Fairness versus Welfare: Notes on the Pareto Principle, Preferences, and Distributive Justice

Louis Kaplow and Steven Shavell

ABSTRACT

In *Fairness versus Welfare*, we advance the thesis that social policies should be assessed entirely on the basis of their effects on individuals' well-being. This thesis implies that no independent weight should be accorded to notions of fairness (other than many purely distributive notions). We support our thesis in three ways: by demonstrating how notions of fairness perversely reduce welfare, indeed, sometimes everyone's well-being; by revealing numerous other deficiencies in the notions, including their lack of sound rationales; and by providing an account of notions of fairness that explains their intuitive appeal in a manner that reinforces the conclusion that they should not be treated as independent principles in policy assessment. In this essay, we discuss these three themes and comment on issues raised by Richard Craswell, Lewis Kornhauser, and Jeremy Waldron.

In *Fairness versus Welfare* (Kaplow and Shavell 2002, hereinafter *FVW*), we advance the thesis that social policies, notably, legal rules, should be selected entirely with regard to their effects on the well-being of individuals. Accordingly, notions of fairness, such as corrective and retributive justice, should receive no independent weight in policy assessment. Our argument is based on the perverse effects on welfare of pursuing notions of fairness, other problematic aspects of the notions, notably, their lack of rationale, and a reconciliation of our thesis with the existence of moral intuitions that seem to favor the notions. Each of these

LOUIS KAPLOW and STEVEN SHAVELL are Professors at Harvard Law School and Research Associates at the National Bureau of Economic Research. We thank the John M. Olin Center for Law, Economics, and Business at Harvard Law School for financial support. We have benefited from exchanges with Richard Craswell concerning his comment and from the research assistance of Zachary Price.

Journal of Legal Studies 32 (January 2003)
 0047-2530/2003/3201-0012$01.50

themes is developed in general terms and in detailed analyses of leading notions of fairness in the areas of torts, contracts, legal procedure, and law enforcement.

In this symposium, Richard Craswell questions our demonstration that endorsement of any notion of fairness sometimes reduces every individual's well-being (that is, violates the Pareto principle). Jeremy Waldron, while commending our incorporation of distributive concerns in the concept of welfare that we defend, voices some disagreement with our treatment of the subject of distributive justice. And Lewis Kornhauser presents arguments concerning our inclusive definition of well-being.

In Section 1, we review the terminology and the main themes of our book. Then, in Sections 2–4, we address the main points of Craswell, Waldron, and Kornhauser, explaining that none of their claims pose significant challenges to our thesis even if they are valid and also that their arguments reflect misinterpretations of our views or are otherwise mistaken. In Section 5, we offer concluding remarks.

1. FAIRNESS VERSUS WELFARE

In this section, we begin by defining what we mean by welfare and notions of fairness because these terms are central to understanding our thesis that policy assessment should be based only on welfare, with no independent weight given to notions of fairness. Then we summarize the three central claims that we offer in support of our thesis.

1.1. Welfare and Fairness

Welfare.[1] Under a welfarist approach to policy assessment, one first determines how a policy affects each individual's well-being and then makes an aggregate (distributive) judgment based exclusively on this information pertaining to individuals' welfare.

The conception of individuals' well-being that we consider, in the tradition of welfare economics, is a comprehensive one. It encompasses not only the direct benefits that individuals obtain from the consumption of goods and services, but also individuals' degrees of aesthetic fulfillment, their feelings for others, and anything else that they value. What factors are included in well-being—and with what weight—is understood subjectively,

1. See *FVW*, chap. 2, sec. A; and chap. 8, sec. B.

in terms of what actually matters to individuals. An implication of our broad definition is that even tastes for fairness are included: Just as an individual might derive pleasure from art, nature, or fine wine, so might an individual feel better with the knowledge, for example, that vicious criminals receive their just deserts. This view, under which tastes for fairness are counted with a weight to be determined empirically, based on the actual weight, if any, that individuals place on such tastes, must be sharply distinguished from the view of notions of fairness as independent evaluative principles, which is the subject of our critique.[2]

