the weight that *P* and *L* give their beliefs in libertarian rights, then all things considered they do not unanimously prefer *p* over *l*, because the decision not to trade reveals a contrary preference. If *p* is not Pareto superior to *l*, then it does not violate the Pareto principle for society to rank *l* over *p* on libertarian grounds.

2. *Utility as Happiness*

A decision by either party not to transact would reveal a preference for *l* over *p*, and we might infer from this "revealed preference" that *L* and *P* do

131. Sen, *supra* note 6, at 482 n.35.

132. *E.g.*, Kaushik Basu, *The Right To Give Up Rights*, 51 ECONOMICA 413 (1984) (exploring problems of incentives under any system in which individuals waive their rights voluntarily); Edi Karni, *Collective Rationality, Unanimity and Liberal Ethics*, 45 REV. ECON. STUD. 571 (1978) (showing how Gibbard's solution may be manipulated by parties who cheat); Jerry S. Kelly, *Rights Exercising and a Pareto-Consistent Libertarian Claim*, 13 J. ECON. THEORY 138 (1976) (exploring difficulties parties would encounter in deciding when to waive their rights).

133. Sen, *supra* note 2, at 224.

134. Sen, *supra* note 10, at 552 n.2.

135. Peter J. Hammond, *Utilitarianism, Uncertainty and Information*, *in* UTILITARIANISM AND BEYOND, *supra* note 25, at 85, 86 (noting that "many common misconceptions have arisen because of a confusion of objectives and constraints").

136. Sen, *supra* note 10, at 551.

not both enjoy higher utility in *p* than in *l*, but Sen resists this inference.[137] He seeks to define utility "in the traditional sense of happiness, or in the sense of a person's own conception of his well-being" rather than in terms of revealed preference.[138] Sen claims that *L* and *P* may still both be happier in *p* than in *l*, even if they would not both choose *p* over *l*. Therefore, the libertarian outcome *l* would still violate the Pareto principle, if we deem *p* Pareto superior to *l* based on "a unanimous utility ranking" rather than by a unanimous preference revealed by choice, so that a refusal to trade would still be a refusal to apply the Pareto principle "without qualification."[139]

Thus, in order to maintain a conflict with the Pareto principle, Sen must depart from economic traditions in two respects. First, he must distinguish between utility and revealed preferences. Second, he must make his definition of the Pareto principle more specific so that it turns on utility levels rather than on preferences that would be revealed by choice.[140] That is, he gives the notion of "individual preference" a "desire interpretation" rather than a "choice interpretation."[141] In effect, he defines utility to exclude preferences based on political commitments, so that *P* or *L* can reveal a preference for *l* over *p* through their decision not to enter a contract, even if *p* gives them more utility than *l*. Thus, as long as "people attach some importance to minding their own business, then that odd contract need not in fact materialize."[142]

It seems, however, that we must include these political commitments in our definition of utility to generate a Pareto principle we would feel an obligation to uphold. If *P* and *L* do not trade, because one of them wants to uphold libertarian values, then it does not seem disturbing if they fail to trade. If *L*, for example, is a libertarian who chooses to place more weight on his libertarian political commitments than his other external preferences, then it would seem illiberal for anyone (say *P*, who wants to trade) to question *L*'s choice by pointing to the utility that *L* would enjoy if *L* and *P* were to trade. Under the principle of preference autonomy, it seems that this utility is for *L* to enjoy or to refuse to enjoy as *L* sees fit. If *L* reinforces his

137. *Id.* at 551-52.

138. *Id.* at 552.

139. *Id.* Thus, whether the refusal to trade is "a violation of the Pareto principle or a consistent application of it depends on whether one regards an individual's liberal values as a part of his preferences or a constraint upon them." DENNIS C. MUELLER, PUBLIC CHOICE II 404 (1989).

140. Sen uses the term "conditional Pareto principle" to describe a Pareto principle that is based on preferences that are conditional on political preferences as well as all other preferences. Sen, *supra* note 2, at 236. He distinguishes this Pareto principle from "Pareto preferences in the traditional sense," by which he means preferences based on happiness. *Id.* at 237. It might be equally accurate, however, to refer to the "conditional Pareto principle" as the "traditional" principle, because the "conditional Pareto principle" turns simply on what outcomes individuals actually prefer, all things considered.

141. Amartya Sen, *Minimal Liberty*, 59 ECONOMICA 139, 145 (1992).

142. *Id.*

personal desire to read the book with a libertarian political commitment, then we might deem it to be *L*'s prerogative to base his personal preference in part on a moral preference.

It robs the Pareto principle of its moral force if one or both of the supposed beneficiaries of the Pareto improvement would prefer, all things considered, to forego the benefit. Thus, it seems one might well accept this violation of Sen's desire-based Pareto principle without qualms, and the supposed conflict with this desire-based Pareto principle does not then amount to much of a paradox.[143] As long as *L* has these political commitments, a liberal respect for *L*'s preferences requires others to respect *L*'s refusal to trade, although they may seek to persuade *L* that these political views are mistaken.

We can make the dilemma more forceful if we take the question to be whether *L* should choose to trade with *P* if *L* believes in liberal values rather than whether others must respect *L*'s decision on this issue. Sen claims that "the dilemma of the Paretian liberal remains" in the form of a "dilemma of personal behaviour."[144] If *L* expresses qualms based on the contract's interference with *P*'s right not to read the book, however, then *P* might respond that the reasons for *L*'s hesitation fail to respect the principle of preference autonomy, because *P* seeks to waive the right that *L* wants to preserve on *P*'s behalf. As a liberal, then, *L* would have to base his objection to the trade on his own desire to be free to read the book, despite his stronger desire to get *P* to read the book. Sen's view of liberalism, for example, prescribes the "good liberal practice of reading what one likes and letting others read what they like," even in the face of a unanimous desire to do otherwise.[145] For *L* to assume that a liberal principle requires him to refrain from the trade that *P* proposes, however, begs the question: Is that interpretation of liberalism correct? Perhaps *L* should declare, as Barry does, "I simply deny that there is any such liberal principle."[146]

This question brings us back to the most important issue that Gibbard's suggestion raises: What does liberalism actually require? Barry responds that liberalism is "a principle that picks out a protected sphere, but one that is protected against *unwanted* interference, not against use in trading with others."[147] Thus, this narrower interpretation of liberalism would never

143. Furthermore, if this particular conflict is all that Sen wants to illustrate, then he can do so using a much simpler hypothetical. It would be sufficient to consider a society of one person who would enjoy reading a book but refrains from doing so in order to comply with a moral commitment. In this case, the desire-based Pareto principle would regard reading the book as the better outcome, but a liberal respect for her right not to read the book would suggest the reverse ranking.

144. Sen, *supra* note 141, at 146.

145. *Id.* at 145.

146. Barry, *supra* note 57, at 14.

147. *Id.* at 19 (emphasis added). Sen complains that "Barry supposes that the only barrier to such contracts can arise from people *not being allowed* to enter into such contracts," whereas

require *L* to violate the Pareto principle, because it would always allow parties to trade their rights. If *L* adopts this view of liberal rights, then perhaps *L* should have no qualms about trading. The lesson we draw from Gibbard's suggestion is that the Pareto principle does not imply, as Sen suggests, that "individual liberty may have to be revoked."[148] If the Pareto improvement does not entail the "deeply illiberal" outcome feared by Sen,[149] then why not trade voluntarily as Gibbard suggests?

Why should we apply Sen's liberal rules to a context in which they would reduce everyone's happiness?[150] Must liberalism require this type of "rule worship," when no one desires such an outcome? From this perspective, the conflict between liberalism and the Pareto principle arises only because Sen extends his liberal principles into circumstances where they do not properly apply. We can instead adopt a more limited view of liberalism that would see no harm in the satisfaction of external preferences if everyone would thereby benefit. This form of liberalism would respect the weak Pareto principle. For example, D.K. Osborne adopts such a view of "the liberal principle" and thus rejects Sen's liberalism:

> The weak Pareto principle operates only in case of unanimity; . . . in that case the liberal principle is empty. On the other hand the liberal principle is forceful only in cases of certain kinds of disagreement; and in those cases the Pareto principle is silent. Thus when the one is binding the other is either empty or silent. If that is true they cannot possibly be inconsistent.[151]

Under a formulation of liberalism that would also respect the strong Pareto principle, the liberal objection to external preferences would apply only if someone is harmed by the satisfaction of those preferences.

"[t]he real issue is . . . whether they will *seek* such a contract." Amartya Sen, *Foundations of Social Choice Theory: An Epilogue*, *in* FOUNDATIONS OF SOCIAL CHOICE THEORY, *supra* note 22, at 213, 225. We can interpret Barry's claim, however, to be that liberalism is silent on the issue of whether people should enter such contracts. If so, then liberalism does not conflict with the Pareto principle on this issue.

148. Sen, *supra* note 2, at 235.

149. Sen, *supra* note 1, at 157.

150. *See, e.g.*, Bernholz, *supra* note 114, at 100 ("[N]o one would dream of applying one decision rule like the rule of liberalism under all circumstances.").

151. D.K. Osborne, *On Liberalism and the Pareto Principle*, 83 J. POL. ECON. 1283, 1286 (1975). Sen responds that Osborne's claim is "based on an unadulterated piece of logical error." Sen, *supra* note 2, at 228. Sen explains that his strict preference cycle arises from comparisons between pairs of alternatives: "The Pareto principle can operate over one or more pairs (without conflicting with the liberal principle over *those* pairs) and the liberal principle can operate over two or more other pairs (without conflicting with the Pareto principle over *those* pairs), and these choices *together* can be inconsistent." *Id.* at 229. Osborne's claim holds true, however, if we interpret it as a challenge to the "liberal principle" that Sen applies to generate the preference cycle. If we apply a more modest liberal principle that declines to rank pairs when the holder of a right wants to waive the right (or perhaps reverses the ranking when the holder wants to waive the right rather than exercise it), then Sen's response to Osborne fails.

If *L*'s decision not to trade disturbs us because it conflicts with Sen's utility-based Pareto principle, then our real quarrel is with the content of the liberal theory used by *L*, not the Pareto principle. If *L* confronts the choice between Sen's version of liberalism and Sen's desire-based Pareto principle, *L* can reject Sen's liberalism and instead adopt a less restrictive form of liberalism that is more sensitive to context and to human happiness. It does not seem "deeply illiberal" to refine liberalism along these lines. Thus, on the issue of whether *L* and *P* should choose to trade, there would be no conflict between liberalism, properly interpreted, and even Sen's desire-based Pareto principle. After all, the mere possibility of a plausible form of liberalism that respects the Pareto principle is sufficient to disprove Sen's impossibility claim regarding the Paretian liberal.[152]

3. *Revealed Preference*

If, on the other hand, we include political commitments in our notion of utility, or if we retain a Pareto principle defined in terms of unanimous preferences that would be revealed by choice, then we have a Pareto principle that seems more plausible as a universal rule. Once we adopt this version of the Pareto principle, however, Gibbard's alienable rights provide a simple solution to the conflict between even Sen's demanding version of liberalism and the Pareto principle. Even if the parties refuse to trade because they are committed libertarians, their refusal would satisfy the Pareto principle based on revealed preference.[153]

152. Sen's impossibility claim is striking only if his notion of liberalism is uniquely plausible. *See* Osborne, *supra* note 151, at 1285 n.3 ("[I]f a person is going to insist on claiming 'The Impossibility of a Paretian Liberal,' he must expect people to examine his concept of liberalism.").