Given this definition of well-being, our advocating that policy assessment be based exclusively on welfare is equivalent to embracing the moral position that policy choice should depend solely on concerns for human welfare. Furthermore, this characterization of our thesis suggests why we choose the inclusive definition of well-being that we do. Ultimately, the moral force of welfarism is grounded in what matters to individuals, and what really matters to individuals is not to be determined on the basis of an analyst's personal judgment, however derived, of what others should or should not value. Nevertheless, little of our analysis depends on this specification of individuals' well-being; as we emphasize in our book, if one favors a different view of well-being, our main arguments imply that policy assessments should be made solely with regard to how policies affect well-being thus construed, with no independent weight given to notions of fairness (see, for example, *FVW*, pp. 23–24 and n. 14, p. 409).

As we noted at the outset, welfare-based policy assessments also require a distributive judgment. We offer a number of remarks on the subject in our book, the most important being that our thesis is independent of how such judgments should be made (see, for example, *FVW*, chap. 2, secs. A.2 and A.3; chap. 3, sec. C.2.e; chap. 4, sec. D.1; and chap. 8, sec. C). The reason is that our thesis is addressed to a prior question, whether policy assessments (including distributive judgments) should be based exclusively on individuals' well-being or also (or instead) on factors that are independent of individuals' well-being.

Fairness.[3] By notions of fairness we include all principles—whether stated in terms of justice (such as corrective or retributive justice), rights (such as a right to a day in court), or cognate concepts (such as the

2. For discussions of issunes concerning possible differences between individuals' preferences and their actual well-being as well as matters involving possible changes in preferences, objectionable preferences, and tastes for fairness, see *FVW*, chap. 8, sec. B.

3. See *FVW*, chap. 2, sec. B.

sanctity of promises)—that may be employed to assess the desirability of policy and that have the following characteristic: At least some weight is given to factors independent of individuals' well-being. That is, we define notions of fairness to include all independent evaluative principles that are not purely welfarist.[4]

Our purpose in adopting this definition is one of convenience. Each of our three main themes, as well as the particular arguments relating to them, distinguishes between approaches that are based solely on individuals' well-being and those that are not. Hence, we find it useful to employ a single, familiar term (one that seems to have no precise, canonical meaning) to refer to all principles that are, in some respect, other than welfarist.

Four points deserve emphasis. First, most notions of fairness are nonconsequentialist: Characteristically, one examines particular features of situations to determine what outcome is most fitting according to a given principle of fairness. For example, if A wrongfully injured B, then B should be compensated by A; if A's action toward B would break a promise, then it is impermissible; if the true nature of the crime was X, then the punishment should be $P(X)$. In particular, whether A should compensate B, A may break a promise, or the punishment should be $P(X)$ does not depend exclusively (or at all) on an assessment of the consequences of doing these things, such as the deterrence of undesirable behavior.

Second, most fairness proponents, including most modern policy analysts who grant importance to notions of fairness, hold mixed normative views, under which some weight is accorded to conceptions of fairness and some to welfare. (At a minimum, most who endorse various nonconsequentialist fairness principles would not adhere to them if the adverse consequences would be extreme.) Our definition of fairness includes such mixed views, and our objections apply to them to the extent that weight is given to considerations other than welfare.

Third, sometimes notions of fairness may be invoked without meaning to contradict a purely welfarist view: A notion of fairness (perhaps a purely distributive notion) might refer exclusively to effects on individuals' well-being, and some notions of fairness may be used as proxy criteria when the concern really is welfarist (requiring that wrongdoers pay for harm could be favored solely on deterrence grounds).

4. For a formal statement of the difference between welfare and fairness, as we define the terms, see *FVW*, p. 24 n. 15, and p. 39 n. 52; and Kaplow and Shavell (2001, p. 283).

We have no per se objection to the use of notions of fairness as stand-ins for welfare since our concern is not semantic (although as a practical matter it is often best for policy analysis to address our actual objectives directly and explicitly). We observe, however, that all of the leading notions of fairness that we analyze in our book are generally offered not as proxy criteria for welfare but rather as independent principles of evaluation.