153. Preferences revealed by choice would be what Harsanyi calls "manifest" preferences. *Supra* notes 85-86 and accompanying text. If these preferences are not "true" preferences, then we have another ambiguity regarding what we mean by the Pareto principle. If we want to uphold the Pareto principle as applied to "manifest" preferences, then alienable rights would ensure compliance with that Pareto principle.

If, however, we choose to uphold the Pareto principle as applied to "true" preferences, then alienable rights would not ensure compliance with that Pareto principle, because parties may choose not to trade even if their "true" preferences would indicate that they should. This Pareto principle could in theory justify paternalistic intervention to satisfy the parties' "true" preferences. This intervention could require a trade when one or both parties prefer not to trade, or it could prohibit a trade when both parties would prefer to trade.

We should not, however, attribute these restrictions on individual choice to the Pareto principle per se. If we object to these paternalistic restrictions on individual liberty, then we object to the reliance on "true" preferences as the morally relevant preferences, not to the Pareto principle itself. If we instead accept the use of "true" preferences rather than "manifest" preferences, then we would not object to such paternalistic interventions in theory. We would understand libertarian rights to be qualified by our theory of justified paternalism and therefore would not consider it an infringement of such rights for the government to implement a Pareto improvement based on "true" preferences. This conflict between libertarian rights and the Pareto principle should disturb us only if we adopt a definition of the Pareto principle inconsistent with our own beliefs regarding paternalistic interventions.

In fact, with this version of the Pareto principle, it is not hard to avoid Sen's problem, even if libertarian rights are inalienable, as long as we qualify the assumption of unrestricted domain in the following respect. Sen notes that as long as "*at least one* person respects the rights of others," such that this person would always prefer the libertarian outcome "and wants that preference to count" in "deciding what is socially better," then libertarian rights will never conflict with the Pareto principle, because the libertarian's preference will always prevent an illiberal outcome from being Pareto superior to the libertarian outcome.[154] That is, it only takes one liberal to veto what would otherwise be a Pareto improvement and thereby eliminate the potential conflict.[155]

In fact, we do not even need one person to be consistently liberal. It would suffice to find one liberal for each decision in which an illiberal outcome would otherwise be Pareto superior to a liberal outcome. As long as this one liberal opposes the outcome on liberal grounds, then the illiberal outcome is not Pareto superior, and the application of the liberal principle does not conflict with the Pareto principle.

What if the liberal in Sen's example is neither *P* nor *L*, however, but some third party that we add to the hypothetical? What if this liberal objects to the alienability of libertarian rights, thus opposes a trade that both *P* and *L* want to make, and thereby prevents application of the Pareto principle to the transaction? Sen suggests that an "outsider may try to judge what should be done and may decide that certain parts of a person's preferences *should not* count in the choice in question," thereby creating new violations of the Pareto principle.[156] Should we be disturbed by the conflict between this liberal preference and the Pareto principle, which would favor the trade?

We should recall that the liberal's opposition merely implies that the trade is not required by the Pareto principle. The liberal's opposition does not actually prevent the trade unless we assume that our system of libertarian rights somehow gives this liberal a right to veto the trade. A libertarian legal system may give the liberal no say in the matter at all and thus leave it entirely up to *P* and *L* whether to trade.

If the only issue is whether we should consider the trade to be a Pareto improvement in light of the liberal's opposition, then it is hard to see why we should exclude the liberal's political views from consideration. The liberal's disapproval of a transaction between *P* and *L* may seem meddlesome under the principle of preference autonomy, but why should we count *P*'s and *L*'s external preferences regarding each other and not

154. Sen, *supra* note 2, at 236.

155. Kotaro Suzumura, *On the Consistency of Libertarian Claims*, 45 REV. ECON. STUD. 329, 332 (1978).

156. Sen, *supra* note 2, at 237.

count the liberal's political preference regarding *P* and *L*'s trade? If we are prepared to count *P*'s and *L*'s other-regarding preferences, then why not count the liberal's as well? Why should we distinguish between these equally external preferences? If we count all these preferences, then the trade between *P* and *L* is not in fact a Pareto improvement.

Suppose, however, that the liberal does have a right to interfere because, say, the liberal is a judge with the power to void the contract between *P* and *L*. If the judge prevents this contract, citing liberal principles, he does not violate a Pareto principle based on revealed preferences, because his decision itself reveals a preference against the contract. As Gibbard suggests, however, the judge's view of liberalism would not seem very liberal at all.[157] This hypothetical takes us back to the substantive content of the theory that we attribute to the liberal. If the judge adopts a view of liberal rights like that suggested by Gibbard, then the judge would not prevent the transaction between *P* and *L*. That is, the trade between *P* and *L* would be consistent with an appropriate notion of liberalism, which would make libertarian rights alienable, so that the parties could waive their rights if they prefer to do so.

B. *The Possibility of a Fair Paretian*

We can formulate a response to Kaplow and Shavell's general claim about fairness that is similar to our response to Sen's claim about libertarian rights. That is, we can object that Kaplow and Shavell have assumed an unreasonable notion of fairness, much as Sen has assumed an unduly demanding notion of liberalism. A more plausible fairness theory would qualify the scope of fairness principles so as not to conflict with the Pareto principle. In fact, the Pareto principle itself could be part of a complete theory of fairness. Any plausible fairness theory includes more than one moral principle, and any such pluralistic theory must specify when some principles take priority over others. There is nothing in the nature of fairness theories that prevents them from including the Pareto principle among their principles and from giving the Pareto principle priority over others. The incorporation of a fairness principle into a pluralistic moral theory does not mean that we must apply this principle automatically, in all cases, as a universal rule that overrides all other principles. Rawls, for

157. Sen agrees with Gibbard on this point, stating, "I see no reason why rights of this kind should not in general be taken to be open to contracting and exchange through mutual agreement." Sen, *supra* note 141, at 145-46.

example, suggests putting principles in "serial or lexical order," so that some take priority over others.[158]

We might, for example, conceive of a fairness theory as endowing individuals with alienable rights that they may invoke to protect their interests from unfair interference by others. If presented with a situation in which it would be in their interests to waive this right, however, then they may do so, as Gibbard would allow individuals to waive libertarian rights. This alienable right would prevent any conflicts with the Pareto principle, because the holder of the right would not invoke it if everyone (including the holder) would prefer that the holder waive the right.

Such a fairness theory would be much less vulnerable to the charge of "rule worship" than the crude fairness notions assumed by Kaplow and Shavell.[159] Under this theory, whenever fairness concerns come into play, these fairness principles serve the interests of the person invoking these principles. If the charge of unfairness is well-grounded, then the unfair outcome would be unfair to at least one person. If fairness principles affect social choices only when an aggrieved individual exists who would prefer to invoke them, then to take these fairness notions into account would never amount to "rule worship." We would never apply an abstract fairness principle simply for its own sake, in violation of the Pareto principle, making everyone in society worse off as a result. Instead, we would employ the fairness principle only in defense of the interests of some identified individual.

Kaplow and Shavell claim that such a qualified fairness theory would not be applying fairness principles on a consistent basis. Thus, they argue that "as a matter of logical consistency, if one is to adhere to the Pareto principle, one cannot give any weight in normative analysis to notions of fairness."[160] A more reasonable theory of fairness, however, would reject the type of mechanical "consistency" that Kaplow and Shavell require. Fairness principles, for example, might be principles designed to resolve disputes only when preferences conflict; these principles would not be relevant when preferences are unanimous. It should seem natural to restrict the domain of fairness principles in this way: If there is unanimous agreement that one state of affairs is better than another, then there would be no issue for fairness principles to resolve.

These fairness principles would be principles of social justice, designed to evaluate when it is legitimate to sacrifice the interests of some in order to satisfy the preferences of others. Rawls, for example, describes "the role of

158. RAWLS, *supra* note 53, at 42; *see, e.g., id.* at 302-03 (giving liberty priority over his second principle of justice and giving his principles of justice priority over efficiency and welfare).

159. Kaplow & Shavell, *supra* note 3, at 72-73.

160. *Id.* at 72.

the principles of justice" as defining "the appropriate distribution of the benefits and burdens of social cooperation" in the face of "a conflict of interests" over this distribution.[161] These principles of distributive justice are necessary to rank two alternatives when neither alternative Pareto-dominates the other. We do not need these principles, however, in order to rank two alternatives when one Pareto-dominates the other. In the case of a Pareto improvement, there is no conflict of interest to resolve.

To take a more concrete example, consider the liberal consequentialism described in Part I. To avoid a conflict with the Pareto principle, we would need to qualify that consequentialism much as we qualified Sen's liberalism to avoid a conflict with the Pareto principle. We would have to count external preferences when we rank a Pareto superior outcome over a Pareto inferior outcome, but we might exclude external preferences when we choose between two alternatives that cannot be ranked using the Pareto principle. That is, as Thomas Scanlon urges,

> [W]hen we set out to compare . . . *conflicting* interests with the aim of supporting a moral judgment as to which should be allowed to prevail, what we do is not compare how strongly the people in question feel about these interests . . . but rather inquire into the reasons for which these benefits are considered desirable."[162]

Kaplow and Shavell claim that it would be "odd" to adopt different principles for normative analysis depending on whether one of the alternatives under consideration Pareto-dominates the other, because "trivial changes in facts would alter the entire basis for assessing legal policies."[163] They illustrate this point using a hypothetical in which everyone is better off under regime *Y* than in regime *Z*, but person *i* is better off by only one penny. Consider regime *X*, which is exactly like regime *Y* except for its effect on person *i*, who is worse off under regime *X* than under regime *Z* by one penny. Why should a fairness principle come into play in comparing *X* and *Z*, but not in comparing *Y* and *Z*, just because one person is better off under *Y* than under *X* by pennies?

The answer is that these two choices are fundamentally different. There is a critical qualitative difference between these choices, even if they seem similar in terms of quantitative welfare effects. There is a risk that choosing *X* over *Z* is unfair to person *i*, who may suffer a loss (albeit a small loss) in order to satisfy the external preferences of someone else. There is no such risk in choosing *Y* over *Z*, because even person *i* benefits from that

161. RAWLS, *supra* note 53, at 4.
162. T.M. Scanlon, *Preference and Urgency*, 72 J. PHIL. 655, 660 (1975) (emphasis added).
163. Kaplow & Shavell, *supra* note 3, at 72 n.20.

reform.[164] Thus, *Y* is unambiguously better than *Z*, but *X* is not, even if *X* and *Y* differ from one another by only two pennies.[165]

To reject a Pareto superior outcome like *Y* for a Pareto inferior outcome like *Z* because we disapprove of the preferences that make *Y* the Pareto superior outcome would fail to show liberal respect for the principle of preference autonomy. We would be telling individuals that they should not seek to satisfy their preferences, even if no one objects, because we do not respect their preferences. This intrusion into individuals' intrapersonal comparison of values would seem inconsistent with liberal principles. Thus, it should seem especially appropriate for a liberal consequentialist to respect the Pareto principle in ranking outcomes.

When comparing outcomes like *X* and *Z*, which do not Pareto-dominate one another, matters are quite different. To reject an outcome that makes some worse off compared to the alternative does not violate the principle of preference autonomy. As Scanlon explains:

> The reasons . . . for excluding "moral" and "antisocial" preferences from the determination of individual utility functions are not strictly inconsistent with this principle, because they do not assert that the fulfillment of these preferences is not good for the individuals in question. All that is asserted is that these preferences "have no claim on us"—that is, on society—for their fulfillment. Denying that they have such a claim need not involve "telling people what is good for them"—it represents a moral judgment, not a judgment of value that is in conflict with theirs.[166]

The exclusion of external preferences is then entirely appropriate when we compare outcomes that the Pareto principle cannot rank, because then we are asking some individuals to make sacrifices to satisfy the preferences of others. In this context, interpersonal comparisons of welfare are necessary. We must ask what preferences give rise to a moral claim, which others are obliged to satisfy, and "differences between preferences . . . are of great relevance when these preferences are taken as the basis for moral claims."[167]

164. Fried states this point "in Kantian terms"; that is, to permit a Pareto improvement "means we are not trading off one person's welfare for another's," and thus "we are not using one person for another's ends." Fried, *supra* note 69, at 24. Where "one person is left worse off," however, "for a Kantian the problem of whether one person's rights are unfairly compromised for the good of others is raised." *Id.* at 25.

165. It is not unusual that the outcome of our normative analysis might turn on two pennies. It would also be true under utilitarianism, for example, that we might rank alternatives differently in marginal cases if we change one alternative by mere pennies.

166. Thomas M. Scanlon, *The Moral Basis of Interpersonal Comparisons*, *in* INTERPERSONAL COMPARISONS OF WELL-BEING 17, 28 (Jon Elster & John E. Roemer eds., 1991).

167. Scanlon, *supra* note 162, at 659.

In the case of a Pareto improvement, on the other hand, no one requires anyone else to make a sacrifice. Thus, no interpersonal comparisons are necessary, and no one needs to assert a moral claim over anyone else. It is noteworthy that when Dworkin explains why a utilitarian should exclude external preferences, he suggests that a utilitarianism that counted such preferences would fail to provide "a justification compatible with the right of those it disadvantages to be treated as equals" if they "will suffer . . . from the fact that others think them less worthy of respect and concern."[168] Similarly, when Harsanyi justifies the exclusion of antisocial preferences, he explains that such preferences would not justify "hurting a third person."[169] The justifications for excluding external preferences, therefore, address the usual case in which satisfying an external preference harms someone else. The underlying reasons for the exclusion of external preferences do not apply in the unusual case in which satisfying these preferences meets with unanimous approval and thus produces a Pareto improvement.[170] This distinction is consistent with Mill's liberal principle that "the only purpose for which power can be rightfully exercised over any member of a civilized community, against his will, is to prevent harm to others."[171] Thus, a reasonable fairness theory, especially a liberal theory, might concede that a Pareto improvement is always a desirable choice, regardless of the nature of the preferences that make the choice a Pareto improvement, but simultaneously maintain that when we make some worse off to satisfy the preferences of others, it must be for a legitimate reason. A liberal theory of fairness might elaborate on what constitutes a legitimate reason by excluding the satisfaction of external preferences from this set of reasons.

A critic of fairness might try to resurrect the conflict between fairness and the Pareto principle as Sen resurrects the conflict between liberalism and the Pareto principle. That is, the critic might suppose that one party is considering a transaction that would violate a fairness principle yet make everyone better off. There is no issue of preference autonomy, because the individual is contemplating imposing fairness principles upon herself, not upon any other person. Our first response might be to assert that the correct fairness theory would not stand in the way of such a transaction. If this party persists in the belief that the transaction is unfair, however, and

168. DWORKIN, *supra* note 40, at 235.

169. Harsanyi, *supra* note 63, at 56.

170. In a similar vein, Posner criticizes utilitarians for ascribing value to "envy and cruelty" but defends a "well-regulated market economy" in which an individual cannot "promote his self-interest without benefiting others as well as himself." POSNER, *supra* note 70, at 83 (noting that "lawfully obtained wealth is created by doing things for other people—offering them advantageous trades"). Posner endorses a "rights system" that would require a sadist to buy his victims' consent and thus "to pay them whatever compensation they demand." *Id.* at 82.

171. MILL, *supra* note 82, at 6.

refuses to transact, then the principle of preference autonomy requires us to respect the preference revealed by that choice, which implies that the transaction is not (all things considered) a Pareto improvement.

Following Sen, the critic of fairness might insist that the transaction remains a Pareto improvement if we define preference in terms of desire rather than choice and thereby exclude political commitments from our definition of utility. Our response would be twofold: First, the conflict identified would be between the Pareto principle and a false theory of fairness, not a correct theory of fairness. Second, in any event, a Pareto principle defined in terms of desire rather than choice would not be a principle that we would be inclined to apply as a universal rule.

In any event, Kaplow and Shavell do not exclude political preferences from their definition of utility. Unlike Sen, they include moral and political preferences (such as "tastes for fairness") in their definition of individual utility.[172] Therefore, the conflict that they describe is even easier to avoid than Sen's in the following sense: As long as one individual is "fair," such that this individual always prefers the outcome dictated by the fairness theory in question, then this fairness theory never conflicts with the Pareto principle, because this preference for fairness always prevents an unfair outcome from being Pareto superior to a fair outcome. It is also sufficient if, for each decision between an unfair outcome and a fair outcome that would otherwise be Pareto inferior to the unfair outcome, there exists at least one person who prefers the fair outcome on fairness grounds. Kaplow and Shavell's claim of conflict with the Pareto principle assumes an unrestricted domain of possible individual preferences and thus requires that neither of these plausible conditions hold for the fairness theory in question.

IV. FAIR SOCIAL WELFARE FUNCTIONS

One feature of utilitarianism that its proponents find appealing is its ability to provide a "complete" ranking of alternatives.[173] That is, utilitarianism assigns each alternative a rank such that in any given pair of alternatives, either one alternative ranks higher than the other or the two alternatives are of equal rank.[174] Can a fairness theory generate a complete

172. Kaplow & Shavell, *supra* note 3, at 75.

173. Sen & Williams, *supra* note 91, at 17 (noting that "completeness is often seen as a merit" and that with full interpersonal comparison of cardinal utility, "utilitarianism yields a complete ordering").

174. *See* ROBIN BOADWAY & NEIL BRUCE, WELFARE ECONOMICS 34 (1984) (defining a "complete" ordering); JOHN VON NEUMANN & OSKAR MORGENSTERN, THEORY OF GAMES AND ECONOMIC BEHAVIOR 26 (3d ed. 1953) (same). The Pareto principle by itself, on the other hand, does not provide a complete ranking of alternatives. The Pareto principle can rank alternatives only if one alternative is Pareto superior to the other. If neither alternative in a pair Pareto-

ranking of feasible alternatives without conflicting with the Pareto principle? That is, can a fair theory of social welfare remain faithful to the Pareto principle over an unrestricted domain of possible individual preferences without generating cycles of preference like those encountered by Sen? If so, what would such a social welfare function look like?

To illustrate the features of such a social welfare function, consider some "crude" fairness criterion that generates a complete ranking of these states of the world, but disregards the Pareto principle and therefore sometimes violates that principle. This fairness criterion might be, for example, the liberal consequentialism outlined in Part I, or it may be a more complex theory of fairness. The precise content of this fairness theory does not matter for our purposes here, as long as it provides a complete ranking and sometimes violates the Pareto principle. Let *F* denote this fairness criterion, and let *S* denote the set of feasible alternatives.[175] That is, *S* is the set of alternatives actually available to us as possible social choices in the real world.[176]

Suppose we reject *F* as a universal rule because we believe in the Pareto principle and seek a more refined fairness theory that respects the Pareto principle. Suppose we adopt the principle that we should rank states

dominates the other, then the Pareto principle does not tell us whether one alternative should get a higher rank or whether the two alternatives are of equal rank.

175. I assume that the purpose of our social welfare function is to guide social choices, and we can make such choices only among feasible alternatives. *See* Kevin W.S. Roberts, *Interpersonal Comparability and Social Choice Theory*, 47 REV. ECON. STUD. 421, 422 (1980) ("We are interested in obtaining a social ordering . . . defined over the set of feasible social states *X*.") (emphasis omitted). The "unrestricted domain" condition then requires that our social choice function generate a ranking of those feasible alternatives for any possible set of individual preferences. *Id*.

176. A social choice can have implications for the feasibility of alternatives in the future. To analyze a sequence of social choices over time, it is useful to conceive of alternatives as a game theorist conceives of strategies chosen by a player in a game. Game theorists define a "strategy" as "a complete plan of action," which "specifies a feasible action for the player in every contingency in which the player might be called on to act." ROBERT GIBBONS, GAME THEORY FOR APPLIED ECONOMISTS 117 (1992) (emphasis omitted); *see also* ERIC RASMUSEN, GAMES AND INFORMATION: AN INTRODUCTION TO GAME THEORY 24 (1989) (defining a strategy for any player as "a rule that tells him which action to choose at each instant of the game, given his information set"); VON NEUMANN & MORGENSTERN, *supra* note 174, at 79 (defining a strategy as "a plan which specifies what choices [a player] will make in every possible situation, for every possible actual information which he may possess at that moment"). We can similarly define an alternative as "a complete contingent plan in advance." DREW FUDENBERG & JEAN TIROLE, GAME THEORY 85 (1991).

At any given decision node, we may also be uncertain regarding whether a specific result is feasible. We can analyze the social choice at that node by considering the attempt to achieve that result as a feasible action with uncertain consequences. We can evaluate actions with uncertain consequences according to the expected values for utility and for fairness, taking the average value over the possible results, with each result given weight in proportion to the probability of that particular outcome. *See* KENNETH J. ARROW, ESSAYS IN THE THEORY OF RISK-BEARING 53 (1971) ("One action will be preferred to another if and only if the expected value of the utility of its consequences is greater."). *See generally* LEONARD J. SAVAGE, THE FOUNDATIONS OF STATISTICS 263 (1954) (defining "expected value").

according to criterion *F* unless to do so would violate the weak Pareto principle. Criterion *F* could violate the weak Pareto principle either directly or indirectly (by generating a preference cycle). Thus, we give the weak Pareto principle lexical priority over criterion *F*. We could also use the strong Pareto principle if we wanted our social welfare function to uphold not only the weak principle but also the strong principle, but the weak Pareto principle is sufficient to illustrate the relevant issues.