Fourth, although we object to the use of notions of fairness as independent evaluative principles, we do not necessarily object to arguments for one or another normative principle that happen to be couched in the language of fairness. Accordingly, our analysis makes no claim about whether an argument that may sound in fairness (or justice or related terms) is thereby untenable. Rather, our claim is that the ultimate criterion for policy assessment should be one that is based exclusively on welfare. To be sure, claims that one or another outcome is unfair are often unhelpful because they convey little information beyond the fact of the author's condemnation. Nevertheless, the object of our book is to show that arguments, however articulated, favoring policy assessment based on fairness principles are deficient, whereas arguments favoring welfare-based assessment are compelling, whether or not phrased as such.

1.2. Conflict between Fairness and Welfare

The first major theme of our book is that pursuit of notions of fairness results in a needless and, at root, perverse reduction in individuals' well-being. That advancing notions of fairness reduces well-being is, as we clearly state in our book (*FVW*, p. 7), a tautology on a general level: Because we define fairness principles as those that accord weight to factors independent of well-being, whenever fairness and welfare assessments differ, it must be that advancing fairness reduces overall well-being.

We nevertheless emphasize this basic conflict for two reasons. First, the depth of the tension between fairness and welfare is not widely appreciated; indeed, in policy analysis that rests on notions of fairness, it usually is not even mentioned. Second, by examining in detail a variety of concrete, paradigmatic settings that lie at the core of the domain of fairness principles, the true nature of the conflict is revealed. One is able to see what it is about leading notions of fairness that makes individuals worse off, and one is thereby better able to assess (see also Section 1.3) what, if anything, may be said to justify sacrificing human well-being.

Reinforcing the second point is that, in each of the situations that we examine, we present the striking argument that promoting any of the pertinent notions of fairness sometimes makes literally everyone worse off, that is, violates the Pareto principle.[5] One way of demonstrating this argument focuses on symmetric cases. For example, in the tort setting, suppose that each individual is equally often an injurer and a victim (and is identical in all other respects, such that the harm that might be caused, the cost of precaution, and so forth are the same for everyone). Further, suppose that there arises a situation in which a favored notion of fairness conflicts with welfare. For instance, a notion of corrective justice might be held to favor the negligence rule (because the notion holds that wrongdoers, and only wrongdoers, should pay) even though a rule of strict liability results in higher welfare (say, because it better controls injurers' activity levels).

Now it should be clear that, in a symmetric case, advancing any notion of fairness will always make everyone worse off when it conflicts with welfare. After all, in a symmetric case, everyone is identically affected, so in comparing two regimes, everyone must be better off under one regime than under the other.[6] A welfare-based analysis, of course, favors the regime under which everyone is better off. Hence, if a notion of fairness conflicts with welfare, it must be that it favors the regime in which every individual's well-being is lower.

This generic symmetric-case demonstration is supplemented by concrete examples and by other demonstrations of the Pareto conflict, some specific to particular contexts and one formal, technical proof that is quite general.[7] Even though as a practical matter it will rarely if ever be the case that one of two policies under serious consideration will literally make everyone better off than the other—a point we have emphasized from the outset of our work on the question (Kaplow and Shavell 1999, pp. 72–74; 2001, pp. 284–85; *FVW*, pp. 55–58 and n. 78)—the result

5. We refer to the weak Pareto principle, which holds that if everyone is strictly better off under one policy than under another, the former should be deemed superior.

6. We are ignoring the uninteresting case in which welfare is identical under the two regimes.

7. We first published the symmetric-case result in Kaplow and Shavell (1999). In *FVW*, the symmetric-case argument appears in general terms in chap. 2, sec. C.1; in the tort setting in chap. 3, sec. C.1.e; in the procedure setting in chap. 5, secs. A.5 and B.3–B.4; and at various other points. We use other demonstrations as well, such as in the contract setting in chap. 4, secs. C.1.f and C.2.e. Our general, formal proof appears in Kaplow and Shavell (2001). For further discussion, see *FVW*, chap. 2, sec. C.1 (especially the notes); and Kaplow and Shavell (2000).

that all notions of fairness sometimes make everyone worse off is of great significance regarding the soundness of these notions as policy-making criteria. This is true for a number of reasons.