We can ensure compliance with the Pareto principle if we allow individuals to alienate the right to fairness. In particular, suppose that each individual enjoys an alienable right to an alternative that is optimal according to criterion *F*. This optimum determines a baseline entitlement for each individual, who can exercise an option to waive the right to this "fair" outcome in exchange for another alternative that the individual prefers over this "fair" benchmark. If the preference for this alternative over the "fair" benchmark is unanimous, then this alternative is Pareto superior to the "fair" outcome, and we must deem social welfare to be higher under the Pareto superior alternative. In effect, if and only if there is unanimous agreement, then we allow parties to trade their rights to the fairness optimum in exchange for an alternative that all prefer, and we require our social welfare function to respect this unanimous preference in ranking these alternatives.[177]

To guard against preference cycles, suppose we also impose the condition of "transitivity" on our social welfare function. "Transitivity" means that if *x* is at least as good as *y*, and *y* is at least as good as *z*, then *x* is at least as good as *z*.[178] Transitivity also implies that if *x* is socially preferable to *y*, and *y* is socially preferable to *z*, then *x* is socially preferable to *z*, which in turn ensures that there are no cycles of strict preference.[179]

Suppose we search through the rankings under *F* for conflicts with the Pareto principle, then revise these rankings to eliminate these conflicts. Given any finite set *S* of alternatives, we can apply the following refinement procedure to implement this revision:

(1) Identify those alternatives in set *S* that rank highest according to criterion *F*. For example, if criterion *F* is the liberal consequentialism outlined in Part I, then these alternatives are those that appear optimal from the perspective of the satisfaction of personal preferences. Each individual

177. Each individual enjoys the same alienable right to the fair outcome and can waive this right only with the consent of all other individuals. Because this alienable right is enjoyed by all on an equal basis, it differs from the alienable right proposed by Gibbard, which "gives a person a special voice on an issue . . . [and thus] cannot be accorded indiscriminately." Gibbard, *supra* note 120, at 404. A right to fairness that requires unanimous consent for waiver is sufficiently alienable to ensure compliance with the Pareto principle. We can relax this unanimity requirement to allow more frequent waivers of the right to the fair outcome. *Infra* Section IV.C.

178. Sen, *supra* note 9, at 1079.

179. *Id.*

is entitled to have society choose one of these "fair" alternatives. Ultimately, however, we can consider these alternatives to be socially optimal only if the Pareto principle does not require otherwise.

(2) Within this "fair" subset of *S*, eliminate those alternatives that are Pareto inferior to any alternative in set *S*. Any alternatives remaining are Pareto optimal. We can assign these alternatives the highest levels of social welfare, pursuant to criterion *F*, because the Pareto principle does not require otherwise. Denote this Pareto optimal set *A* and assign these alternatives the highest rank. We deem these alternatives to be socially optimal.

(3) Consider the alternatives within the "fair" subset of *S* that are not Pareto optimal. Within this new subset of "fair" alternatives, eliminate those alternatives that are Pareto-dominated by other alternatives in the same subset. All individuals would prefer to waive their rights to these Pareto inferior alternatives in exchange for the Pareto superior alternatives. We must assign these Pareto inferior alternatives a lower level of welfare than the Pareto superior alternatives, pursuant to the Pareto principle.

(4) Let *C* denote the remaining subset of "fair" alternatives, and assign these alternatives equal levels of social welfare. The Pareto principle does not require different rankings for these alternatives, and we are indifferent among these alternatives according to criterion *F*. Among the alternatives falling outside set *A*, these alternatives are optimal only provisionally, subject to further application of the Pareto principle, because other alternatives that appear inferior according to our fairness criterion *F* may Pareto-dominate some of the alternatives in set *C*.

(5) Identify those alternatives in set *S* that Pareto-dominate at least one alternative in set *C*, and let *B* denote this new subset of *S*. For example, if criterion *F* is the liberal consequentialism we applied to Sen's example, then alternative *l* (in which *L* reads the book) may fall in set *C* but is nevertheless Pareto inferior to alternative *p* (in which *P* reads the book). If *l* falls in set *C*, then we must place *p* in set *B*. All individuals would prefer to waive their rights to the Pareto inferior alternatives in set *C* in exchange for the corresponding Pareto superior alternatives in set *B*. We assign all alternatives in set *B* higher levels of social welfare than assigned alternatives in set *C*, pursuant to the Pareto principle.

(6) Let *D* denote the set of all alternatives in set *S* that fall outside sets *A*, *B*, and *C*. All alternatives in set *D* either are Pareto-dominated by at least one alternative in set *C* or receive lower rankings than alternatives in set *C* according to criterion *F*. Either the Pareto principle or criterion *F* requires us to deem social welfare to be lower in alternatives in set *D* than in alternatives in set *C*. Therefore, we assign all alternatives in set *D* lower levels of social welfare than assigned alternatives in set *C*. For example, suppose criterion *F* is the liberal consequentialism we applied to Sen's

example. The Pareto principle cannot rank alternative *o* (in which neither *P* nor *L* read the book) and alternative *l* (in which *L* reads the book), but criterion *F* would rank *l* over *o* based on the satisfaction of *L*'s personal preferences. If alternative *l* falls in set *C*, then we place *o* in set *D* and assign it a lower level of social welfare than that attributed to *l*.

(7) To rank alternatives within set *B* or set *D*, repeat steps (1) through (6) for each set. That is, apply the same steps to each set that we applied to set *S*. For example, use steps (1) through (6) to partition set *B* into subsets *BA*, *BB*, *BC*, and *BD*. In deciding which alternatives within set *B* are socially optimal, we continue to use both the fairness criterion *F* and the Pareto principle in ranking the alternatives in set *B*. Criterion *F* determines the baseline entitlements (subset *BC*) in set *B* that individuals are allowed to trade for Pareto superior alternatives (in subset *BB*). Continue to partition into smaller subsets until all alternatives in set *S* have been ranked. Because each partition complies with the Pareto principle, the ranking that results from this series of partitions will also comply with the Pareto principle.

Let *F** denote this procedure and the refined fairness theory that it embodies. This *F** is a social welfare function, which takes information about alternatives in set *S* as inputs and generates a ranking of these alternatives as its output. Because we construct the rankings under *F** through a series of partitions, these rankings cannot feature any preference cycles. The function *F** requires fairness information to determine the location of these partitions, which are based on the identification of fairness optima, and cannot generate its rankings using utility information alone. Therefore, *F** embodies a fairness theory and is not a welfarist social welfare function.[180]

Thus, the social welfare function *F** incorporates fairness principles and satisfies the Pareto principle without generating preference cycles.[181] In

180. The rankings under *F** depend on fairness considerations. One can of course contrive special examples in which the rankings under *F** will not depend on fairness information. For example, consider a set *S* containing only two alternatives, one Pareto superior to the other. Any Paretian theory, including not only welfarism but also *F**, will rank the Pareto superior alternative higher than the Pareto inferior alternative, regardless of fairness considerations. Such examples do not show that *F** is welfarist, however, because *F** would be welfarist only if its rankings were independent of fairness information for any set *S*. As long as *F** incorporates fairness considerations anywhere, under any circumstances, it is a fairness theory and not a welfarist theory. Kaplow and Shavell define "conceptions of fairness" to be principles based "at least in part" on factors unrelated to individual utility, Kaplow & Shavell, *supra* note 3, at 65, so that "fairness" theories include "mixed views," which give weight not only to notions of fairness but also to individual utility, *id.* at 67.

181. The *F** described above is only one example of a fair Paretian social welfare function. We can describe other fair ranking procedures that we can apply to finite sets of alternatives while complying with the Pareto principle. Consider, for example, the following procedure for ranking alternatives in set *S*. Rank all Pareto optimal alternatives according to our fairness criterion *F*. Give the highest rank to those Pareto optimal alternatives that rank highest according to this fairness criterion and declare them to be socially optimal. To rank the remaining alternatives, delete these socially optimal alternatives from set *S*, and repeat the procedure. That is, apply the

fact, it also satisfies the requirement of transitivity. Our compliance with the Pareto principle ensures that the alternatives that rank highest under this definition of social welfare are also Pareto optimal, but *F** also allows fairness principles to set baseline entitlements and thus to determine *which* Pareto optimum we will consider socially optimal. Thus, fairness principles and the rights they confer upon individuals play an important role in determining the optimal distribution of benefits and burdens in society.

As long as set *S* remains finite, we can apply the function *F** to an arbitrarily large number of alternatives.[182] We cannot, however, apply *F** to all infinite sets of alternatives.[183] If we want a social welfare function that we can apply to any infinite set *S*, then we can consider the following procedure, which is analogous to *F**:

(1) Rank all alternatives according to criterion *F*. To eliminate the possibility that a Pareto inferior alternative and a Pareto superior alternative have the same rank, apply a welfarist principle to break ties in fairness rankings.[184] The use of a welfarist tiebreaker creates a finer set of rankings within each fairness rank, much as *F** identifies set *C* as a subset of alternatives within a fairness rank.

(2) If any of these ranks within a fairness rank includes Pareto optimal alternatives, then use Pareto optimality as a further tiebreaker. That is, give these Pareto optima a separate higher rank, much as *F** separates set *A* from set *C*. If the highest rank includes such Pareto optima, then these optima are also social optima.

(3) Within each fairness rank, finer rankings based on welfarist principles ensure consistency with the Pareto principle. These rankings, however, may still conflict with the Pareto principle across different fairness ranks. To eliminate the possibility that a Pareto superior alternative has a lower rank than a Pareto inferior alternative in some other fairness rank, give each such Pareto superior alternative a new rank. In particular,

procedure to the remaining alternatives, assuming that the higher ranking alternatives are unavailable. Give the next highest rank to the socially optimal alternatives for the remaining set. Continue until we have ranked all alternatives in set *S*.

182. Finite sets of alternatives have been the traditional focus of social choice theorists. Graciela Chichilnisky, *Social Aggregation Rules and Continuity*, 97 Q.J. ECON. 337, 337 (1982) (noting that in the "vast literature" on social choice, "most existing work has focused on problems where the individuals face a finite number of choices").

183. For example, a technical problem arises if set *D* does not contain its own boundary, such that we cannot identify an optimal alternative within that set. Savage describes this problem: "Suppose, for example, that you were free to choose any income, provided it be definitely less than $100,000 per year. Precisely which income would you choose, abstracting from the indivisibility of pennies?" SAVAGE, *supra* note 176, at 18; *see also id.* ("If infinite sets of available acts are set up and interpreted without some mathematical tact, unrealistic conclusions are likely to follow.").

184. The welfarist principle does not have to be utilitarian. For example, it can be the "leximin" rule. *Infra* note 198. Allowing a welfarist principle to play some limited role in determining rankings does not render this procedure a welfarist social welfare function. *Supra* note 180.

for each alternative, consider the set of all Pareto inferior alternatives. If any Pareto inferior alternative has a higher rank than the alternative in question, then give that alternative a new rank that is higher than the highest ranking Pareto inferior alternative but no higher. Thus, for each rank, create a set of alternatives that ranks higher than the rank in question but lower than any higher rank. This set collects all alternatives that are Pareto superior to any alternative in the rank in question but not to any higher ranking alternative, much as set *B* in *F** collects alternatives that are Pareto superior to at least one alternative in set *C*.[185]

(4) To rank alternatives within these newly created sets in a manner consistent with the Pareto principle, repeat steps (1) through (3) for each set. Continue to partition into smaller subsets until all alternatives of interest have been ranked.[186]

Let *F*** denote this procedure and the fairness theory that it embodies. The *F*** procedure is a social welfare function with features similar to those of *F**.[187] Like the *F** procedure, the *F*** procedure makes its rankings a function of fairness information but nevertheless complies with the Pareto principle. Thus, *F*** is a second example of a fair Paretian social welfare function.

How do our new rankings under *F** and *F*** compare with our original rankings under *F*? The difference between the rankings under either *F** or *F*** and the rankings under *F* will depend on how frequently criterion *F* actually violates the Pareto principle in set *S*. For example, suppose that

185. These new rankings will not rank any Pareto superior alternative below a Pareto inferior alternative, because Pareto superiority is transitive. Thus, when we move an alternative up the rankings to outrank all Pareto inferior alternatives, the alternative in question will not outrank any Pareto superior alternative, because any Pareto superior alternative would have to move at least as high in the rankings. Any Pareto superior alternative would be Pareto superior to any alternative that is Pareto inferior to the alternative in question. Similarly, any Pareto inferior alternative would receive a rank no higher than the alternative in question, even if the Pareto inferior alternative moves up the ranks, because the alternative in question would have to move at least as high in the rankings. The alternative in question would be Pareto superior to any alternative that is Pareto inferior to the Pareto inferior alternative.

186. Ranking an infinite number of alternatives under this procedure could require an infinite number of steps. This feature, however, does not distinguish this procedure from welfarist ranking procedures. A utilitarian, for example, would have to calculate average or total utility for an infinite number of alternatives in order to rank them. Thus, the proposed procedure can provide a complete ranking of an infinite number of alternatives in the same sense that a utilitarian theory can. Using the proposed procedure to rank a *finite* number of alternatives within set *S*, however, could also require an infinite number of steps if set *S* contains an infinite number of alternatives. This feature distinguishes this procedure from welfarist ranking procedures and raises practical problems in actually implementing the procedure. *See infra* Section IV.B (discussing the implications of this analysis for making social choices in practice).

187. The *F*** procedure gives welfarism a somewhat larger role than the *F** procedure does, insofar as *F*** uses a welfarist principle to break ties under criterion *F*. Nevertheless, *F*** never ranks a less fair alternative above a more fair alternative unless the Pareto principle requires this ranking. In this sense, *F*** respects each individual's alienable right to a fair alternative, because when there are multiple fairness optima, no one enjoys an entitlement to any *particular* optimum on fairness grounds.

among all feasible alternatives, we find exactly one conflict between the rankings produced by criterion *F* and the Pareto principle. That is, criterion *F* ranks only one alternative below a Pareto inferior alternative, contrary to the Pareto principle. The only effect of the *F** procedure would be to move that one alternative up the rankings just far enough to outrank the Pareto inferior alternative. In that case, the refined ranking under *F** would differ from the ranking under criterion *F* with respect only to this one alternative. If conflicts between criterion *F* and the Pareto principle are rare, then the social welfare function *F** will resemble fairness criterion *F*. If these conflicts are common, however, then the differences between *F** and *F* will multiply. Therefore, we cannot describe the rankings under *F** without knowing not only criterion *F* but also how often this criterion violates the Pareto principle in set *S*. Similarly, for any given *F*, the change in the rankings produced by the *F*** procedure would be sensitive to the assumptions we make regarding set *S*.

A simple example, however, illustrates how we would expect both *F** and *F*** to preserve an important role for fairness even in the presence of a conflict between criterion *F* and the Pareto principle. Suppose criterion *F* is the liberal consequentialism outlined in Part I. In particular, suppose *F* calls for the maximization of total satisfaction of personal preferences. Suppose also that *P* and *L* are the only two individuals in our society and that we must make exactly two social choices. First, we must choose between alternative *p* and alternative *l* in Sen's example, where both *P* and *L* favor *p* over *l* based on their external preferences.[188] Second, we must distribute wealth to *P* and *L* subject to the constraint that we cannot distribute more than the total amount of wealth in the economy, which we take to be a fixed amount. Thus, each alternative in set *S* specifies not only a choice between *p* and *l* but also the distribution of wealth.

Suppose that *P* would derive the same satisfaction from consuming any given amount of wealth that *L* would and that each individual would derive diminishing marginal utility from consuming this wealth. Then our fairness criterion *F* would not only favor *l* over *p* but also call for an equal division of all wealth between *P* and *L*, which would maximize the satisfaction of personal preferences. Suppose, however, that *L* also harbors malice toward *P*, such that it satisfies *L*'s malicious preference to reduce *P*'s utility. The criterion of utility maximization would take such external preferences into account and thus not only favor *p* over *l* but also call for a division of wealth biased in favor of *L* and against *P*, which would cater to *L*'s malicious preference.

188. For simplicity of exposition, assume that alternative *o* is not a feasible social choice. Adding a third alternative like *o* would make the example more complicated without affecting any important results.

What alternative would we choose under *F** (assuming a finite set of alternatives) or under *F*** (allowing for an infinite set of alternatives) in this example? Our fairness criterion *F* favors outcome *l* combined with an equal division of all wealth. This unique fairness optimum would not be Pareto optimal, however, and would fall into set *C* in the *F** procedure, leaving set *A* empty. Set *B* would be a set of Pareto superior alternatives featuring outcome *p* rather than outcome *l* and would rank higher than set *C*. The social optimum within this set *B* would be outcome *p* combined with an equal division of all wealth, which would be Pareto optimal and therefore fall in set *BA*.[189] Similarly, *F*** would select the same alternative as the social optimum. Thus, both *F** and *F*** would favor *p* over *l*, like any Paretian theory would, but they would also protect *P* from any disadvantage based on *L*'s malice. That is, both *F** and *F*** would ensure that both *P* and *L* enjoy at least the utility that they would enjoy under the fairness optimum under criterion *F*. Welfarist theories cannot guarantee this baseline entitlement, because they must ignore all fairness concerns and thus take all of *L*'s external preferences to be morally relevant.

That is, any given welfarist theory could make an individual worse off than he would be under the fairness optimum under criterion *F*.[190] For example, as long as *L*'s malice is sufficiently intense and *P*'s gain in moving from *l* to *p* is sufficiently small, utility maximization would call for a distribution of wealth so biased against *P* that *P* would prefer the fairness optimum, despite its choice of outcome *l* rather than *p*, because it would also provide *P* with an equal share of wealth. Under either *F** or *F***, fairness entitlements protect *P* from such a biased outcome, even if that outcome is optimal from a welfarist perspective, because that outcome would not be a Pareto improvement over the fairness optimum. In this sense, fairness concerns under *F** or *F*** still trump welfarist considerations even if our fairness criterion *F* yields in the face of a Pareto improvement.[191] Although the right to fairness is alienable under *F** or *F***, an individual like *P* can still invoke this right to veto the social choice favored by a welfarist. Similarly, this right to fairness, for example, would

189. There would be no conflict between criterion *F* and the Pareto principle within this set *B*, because all alternatives within this set would feature outcome *p*. Thus, we would rank alternatives within set *B* simply according to our fairness criterion *F*, which would favor an equal division of wealth.

190. This feature is inherent in welfarism, because any welfarist theory must ignore all fairness considerations and select a social optimum based only on utility information. Given this constraint, no welfarist theory can ensure that the optimum it selects will have any particular relationship with the fairness optimum. Given utility information, the welfarist is committed to the same social choice, regardless of fairness information. Thus, the welfarist cannot adjust this social choice in light of new information regarding the fairness optimum. Only a fairness theory can make its social choice a function of such information.

191. I have described *F** and *F*** as giving the Pareto principle priority over fairness principles, but we can also describe *F** or *F*** as giving alienable rights to fairness priority over welfarist considerations. Both characterizations describe the same theory.

protect the Jew from the Nazi's external preference in Hare's hypothetical, whereas welfarism cannot offer the same guarantee. Thus, the right to invoke fairness concerns under *F** or *F*** ensures that individuals enjoy at least the level of utility to which they would be entitled under *F*, and in this sense protects individual interests to the same degree that *F* would.[192]

A. *Continuity*

These examples are sufficient to disprove Kaplow and Shavell's claim that any social welfare function that incorporates fairness concerns must violate the Pareto principle. Neither *F** nor *F*** is "individualistic" as Kaplow and Shavell use this term.[193] That is, it is possible for two

192. Ranking alternatives in set *D* under *F** in our example requires us to move all alternatives featuring outcome *p* up the rankings so that they outrank any Pareto inferior alternatives featuring outcome *l*. The alternatives featuring outcome *l* would not change rankings with respect to one another, but all alternatives featuring outcome *p* would move up the rankings with respect to at least some alternatives featuring outcome *l*. With each iteration of our refinement procedure, set *C* for the current iteration moves further down the rankings under criterion *F*, further away from the fairness optimum for set *S* as a whole. As we move away from the fairness optimum for set *S*, each iteration allows a greater range of possible levels of utility for each individual in set *C*. This greater range implies that criterion *F* in this example will provide a lower constraint on the utility to which each individual is entitled as we move away from the fairness optimum for set *S*. The rankings under *F*** for the alternatives in set *D* are similar but complicated further by the use of a welfarist principle as a tiebreaker.

The rankings within set *D*, however, would become relevant for the social choices we actually make only if the alternatives in higher ranking sets are not available. If those alternatives are not available, however, then set *S* is not as we have assumed, and we would have to apply our refinement procedure to a different set *S*. A different set *S*, however, would generally imply different rankings for alternatives within set *D*. *See infra* Section IV.B (discussing how the rankings under *F** are generally a function of the entire feasible set *S*). Our rankings within our original set *D* would become relevant only in the very special case that the actual set *S* coincides exactly with this set *D*. In general, however, we would expect criterion *F* and the constraints on set *S* to produce a unique fairness optimum such that fairness principles would give each individual a substantial entitlement. Thus, under *F** or *F***, fairness considerations generally exert the most force precisely where it counts the most as a practical matter: in actually making a social choice from any given set *S*.

193. KAPLOW & SHAVELL, *supra* note 4, at 2. There is some ambiguity regarding what Kaplow and Shavell mean by "individualistic." In general, they appear to use the term to refer to welfarist social welfare functions. They also use the term, however, to refer to all social welfare functions that comply with the Pareto indifference rule, which holds that if every individual is indifferent between two alternatives, then we should be socially indifferent between those two alternatives. *See* Sen, *supra* note 9, at 1075. Kaplow and Shavell assume that compliance with the Pareto indifference rule is equivalent to welfarism, which requires "strong neutrality," such that "the utility information regarding any two social states is all that is needed for ranking that pair." *Id.* at 1121. These two concepts, however, are in general not equivalent. *Id*. at 1154-55. Strong neutrality implies the Pareto indifference rule, but the converse implication does not hold unless we make "quite demanding" assumptions. *Id.* at 1155. That is, social welfare functions may comply with the Pareto indifference rule without being welfarist.