First, even most proponents of notions of fairness will find it deeply troubling that adherence to their principles entails endorsing the principle that sometimes it should be deemed socially desirable to make everyone worse off. Indeed, one can ask to whom one is being fair if every individual is made worse off. Furthermore, proponents of notions of fairness who ground the notions in ideas of freedom and autonomy should have particular difficulty with our demonstration since individuals would unanimously reject a fairness notion if it makes them all worse off.

Second, there is an important matter of logical consistency. If indeed a principle is shown to be deficient, one cannot consistently adhere to it on the ground that the case in which its deficiency is glaringly apparent is not the case one is considering at the moment. This point about logical consistency, which is a staple of argument in moral philosophy, has all the more force because the cases in which we demonstrate the Pareto conflict in our book—often symmetric cases—are simple, basic, paradigmatic, clear cases in which notions of fairness apply. Our judgments about such cases should carry more weight, decisive weight, compared with our judgments in cases with many potentially conflating factors, in which it is difficult to reach conclusions with confidence.

Third, as we elaborate in our book and elsewhere, our demonstration that all notions of fairness always make everyone worse off in the symmetric case is especially significant under many broadly endorsed normative frameworks (see, for example, Kaplow and Shavell 1999, pp. 73–74; *FVW*, chap. 2, sec. C.1; chap. 3, sec. C.1.e.iii). We show that the Golden Rule, Kant's categorical imperative, and the veil-of-ignorance construct each require that normative principles be tested as if one is in a symmetric setting. (The reason, in brief, has to do with the need for impartiality: One could, for example, favor "might makes right" as a general rule if one were unusually mighty, but if one is forced to assume that just as often someone else will be mightier, one would reject such a rule.) Accordingly, if one adheres to any or all of these normative frameworks—as most fairness proponents in fact do—one is forced either to reject all nonwelfarist principles or to endorse the view that making everyone worse off should be the core feature of any sound normative principle.

Although we regard the conflict between notions of fairness and the

Pareto principle as powerful, we also should emphasize that much of our book's analysis of the broader conflict between fairness and welfare does not focus on this particular argument. The second part of our book—chapters 3–6, constituting nearly two-thirds of the total—examines how fairness reduces welfare in detail and often without regard to the Pareto principle. For example, chapter 3, which is on torts, fully considers nonreciprocal cases in which Pareto conflicts do not typically occur, and chapter 6, which is on law enforcement and retributive justice, the longest chapter in our book, makes almost no reference to the Pareto principle in arguing that notions of fairness are perverse with regard to their implications for individuals' well-being (for example, the only beneficiaries of a more fair system may be criminals who escape punishment and thus profit from their crimes). That is, we rely not only on an abstract, if extremely important, general argument, but also on particular analysis of how leading notions of fairness play out in important legal settings.

1.3. Further Deficiencies in Notions of Fairness[8]

Our second central theme in *FVW* concerns largely internal deficiencies in notions of fairness themselves. Perhaps the most important is their lack of rationale. We develop this theme by considering in great detail justifications that have been offered both by moral philosophers and by legal scholars. In addition, with regard to the scenarios in each of the legal contexts that we examine in depth, we further consider what warrant might be offered for giving weight to notions of fairness. We consistently find that none of the possible rationales is convincing.

For example, upon reviewing the literature supporting retributive justice, from Aristotle to Kant and Hegel to moderns, it is difficult to identify the affirmative case for the principle. Instead, one finds, for example, reference to the need to restore some sort of moral balance in the world, more a conclusory metaphor than anything else. When one adds that even retributivists would not punish all wrongs (most lies, for example) and that, more broadly, different theories with different metaphors are applied in different contexts (even though, on their face, many of the theories and metaphors apply to the other contexts as well), it becomes difficult to imagine how plausible justifications could be offered for these theories.

8. For a summary of many of our points, see *FVW*, chap. 2, secs. B.2 and C.2. Fuller statements appear throughout chaps. 3–6.

We also identify a number of difficulties that, although not necessarily inherent in notions of fairness, seem to be endemic. There are matters of definition, often so serious as to leave basic statements of fairness notions highly incomplete if not entirely empty. Corrective justice purports to answer questions of tort law by holding that wrongdoers should compensate victims, but then a theory of wrongdoing must be supplied and it is that theory that ultimately indicates who should compensate whom. Likewise for promise-keeping (because promises must be interpreted) and for retributive justice (because the wrongfulness of acts that vary on many dimensions must be measured with a common denominator and equated, in some often unspecified proportion, to measures of punishment). We further observe that such fundamental problems in stating and applying the principles are probably related to their lack of affirmative rationale; it is difficult to imagine that, if there were readily identifiable reasons for pursuing the principles, we would have so little idea of what they mean.

In addition, many notions of fairness involve adopting an ex post perspective, asking, for example, what punishment is appropriate given that a crime has been committed and the criminal apprehended and convicted. This perspective tends to undervalue or ignore other outcomes, including more likely ones, such as the fact that for many crimes most criminals go scot-free. In a related way, behavioral effects, such as whether crimes are committed, are downplayed if even considered at all. Principles that require incomplete assessments of situations are unlikely to lead to sound policy choices. (One of our subthemes is that notions of fairness often serve as proxy indicators of welfare, which in turn helps to explain their appeal. Indeed, in our analysis of leading fairness principles, we often find that, with regard to aspects of situations that the principles do lead one to consider, they often point toward welfare-relevant effects; moreover, when the principles conflict with welfare, it is often precisely on account of the welfare-relevant factors that the principles lead the analyst to ignore.)

Yet another problem arises from the nonconsequentialist nature of most notions of fairness. A common repercussion is that a rule can be deemed more fair even though it results in more unfair outcomes or a greater incidence of the behavior whose wrongfulness underlies the motivation for the theory. For example, we show how insisting on the fair punishment for, say, murder, can result in more actual instances of unfair punishment of murderers and in a greater number of innocent people mistakenly accused of murder, and also in more murder, which retrib-

utive theory deems to be a serious wrong such as to demand punishment in the first instance.

These points—especially that concerning the lack of affirmative rationale for notions of fairness—are difficult to capture well in this recapitulation. Their development occupies much of our book and a disproportionate share of the extensive notes and references. The extent of these deficiencies is notable in view of the centuries of attention given to the notions of fairness by scholars, the vast majority of whom are proponents of the notions. In summary, notions of fairness are quite difficult to defend even aside from the adverse, often perverse, implications of pursuing notions of fairness for human welfare.

1.4. Social Norms and the Reconciliation of Fairness's Appeal with Welfarism

The third major theme of our book involves providing an answer to the question of how it can be that notions of fairness appeal to our moral instincts and intuitions and yet should not be given any independent weight in policy assessment. An important part of our answer, which is social-scientific in nature, involves social norms.[9]

First, we observe that principles of fairness tend to correspond to internalized social norms, such as keeping promises and holding wrongdoers accountable. These social norms are commands that people want to obey because the maxims have been inculcated or are inborn. Thus, one might feel guilty for telling a lie (and also may fear social disapprobation). Social norms appeal to us both because of their internalization and related social reinforcement and because they valuably guide our behavior and curb opportunism in everyday life. Given their internalized character and instrumental value, it is not surprising that individuals who engage in policy analysis, themselves well-socialized members of society who attach importance to these social norms, will be inclined to accord weight to corresponding notions of fairness.[10] For example, our attachment to the social norm that promises should be

9. See *FVW*, chap. 2, sec. D, for explanation of the argument in general terms; chap. 3, sec. E; chap. 4, sec. C.2.g; chap. 5, secs. A.6 and B.3.f; and chap. 6, sec. D, for development of the argument in each of the legal contexts that we examine; and these sections as well as chap. 2, sec. B.2.c, and chap. 7, sec. B.1, for additional explanations for the appeal of fairness principles that are consistent with our thesis.

10. This explanation also indicates why individuals may have a taste for notions of fairness in the sense discussed in Sec. 1.1.

kept naturally disposes us to favor the promise-keeping notion of fairness when we assess contract law.