For example, we could have qualified our fairness principle *F* further, so as to comply with not only the weak Pareto principle but also the Pareto indifference rule. To implement this additional qualification, we could modify steps (2) and (4) in the *F** procedure to expand sets *A* and *C* accordingly. For example, if each individual is indifferent between a "fair" alternative in set *C* and another alternative, then we would include both alternatives in set *C*. The rationale

alternatives, *y* and *z*, to yield the same levels of utility for each individual in society yet receive different rankings in terms of social welfare. Suppose, for example, that *y* fares better than *z* according to criterion *F*, such that when we apply the *F** procedure, *y* falls in set *C*, but *z* falls in the lower ranked set *D*. Similarly, it is possible for *F*** to rank *y* above *z*. Kaplow and Shavell offer a formal proof to support their claim that any such social welfare function violates the Pareto principle.[194]

How then do *F** and *F*** incorporate fairness concerns without violating the Pareto principle? Our functions *F** and *F*** do not come under Kaplow and Shavell's formal proof because both functions violate an assumption of continuity that Kaplow and Shavell use to prove their claim.[195] That is, Kaplow and Shavell assume that we can achieve an arbitrarily small change in the ranking of any given alternative by making a sufficiently small change in that alternative in the appropriate respect. Consider, for example, states *y* (in set *C*) and *z* (in set *D*) as described above. Suppose that an arbitrarily small increase in some good for all individuals in state *z* yields a third state *x* in which each individual enjoys only a slightly higher level of utility in *x* than in *z*. Kaplow and Shavell assume that the social welfare function must increase continuously such that a sufficiently small increase in some good for all individuals in state *z* would still yield a lower level of social welfare in the new state *x* than in the "fair" state *y*. This ranking would violate the Pareto principle, which would instead rank the Pareto superior state *x* above the Pareto inferior "fair" state *y*. The *F** procedure, however, would place any such Pareto superior state *x* in set *B*, giving *x* a higher ranking than *both* state *y* (in set *C*) and state *z* (in set *D*), consistent with the Pareto principle. Similarly, the *F*** procedure would place state *x* above not only state *z* but also state *y* in the rankings. That is, any change in state *z* that creates a Pareto improvement over state *z*, no matter how small in quantitative terms, would yield a discontinuity, as the new state *x* would jump past state *y* in the rankings.

Kaplow and Shavell do not explain, however, why we should require a social welfare function to feature their particular continuity property. It is certainly a stronger requirement than the weak Pareto principle. Although a

would be that if all are willing to exchange the "fair" alternative (to which they are entitled) for another alternative, but no one would strictly prefer to make this trade, then we should assign the same level of social welfare to these two alternatives. Similarly, we could modify *F*** to use the Pareto indifference rule to move alternatives up the rankings. I have, however, designed both *F** and *F*** to be "non-individualistic" in both senses used by Kaplow and Shavell, so that they respond specifically to their formal claim without any ambiguity.

194. KAPLOW & SHAVELL, *supra* note 4, at 3. Kaplow and Shavell interpret their proof as if it applied to any social welfare function that incorporates notions of fairness. This proof, however, assumes that all such functions violate the Pareto indifference rule, whereas only a *subset* of all social welfare functions incorporating notions of fairness violate the Pareto indifference rule. *Supra* note 193. Thus, Kaplow and Shavell's formal proof is not as general as they seem to claim.

195. KAPLOW & SHAVELL, *supra* note 4, at 3.

utilitarian social welfare function would be continuous in this sense, other social welfare functions need not exhibit this specific feature. There 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. Indeed, if we accept both the Pareto principle and fairness principles, then we might most reasonably view Kaplow and Shavell's proof as a decisive reason to reject such a continuity requirement as unacceptable, not as a reason to reject fairness principles. At most, Kaplow and Shavell have shown that a social welfare function cannot simultaneously comply with the Pareto principle, incorporate fairness principles, and exhibit their particular continuity feature. If we must reject one of these three features of a social welfare function, however, then we may regard this continuity as the least important of these features. Kaplow and Shavell could have, with equal accuracy, declared their result to be that any "non-individualistic" social welfare function must violate their continuity assumption, but such a result would not be very provocative.

Indeed, any theory based on lexical priority will feature the general type of discontinuity exhibited by F^* and F^{**}.[196] These social welfare functions are discontinuous precisely because they give priority to one principle (the Pareto principle, as embodied in the alienability of individual rights to fairness) over another (the fairness criterion F). Thus, even a small change in circumstances (in our example, in utility) can trigger application of the higher-priority principle (the Pareto principle), which then overrides rankings based on any lower-priority principles (like criterion F). This discontinuity may seem strange to those who are accustomed to weighing all values on a single scale, but those who believe that some values take absolute priority over others should not find such discontinuities disturbing. Continuity is a strong assumption that social choice theorists rarely impose on social welfare functions.[197]

196. *See* BOADWAY & BRUCE, *supra* note 174, at 147 (noting that "lexicographic orderings . . . are precluded by the requirement of continuity").

197. A survey of the literature on social choice theory reveals few examples of theorists who impose any continuity assumptions on social welfare functions. Chichilnisky, *supra* note 182, at 337 ("The vast literature on Social Choice makes little use of the continuity properties of the social aggregation rule."); *see also* Sen, *supra* note 9 (surveying the literature on social choice theory). In this literature, theorists have used continuity assumptions to derive welfarist social welfare functions. *E.g.*, Eric Maskin, *A Theorem on Utilitarianism*, 45 REV. ECON. STUD. 93, 94 (1978) (showing that a social welfare function must be utilitarian if it features several properties, including a continuity property); Roberts, *supra* note 175, at 428 (showing that a social welfare function must be welfarist if it satisfies various conditions, including a continuity condition). Imposing continuity assumptions in conjunction with other conditions is often tantamount to assuming that only welfarist theories are reasonable.

Chichilnisky defends her continuity assumption as "natural," arguing that "it is desirable for the social rule to be relatively insensitive to small changes in individual preferences" so that "mistakes in identifying preferences" are "less crucial." Chichilnisky, *supra* note 182, at 337.

Kaplow and Shavell, however, do not rule out all types of discontinuities. They require only that a social welfare function be continuous in the consumption of some good, but even this particular continuity condition is a strong assumption. It would rule out, for example, giving any welfarist principle lexical priority over any other principle, because an arbitrarily small Pareto improvement would override rankings based on the lower-priority principle, which would create discontinuous changes in rankings. This continuity assumption would preclude not only the use of fairness considerations to break ties in an otherwise welfarist theory but also the use of *any* principles as tiebreakers in a welfarist theory, including welfarist principles. That is, the continuity assumption would even rule out a large class of purely welfarist theories.[198] Most important for our purposes, Kaplow and Shavell's continuity assumption rules out alienable rights to fairness, which in effect give the Pareto principle lexical priority over fairness principles. Such a system of alienable rights would violate this continuity assumption because even a small change in the consumption of some good can be sufficient to create a Pareto improvement that induces all individuals to waive their rights to the "fair" alternative. Given that this continuity assumption rules out many plausible moral

This pragmatic defense of continuity, however, lacks force as a reason to impose a restriction on our ideal social welfare function, which indicates what welfare we would assign to each alternative under conditions of complete information. We must first establish what welfare we would assign under such ideal conditions before we decide what policy is the optimal response to the constraints imposed by imperfect information. Unless we establish our ideal objective first, we cannot decide how we can best pursue that ideal imperfectly in practice. Both F^* and F^{**}, for example, are sensitive to small changes precisely because such small changes may determine whether parties agree to waive their rights. If our moral ideal takes rights seriously in this way, then it is entirely appropriate that we attach such importance to discovering the truth regarding the parties' preferences, and it is inappropriate to pretend that this information is insignificant.

198. It would, for example, rule out a utilitarian theory that first applied the criterion of average utility to rank alternatives and then applied the criterion of total utility as a tiebreaker when the principle of average utility yields a tie. Suppose two states *y* and *z* yield the same average level of utility, but state *y* features a larger population, so that *y* yields greater total utility than *z*. Our tiebreaking rule would rank *y* higher than *z*. Now consider a third state *x*, which is just like *z*, except it includes a slightly higher level of consumption of some good and thus yields a slightly higher level of utility for each individual in *z*. Our hypothetical theory would rank *x* above not only *z* but also *y*, no matter how small the increase in average utility in *x* over *z*, in violation of Kaplow and Shavell's continuity assumption.

As another example, consider the "lexical difference principle" proposed by Rawls. RAWLS, *supra* note 53, at 83. Sen applies this principle to individual utilities and refers to the result as the "leximin" rule. Sen, *supra* note 9, at 1115-21. This rule first ranks states according to the utility of the worst-off individual. If that individual is indifferent among some states, this rule then ranks *those* states according to the utility of the second-worst-off individual, and so forth. The leximin rule also violates Kaplow and Shavell's continuity assumption. Consider a society with two individuals and two states *y* and *z*. Suppose the worst-off person is indifferent between these two states, but the best-off person prefers *y* over *z*. The leximin rule will rank *y* higher than *z*. Now consider a third state *x*, which is just like *z* except that a slightly higher level of consumption of some good produces a slight increase in utility for each individual. The leximin rule will rank *x* higher than not only *z* but also *y*, no matter how small the increase in utility in *x* over *z*. Thus, Kaplow and Shavell's continuity assumption implies that "their real quarry is much broader" than their focus on fairness principles would suggest. Fried, *supra* note 69, at 39.

theories, Kaplow and Shavell bear a heavy burden in justifying their use of this assumption.[199]

B. *Independence*

Both *F** and *F*** violate another condition that a utilitarian social welfare function would satisfy. We might call the following restriction the "independence" condition: The social ranking of any pair of alternatives must be the same as long as the individual utility and fairness (*F*) information about the pair remains the same.[200] We might view this condition as an analogue of the "independence of irrelevant alternatives" condition that Kenneth Arrow used to prove his famous impossibility theorem.[201]

To see how both *F** and *F*** violate this condition, consider the problem of ranking two alternatives *x* and *y* that cannot be ranked by the Pareto principle. In general, we cannot rank these two alternatives without information about the other alternatives in set *S*. We cannot, for example, simply apply criterion *F* to rank the two outcomes unless one of the alternatives is the unique optimum under criterion *F*, in which case the fairness optimum is socially preferable. To apply criterion *F* mechanically when ranking any pair of alternatives that cannot be ranked by the Pareto principle might generate a preference cycle when combined with the Pareto principle. For example, *x* may rank higher than *y* according to criterion *F*, but *y* may be Pareto superior to a third alternative *z*, which in turn may rank higher than *x* according to criterion *F*. By assumption, criterion *F* may rank *z* higher than *y* even though *y* Pareto-dominates *z*. If neither *x* nor *z* Pareto-dominates the other, then our proposed simple decision rule would generate a preference cycle. If our social welfare function is to operate over an unrestricted domain, it cannot use such a simple rule to rank alternatives.[202]

199. *See* Fried, *supra* note 69, at 21-32 (questioning Kaplow and Shavell's continuity assumption).