Second, we examine the implications of this phenomenon. Most important, it provides an explanation for the attraction of notions of fairness, but an explanation that offers no justification for according the notions independent evaluative weight. That individuals are in a sense programmed to conduct their everyday lives in accordance with social norms does not warrant elevating these norms to the status of independent evaluative principles in the qualitatively different context of legal policy design. Moreover, given that the raison d'être of social norms is functional, to promote welfare, it would be an ironic mistake for the analyst to treat them as if they were independent principles to be pursued at the expense of well-being. Indeed, in our book we repeatedly show that divergences between the prescriptions of notions of fairness and those of welfare-based analysis can be traced to differences between the realm of everyday life, the appropriate domain of social norms, and that of regulation through the apparatus of control of the modern state, where we argue that analysis should be based exclusively on welfare.[11]

Thus, if we are self-conscious about the role of social norms and the corresponding origins of our instincts and intuitions about notions of fairness, we would not be led to attach independent weight to notions of fairness when assessing policy.[12] This theme, like our other two principal themes, is developed both in general terms (in this instance drawing on a range of literatures in the social and natural sciences) and in great detail in chapters 3–6 in each of the legal contexts that we examine and with respect to each of the leading notions of fairness that we consider.

Our first two themes—involving how giving importance to notions of fairness leads to needless sacrifices in our well-being, and the lack of affirmative rationale for (and other difficulties with) notions of fairness—indicate why our normative thesis that policy assessment should be based exclusively on considerations of individuals' well-being is cor-

11. Relatedly, this relationship between functional social norms and notions of fairness explains why notions of fairness tend to have the aforementioned proxy characteristic that they tend to indicate some respects in which policies promote welfare (but they also tend to be incomplete and sometimes misleading in large part because important context differences render fairness principles imperfect proxies).

12. As should be clear from the text and as we emphasize in the book, this argument demonstrates that policy analysts should not be guided by notions of fairness but in no way indicates that individuals in everyday life should cast aside social norms. See chap. 7 of *FVW* for discussion of the differing implications of our analysis for ordinary individuals, policy analysts, and government decision makers.

rect. Our third theme is complementary in that it reconciles our thesis with widely held moral instincts and intuitions and, in a related way, it shows why contrary moral arguments, substantially grounded in such instincts and intuitions, should not be seen as posing a real challenge to our thesis.

2. CONFLICT WITH THE PARETO PRINCIPLE

Richard Craswell (2003) suggests that our argument that pursuing all notions of fairness sometimes makes everyone worse off might be avoided under certain "hybrid" fairness theories of the following sort: Apply the initial fairness theory in choosing between two regimes unless this would make everyone worse off, in which case abandon the fairness theory and instead use a purely welfare-based assessment, which of course would favor the regime under which everyone is better off.[13] For reasons given in our original articles on the subject, in *FVW*, and in our reply to Howard Chang, who previously advanced a similar view,[14] this attempt to circumvent our argument fails.[15] The possibility of such hy-

13. Craswell also cites to similar effect Chang (2000a) and an unpublished manuscript by Barbara Fried, "Can We Really Deduce Welfarism from the Pareto Principle" (which we have not read, although we have seen an earlier version). We do not address here most of the second section of Craswell's comment, which consists of his own version of aspects of some of our arguments regarding our first theme.

14. See Chang (2000a, especially secs. 3 and 4); see also Chang (2000b).

15. See, for example, Kaplow and Shavell (1999, pp. 72–74, including nn. 20 and 23); *FVW*, chap. 2, sec. C.1 (including nn. 75–76 and 78–80); *FVW*, chap. 3, sec. C.1.e; and Kaplow and Shavell (2000) (replying to Chang). We observe that, although Craswell advances an argument very similar to Chang's and makes numerous references to Chang's article, he refers to our reply to Chang only twice, both times with regard to points he raises in footnotes. See Craswell (2003, p. 255 n. 13, and p. 259 n. 18). We also do not understand why Craswell asserts that in our "initial presentation" of our Pareto argument, we "do not even address hybrid theories," but rather "limit [our] initial argument to fairness theories in which unfairness is given a constant weight or a weight that is independent of whether any victims of the alleged unfairness are made worse off" (Craswell 2003, p. 253). To be sure, we occasionally, for ease of exposition, offered examples with the latter quality, but virtually all of our analysis, including our particular discussion of mixed views (under which both fairness and welfare may receive weight), is quite general. Thus, our first published paper, Kaplow and Shavell (1999), never makes the posited restriction and specifically addresses hybrid theories. See Kaplow and Shavell (1999, p. 72 n. 20) (which Craswell himself later quotes, at p. 255). Our formal article, Kaplow and Shavell (2001), explicitly considers any consistent notion of fairness, allowing weight to vary in any manner as long as it is not discontinuous (which intentionally rules out hybrid theories on the ground that they are incoherent). *FVW* defines fairness in a general manner that includes both hybrid theories and theories under which the weight given to fairness may vary, even

brid fairness theories does not, upon reflection, fundamentally challenge our thesis. Moreover, as moral theories, the hybrid schemes are incoherent because of what turns out to be their inability to make consistent choices among alternative regimes and also on account of their discontinuous nature.

Consider first the relationship between the hybrid fairness theories and our thesis. As we explain in Section 1, the Pareto argument—despite its importance—relates to but one of three themes in *FVW*, and with regard to that theme actually occupies only a part, often a small part, of our analysis.[16] (As noted in Section 1.2, for example, in our longest chapter, on law enforcement and retributive justice, we give extensive attention to how adherence to retributive justice leads to perverse sacrifices of welfare, while we make virtually no reference to the Pareto argument.)

More directly, our elaboration of the significance of the Pareto conflict, as summarized in Section 1.2, makes clear that Craswell's suggested hybrid theory is largely nonresponsive. As we explained there and emphasized in our book, the import of the Pareto conflict has nothing whatsoever to do with it arising frequently in practice (it does not) or in any particular case under consideration. Rather, all of our arguments have to do with the implications of the conflict for choosing normative criteria. That most proponents of notions of fairness should find the conflict with the Pareto principle disturbing Craswell concedes; indeed, this motivates his consideration of hybrid theories under which fairness is trumped by welfare in those rare cases in which Pareto conflicts arise. But if the initial fairness theory is such that it would indeed sometimes deem making everyone worse off to be socially desirable (morally correct), does this not suggest that there is something fundamentally wrong with the theory, rather than merely signify a trivial blemish to be cured through a bit of fine-tuning?

---

discontinuously (see, for example, *FVW*, chap. 2, sec. B.1), and explicitly addresses hybrid theories (see, for example, *FVW*, pp. 53–55 and n. 76, which Craswell also cites later, at p. 257 n. 16).

16. Craswell suggests otherwise. For example, he asserts that our argument about conflict with the Pareto principle "receives most of the emphasis" in our work (Craswell 2003, p. 245; see also pp. 246 and 273). He further states that "we do not make . . . [the substantive problems] the focus of our critique" (Craswell 2003, p. 257, brackets and words therein in original). Yet "substantive problems" are his words, not ours. As the prominence of chap. 2, secs. C.2 and D, in *FVW* indicates (as well as our preface, pp. xviii–xx), we regard all three themes to be central; indeed, what Craswell says we do not emphasize in fact constitutes the substantial majority of our book.

Furthermore, we emphasize that the cases in which we identify conflicts with the Pareto principle are simple, basic, paradigmatic, clear cases—ones in which we can be confident of our normative judgments—in contrast to more complex settings where possibly confounding factors make normative assessment more difficult, contestable, and prone to error. Given that this is so, the idea embedded in Craswell's hybrid theories, that we should virtually always follow the questionable guidance derived from more opaque cases while only rarely following the contrary confident lesson from the clear cases, seems backwards. As John Rawls (1980, p. 546) has stated, "[A] theory that fails for the fundamental case is of no use at all."