200. I use the term "fairness information" here to refer to the information necessary to rank alternatives under criterion *F*.

201. KENNETH J. ARROW, SOCIAL CHOICE AND INDIVIDUAL VALUES 26 (2d ed. 1963). Others have since restated Arrow's formal condition in informal terms similar to those used in our independence condition. *E.g.*, ALFRED F. MACKAY, ARROW'S THEOREM: THE PARADOX OF SOCIAL CHOICE 8 (1980) ("This requires that the social ordering of a given set of alternatives depend only on the individuals' preference orderings of those alternatives."); MUELLER, *supra* note 139, at 386 ("The social choice between any two alternatives must depend only on the orderings of individuals over these two alternatives, and not on their orderings over other alternatives."); Sen, *supra* note 10, at 539 ("The social ranking of any pair of states must be the same as long as the individual utility information about the pair remains the same . . . ."). Our condition is more general in that it allows the use of more information in ranking a pair of alternatives.

202. In her critique of Kaplow and Shavell, Fried also notes that combining fairness principles with the Pareto principle can produce conflicts with the requirement of transitivity. Fried, *supra* note 69, at 20-21, 33-35. Fried proposes to solve this problem by ruling out any

Instead, under either *F** or *F***, one could only use criterion *F* to rank *x* provisionally higher than *y*. Alternative *x* would enjoy a rebuttable presumption of higher social welfare, but a proponent of alternative *y* could rebut, for example, by showing that *y* Pareto-dominates *z*, which is not Pareto inferior to *x* and which ranks higher than *x* according to criterion *F*. Upon this showing, both *z* (by criterion *F*) and *y* (by the Pareto principle and by transitivity) would rank higher than *x*, at least provisionally, subject to further rebuttal by a proponent of *x*, for example, based on a similar showing with respect to some other alternative that is Pareto inferior to *x*. Thus, one could not conclusively rank alternatives *x* and *y* without knowing at least the features of all alternatives Pareto inferior to either *x* or *y*. This information requirement violates our independence condition.

Should this violation disturb us? Arrow defends the "reasonableness" of his own independence condition as an axiom that ensures that the ranking of alternatives *x* and *y* are independent of the existence of other alternatives like *z* in our example.[203] He would rule out, for example, a voting procedure in which the ranking of two candidates changes when a third candidate is deleted from the ranks of the candidates. Instead, "the system applied to the remaining candidates should yield the same result."[204] Thus, perhaps the independence condition is appealing because we do not want the rankings of feasible alternatives to change as other alternatives join or leave the feasible set.

A violation of the independence condition may sound odd in the abstract, but a concrete example may make such a violation seem reasonable and illustrate why such a violation would be necessary to maintain compliance with the Pareto principle. Consider Sen's example and assume that *F* is the liberal consequentialism outlined in Part I. Consider the problem of ranking alternatives *p*, in which only *P* reads the book, and *o*, in which no one reads the book. Neither alternative is Pareto superior to the other, because *P* has a personal preference for *o*, and *L* has an external preference for *p*, so the Pareto principle cannot rank these alternatives. Under the liberal principle *F*, we would rank *o* higher than *p*, because *P*'s personal preference is not to read the book. If these two states were the only feasible alternatives, then we could simply choose *o* accordingly. Suppose, for example, that no version of the book exists in a language that *L* can read, so that *l*, in which only *L* reads the book, is not a feasible alternative.

Suppose, however, that *L* discovers a dictionary that would enable *L* to translate the book, making *l* a feasible alternative. If we are to avoid Sen's preference cycle, then we cannot continue to rank *o* over *p* once we

preferences that produce conflicts between the Pareto principle and fairness principles. *Id.* at 35-36. The solution that I propose solves the same problem without ruling out such preferences.

203. ARROW, *supra* note 201, at 27.

204. *Id.*

introduce *l* as a feasible alternative, for *l* would then rank higher than *o* under principle *F*, because *L*'s personal preference is to read the book. The Pareto principle, however, would rank *p* higher than *l*, because *L* and *P* have such intense external preferences that they unanimously prefer *p* over *l*, which completes Sen's strict preference cycle.

Under either *F** or *F***, we choose to break this preference cycle in the same way that Gibbard breaks it: We rank *p* as the optimum among the three alternatives and thus rank *p* over *o*, reversing the ranking that applies in the absence of *l* as a feasible alternative. Following the Pareto principle, we rank *p* higher than *l*, which in turn ranks higher than *o* under criterion *F*, because *L* prefers to read the book. Under Gibbard's scheme of alienable rights, *L* and *P* both agree to waive their rights because they see that the exercise of their libertarian rights would produce outcome *l*, which no one prefers over *p*. Under either system of alienable rights, the feasibility of *l* becomes morally relevant to the ranking of *o* and *p*, because *P* would choose to waive his right to *o* over *p* only if *l* is in fact available as an alternative. The threat of *l* as a social choice is what leads *P* to waive his right, and this threat is credible only if *l* is feasible.

Gibbard's solution to Sen's problem also violates the independence condition, because in deciding whether to waive libertarian rights, each party needs to know how other parties would choose to exercise their rights. One cannot determine the ranking of two alternatives (like *o* and *p*) without knowing the parties' preferences regarding other alternatives (like *l*). Thus, "the outcome of the choice process is not independent of the set of feasible alternatives," and if "the outcome represents the just social state, then justice becomes a relative concept (i.e. relative to the set of feasible alternatives)."[205] If we believe that our social welfare function must respect these types of preferences to waive rights, then we cannot accept the independence condition as a constraint on the social welfare function. Sen's example shows that any solution to Sen's problem that uses the liberal principle to rank alternatives as a general matter but gives priority to the Pareto principle must violate our independence condition over an unrestricted domain.

Given the sweeping effects of the independence condition, should we consider it an essential feature of a social welfare function? It is certainly a stronger assumption than the Pareto principle. Independence conditions have been controversial as restrictions on social welfare functions. Of all Arrow's axioms, for example, his "independence of irrelevant alternatives has been the subject of the most discussion and criticism."[206] Many have

205. Karni, *supra* note 132, at 573 (emphasis omitted).

206. MUELLER, *supra* note 139, at 393; *see also* MACKAY, *supra* note 201, at 48 (describing Arrow's independence condition as "the most controversial of the lot").

found the relaxation of this "much debated condition" to be "an appealing way" of escaping Arrow's impossibility result.[207] Alfred MacKay, for example, has suggested that the assumption that "in socially ranking any pair of candidates a device cannot respond to information about other candidates" is a condition that "does not appear to have much intrinsically to recommend it," so that if it has "undesirable side effects," we could dispense with it "without much regret."[208] Sacrificing the independence condition may be a more attractive option than sacrificing either the Pareto principle or liberal principles.

Perhaps the objection to a social welfare function that violates the independence condition is pragmatic. For example, Arrow cites "a strong practical advantage" in defense of his own independence condition.[209] A system that allows us to rank *x* and *y* without knowing information on the rest of the agenda "has an appealing economy to it."[210] It would be a complex task to use a function like *F** or *F*** to rank any pair of alternatives, because one may have to generate a ranking of many alternatives in order to rank those two alternatives. Fortunately, in order to make social choices under either *F** or *F***, we do not have to rank arbitrary pairs of alternatives. To make such a choice, we could use a simpler procedure than the procedure necessary to rank all alternatives in set *S*. For example, under *F** we would use the following procedure:

(1) Identify the alternatives in set *S* that are optimal under *F*. If any of these "fair" alternatives are also Pareto optimal, then these Pareto optimal alternatives constitute set *A* and are our optimal social choices.

(2) If none of these alternatives are Pareto optimal, then we eliminate any "fair" alternative that is Pareto inferior to any other "fair" alternative. The remaining "fair" alternatives constitute set *C*.

(3) Identify all alternatives that are Pareto superior to at least one of the alternatives in set *C*. This set of Pareto superior alternatives would constitute set *B* and would include our optimal social choices. To identify the optimal choices within set *B*, repeat steps (1) through (3) for set *B*. That is, apply the same steps to set *B* that we applied to set *S*. Continue this procedure until a "fair" alternative is also Pareto optimal. Choose such an alternative as a social optimum.

207. Sen, *supra* note 1, at 156.

208. MACKAY, *supra* note 201, at 92-93.

209. ARROW, *supra* note 201, at 110.

210. MUELLER, *supra* note 139, at 394. Mueller stresses that social choice processes that violate Arrow's condition might be subject to "abuse" or "strategic misrepresentation of preferences." *Id*. at 395. Similarly, MacKay notes that Arrow's condition requires a social welfare function to use only information that can be "reliably ascertained" and avoids "strategic manipulation" of the social choice system. MACKAY, *supra* note 201, at 36.

We can describe a similar procedure under *F***.[211] Even these procedures, however, might be cumbersome if set *S* includes a large number of alternatives. The difficulty of finding the optimum under *F** or *F*** in practice, however, is beside the point. Conditions of "computational convenience and simplicity" address pragmatic considerations that are not directly relevant to the issue of moral philosophy that is at stake.[212] I present *F** and *F*** simply to demonstrate that a social welfare function can, in theory, incorporate fairness concerns without violating the Pareto principle or otherwise generating preference cycles. A theoretical counterexample is sufficient to disprove the logical impossibility of constructing such a social welfare function. This procedure addresses the ethical question of what function we ought to maximize ideally, not the pragmatic question of how we might maximize that function imperfectly in practice, given the constraints imposed by the costliness of gathering and processing information.

In reality, the identification and analysis of alternatives requires an investment of scarce resources. We must devote resources, for example, to the analysis of the consequences of different policies. We must also bear costs to gather information on individual preferences regarding these consequences. The allocation of these scarce resources for policy analysis is itself a social choice subject to normative analysis. Thus, practical considerations militate in favor of a simplified procedure for policy analysis.

From a practical standpoint, it is likely that a reasonable rule of thumb for making social choices under *F** or *F*** would be simply to choose the optimum under our second-best criterion *F*, unless we can readily identify a Pareto improvement over that optimum. Suppose, for example, that *F* is the liberal consequentialism outlined in Part I. We would in general expect to identify a unique optimum under that *F*.[213] Everyone would enjoy an entitlement to this fairness optimum, which would itself include the procedure necessary to identify the optimal social choices. Deviations from this optimum would be justified under *F** or *F*** only if all individuals prefer another alternative. Given the costs of finding such a Pareto superior

211. We would first identify the alternatives that are optimal under *F*. Second, if there is more than one optimum, then we would apply a welfarist principle to rank the fairness optima and to identify a smaller subset of optima. If any of these optima are also Pareto optimal, then declare these alternatives to be our optimal social choices, much as *F** would choose an alternative in set *A*. If none are Pareto optimal, then this subset corresponds to set *C* under *F**. Move all alternatives that Pareto-dominate at least one of these alternatives in this subset to a higher ranking set, corresponding to set *B* under *F**, and repeat the entire procedure on this Pareto superior set. Continue these iterations until an optimum is also Pareto optimal.

212. Kelly, *supra* note 132, at 139.

213. The probability of an exact tie for the optimum is virtually zero.

alternative in the real world, we may expect it to be rarely worth investing much of our scarce resources in the search for such Pareto improvements.

We may expect a limited investment of resources to be unlikely to identify a Pareto improvement. At best, we might identify a limited set of possible Pareto improvements, but given the costs of discovering the preferences of each individual in the world, we would be uncertain whether any given candidate for a Pareto improvement is in fact Pareto superior to the fairness optimum. As the number of individuals in the world increases, it becomes increasingly unlikely that any given candidate is Pareto superior to the fairness optimum in reality. After all, for such an alternative to Pareto-dominate the fairness optimum, everyone must prefer this alternative over the fairness optimum (under the weak Pareto principle) or no one must prefer the fairness optimum over this alternative (under the strong Pareto principle). Given the low probability of actually choosing a Pareto improvement, the selection of some alternative other than the fairness optimum is likely to decrease social welfare under F^* or F^{**}.

Almost certainly at least one person in the world will prefer the fairness optimum over any other alternative that we can readily identify. First, given the diversity of moral views in the world, at least one person is likely to believe in criterion F as a fairness principle and thus prefer our fairness optimum over other alternatives on moral grounds. As long as we include moral preferences in our definition of the Pareto principle, these alternatives will not be Pareto superior to our fairness optimum. Second, even if we were to exclude moral preferences from our definition of the Pareto principle, it would still seem unlikely that we could readily identify a Pareto improvement over our fairness optimum, as long as our fairness principles tend to promote human welfare. The more unlikely criterion F is to rank a Pareto inferior alternative over a Pareto superior alternative, the more difficult it will be to find a Pareto improvement over the fairness optimum. Thus, as long as our fairness principles are likely to make someone somewhere better off when applied, as any reasonable fairness theory would, then anyone who seeks to maximize either F^* or F^{**} is likely to make the same social choices in practice as someone who seeks to maximize F.

Therefore, the difficulty of maximizing a more complex social welfare function like F^* or F^{**} may simply militate in favor of maximizing F instead as a less costly procedure that comes as close to maximizing the ideal function as we can in the real world. That is, in making social choices, we might use criterion F as a reasonable proxy for the type of ideal social welfare function proposed here, departing from criterion F only in those unusual circumstances when we can identify a Pareto improvement over the fairness optimum. If we are unlikely to identify an alternative Pareto superior to the alternative that appears optimal under criterion F, we may in

general use criterion *F* as the second-best rule that we apply in practice to make social choices.

Kaplow and Shavell assert that it is "simply irrelevant . . . that Pareto dominance will be rare among actual policy alternatives," because their point is that "the Pareto principle not only rules out choice of Pareto-dominated policies; it also renders inadmissible certain criteria for assessing policy."[214] My response takes seriously the theoretical significance of the Pareto principle for the question of the "criteria for assessing policy" but proposes ideal social welfare functions that not only incorporate fairness criteria but also comply with the Pareto principle. The fact that "Pareto dominance will be rare" is relevant, however, to whether we should make social choices using cruder fairness principles in practice as an imperfect but reasonable approximation of the choices we would make using the ideal social welfare function.

Furthermore, even if we can commonly and easily identify Pareto improvements over the fairness optimum under criterion *F*, a belief in either *F** or *F*** as the ideal criterion for social choice would still make the evaluation of policies under criterion *F* morally relevant. We would need to apply criterion *F* to determine the entitlements that we would allow individuals either to enjoy or to alienate. Thus, the maximization of either *F** or *F*** would preserve an important role for fairness principles in practice, because criterion *F* would specify what rights individuals can invoke to trump welfarist concerns. This analysis under fairness principles would be an essential first step in the identification of the optimal social choice under either *F** or *F***.

In contrast, the identification of the optimum under welfarist criteria would be morally irrelevant. The welfarist optimum could possibly coincide with the social optimum under *F** or *F*** in some cases, but we could not count on such a coincidence. In fact, pursuit of the welfarist optimum could lead us seriously astray and could easily produce outcomes that rank lower under either *F** or *F*** than the fairness optimum identified by criterion *F*. On those occasions when the welfarist optimum does rank higher under *F** or *F*** than does the fairness optimum under *F*, it will do so by virtue of Pareto superiority, not by any other application of welfarist analysis. Under *F**, we would have no need to apply welfarist criteria in the evaluation of policy alternatives, except to identify Pareto improvements. Thus, a belief in *F** leaves us with good reasons to concern ourselves with fairness principles and no reason at all to heed welfarist analysis beyond the application of the Pareto principle. Similarly, a belief in *F*** gives fairness principles priority over welfarism, except for the application of the Pareto principle.

214. KAPLOW & SHAVELL, *supra* note 4, at 4.

C. *A Liberal Theory of Social Welfare*

I have suggested that the liberal consequentialism described in Part I is a plausible candidate for a second-best fairness principle *F*, but it is not my objective in this Article to defend either *F** or *F*** as the ideal social welfare function derived from that *F*. If we were to adopt that consequentialism as our principle *F*, then liberal principles might call for more frequent departures from *F* than those suggested by the weak Pareto principle alone. For example, we might substitute the strong Pareto principle for the weak Pareto principle, or we might want to depart from criterion *F* in other circumstances when all those individuals with standing to complain about an alternative nevertheless favor that alternative.

For example, all those who would suffer under a particular alternative in terms of their personal preferences may nevertheless favor that alternative on the basis of their external preferences. In Sen's example, *P* and *L* prefer *p* over *l* on this basis. Alternative *p* might fail as a Pareto improvement, however, if those whose personal preferences are unaffected by that alternative exercise a veto based on their external preferences. Consider the political preferences of some third party added to Sen's example, who may veto *p* based on a belief in the type of liberalism proposed by Sen. If this veto implies that we must apply the liberal consequentialism described in Part I, then the veto would prevent *P* and *L* from waiving their rights, because, under that criterion *F*, the alternative that interferes with the satisfaction of personal preferences (*p*) would rank lower than the alternative that satisfies those preferences (*l*).

Should the trade that produces *p* go forward despite the objections of some based on their external preferences? Those who object may argue that those who favor the trade do so on the basis of external preferences and thus forfeit any right to have external preferences excluded from a calculation of the welfare effects of the trade. Both the proponents and the opponents of the trade base their positions on external preferences rather than personal preferences. The opponents of the trade would claim that their external preferences should receive just as much weight as the external preferences satisfied by alternative *p*. If we give all preferences equal weight in these cases, we would allow such trades to go forward only if they increase a measure of welfare that includes the satisfaction of all external preferences as well as personal preferences. This welfarist qualification would also prevent a series of such trades from producing a strict preference cycle, because if they go forward only when they increase

this measure of welfare, then a series of these trades cannot lead back to the origin.[215]

A Pareto improvement would be a special case of such a trade, an example that all individuals prefer. We could depart from criterion *F* not only to implement Pareto improvements but also to permit any other transaction that is both an improvement from the standpoint of welfarism and favored by all who would suffer in terms of their personal preferences. We might consider any such transaction or any series of such transactions to be a social improvement. To implement this qualification in a social welfare function, we would follow a procedure like that used to produce *F** or *F***, but we would substitute this more general notion of a social improvement for the more specific notion of a Pareto improvement.

My goal, however, has not been to identify the ideal social welfare function. I have merely sought to show that the Pareto principle does not by itself end the search for a fair or liberal social welfare function. On the contrary, reconciliation with the Pareto principle may simply be the first and most obvious refinement we would want to make in developing a social welfare function based on liberal principles.

This search for the ideal conception of social welfare is an example of the pursuit of what Rawls calls "reflective equilibrium," which one reaches "after a person has weighed various proposed conceptions," seeking those "that match our considered judgements," and "has either revised his judgments to accord with one of them or held fast to his initial convictions."[216] I mean to suggest that the type of liberal theory of social welfare sketched here may "give a better match with our considered judgments on reflection"[217] than the welfarism proposed by Kaplow and Shavell, which rejects all notions of fairness. In our pursuit of reflective equilibrium, we may more easily sacrifice features like independence and continuity in our social welfare function than fairness principles or the Pareto principle.[218] In deciding whether independence and continuity are

215. As Gibbard points out, libertarian rights may be inconsistent with one another, and we have to qualify these rights in some way to ensure a system of rights that is consistent with itself, even before we consider the conflict between these rights and the Pareto principle. Gibbard, *supra* note 120, at 388-97. Given that "[t]he problem of *internal* consistency of the kind with which Gibbard is concerned arises only with preference configurations requiring rather other-oriented motivations," Sen agrees that "the weakening of these 'rights' in the presence of other-oriented motivations would not seem to involve any great violations of libertarianism." Sen, *supra* note 2, at 235. Christian Seidl explores the necessary conditions for liberalism, which he takes to be a system that allows any "*coalition*" to do what it wants such that "*the joint outcome can actually be put into action*." Christian Seidl, *On Liberal Values*, 35 ZEITSCHRIFT FUR NATIONALÖKONOMIE 257, 279 (1975).

216. RAWLS, *supra* note 53, at 48.

217. *Id.* at 50.

218. Any Paretian social welfare function that incorporates fairness notions must violate independence or continuity conditions or both over an unrestricted domain. *See* Roberts, *supra*

necessary conditions for a plausible social welfare function, we must consult our moral intuitions. Given the defects of welfarism as a moral theory, we may find that a fairness theory on balance yields a better fit with our moral intuitions, and we might reasonably opt for the fairness theory. I have outlined a solution that I suggest "moves us closer to the philosophical ideal," but may not, "of course, achieve it."[219]

V. CONCLUSION

I have shown that the Pareto principle by itself does not imply that we must abandon principles of fairness, including principles of liberal toleration. A theory of fairness can incorporate the Pareto principle, and indeed can generate a complete ranking of alternatives, much as a utilitarian social welfare function can. A fairness theory that complies with the Pareto principle would never sacrifice the interests of all individuals and is thus not vulnerable to the charge that it stands in opposition to human welfare. Such a fairness theory would apply fairness principles only to resolve conflicts between individuals, so that these principles would always serve the interests of at least one person. Fairness principles under this theory would address the just distribution of benefits and burdens in society.

I have not addressed precisely what principles a theory of fairness should include. My goal in this Article has been merely to demonstrate that a plausible theory can capture our strong moral intuitions regarding liberal toleration and also respect the Pareto principle. Utilitarians need stronger assumptions than the Pareto principle to preclude the incorporation of fairness concerns in a social welfare function.

note 175, at 428 (showing that imposing independence and weak continuity conditions along with the weak Pareto principle over an unrestricted domain implies a welfarist social welfare function).

219. RAWLS, *supra* note 53, at 50.