Another, related problem is the matter of logical consistency, which has long been a primary ingredient of moral argument.[17] Craswell's proposed hybrid schemes would have the entire basis for normative assessment shift depending on whether a Pareto conflict happens to arise. Additionally, as we examine further below, whether or not a Pareto conflict arises can be a matter of a 1-cent (indeed, one-billionth of 1 cent) difference in outcome to a single person. It is as if Craswell were proposing a moral theory that applied entirely different principles on Tuesdays, requiring engineers to develop ever more precise recording devices for all human activity, linked to atomic clocks, because whether an event occurs a nanosecond before or after the stroke of midnight could fundamentally alter how the social response should be determined.[18]

Additionally, our demonstration of the Pareto conflict in all symmetric cases is telling for hybrid theories. Craswell dismisses the significance of this demonstration because fully symmetric cases rarely arise in practice.[19] But, once again, our point was never that Pareto conflicts or symmetric cases occur frequently; we clearly stated that they do not (as noted in Section 1.2). Instead, we explained that prominent moral

17. See, for example, *FVW*, pp. 56–57 n. 78; and Kaplow and Shavell (2000, p. 244).

18. One could argue that this hybrid approach is not a melding of two inconsistent theories but a single, conditional theory. That is, one could have a single "apply X except when it is Tuesday in which case apply not-X" moral theory. Yet the underlying inconsistency of the criteria for normative assessment would remain despite calling it one theory rather than an admixture of two conflicting theories.

19. After briefly stating our argument in the symmetric case, he refers to it as a "special case of rules that affect everyone in society identically. Most rules, however, do not affect everyone in society identically." Then, he proceeds without further comment to consider our more general argument outside the symmetric context (Craswell 2003, p. 247).

frameworks—the Golden Rule, categorical imperative, veil of ignorance—all require that every moral principle be tested as if one is in the symmetric case; it is usually a hypothetical case, but one that is ideally suited for discerning which moral principles are correct. Hence, for those not ready to discard, indeed reverse, all of these teachings, every notion of fairness must be rejected.[20]

In sum, Craswell's hybrid theory does not avoid the force of our arguments based on the conflict between notions of fairness and the Pareto principle. Independently, we now explain how the sort of hybrid schemes that Craswell proffers are internally incoherent as moral theories.[21] With regard to this part of our argument, we observe that Craswell (like Chang) is really responding not to *FVW*, but rather to a short, technical, economics journal article containing our general demonstration (that is, without regard to symmetric settings) of the conflict between all notions of fairness and the Pareto principle. Furthermore, Craswell (like Chang) offers an informal, not entirely specific or consistent counterexample to our formal analysis and proof (the validity of which is not contested). For present purposes, we will focus on our two most pertinent objections that, although technical, are of decisive importance.[22]

20. To elaborate, Craswell's hybrid theories give no weight to their underlying fairness principles whenever they would conflict with the Pareto principle, and we demonstrate that such conflicts always arise when fairness is decisive in a symmetric setting. Therefore, in symmetric settings, Craswell's hybrid theories can never give any weight to fairness. Accordingly, if principles for all asymmetric settings are to be derived from what principles should govern in symmetric ones, it follows that one should never give decisive weight to a notion of fairness in any asymmetric setting either.

21. It is unclear the extent to which Craswell disagrees with us regarding these arguments. Often, he merely seems to suggest that various views are plausible and that our particular position really involves "substantive" arguments (Craswell 2003, pp. 256–57). We see the arguments to follow in the text as going more to whether a hybrid moral theory is coherent or even can be said to exist in any meaningful sense rather than to what is ordinarily meant by substantive moral argument between competing theories, but ultimately the question of categorization is purely semantic.

22. A third problem is that it is unclear whether Craswell's hybrid theories, whatever their other infirmities, even count as nonwelfarist theories. He refers to our observation in our technical article (Kaplow and Shavell 2001, p. 283) to the effect that under any nonwelfarist theory it must sometimes be true that there will exist two regimes that are judged differently even though every individual's level of well-being is identical under each regime. Craswell states that this description "does not include any hybrid fairness theories"(Craswell 2003, p. 254 and n. 11). In other words, if welfare levels are identical in two regimes and one regime is in all possible respects grossly unfair compared to the other, under Craswell's hybrid "fairness" theories he is implicitly asserting that he requires a judgment of indifference. However, some reflection reveals that such a theory really is a

First, if any hybrid fairness theory (or any other theory) is even to be considered as a candidate for an ideal normative criterion for policy assessment, it must be capable in principle of making consistent choices among possible regimes. As it turns out, the sorts of hybrid theories that Craswell offers to circumvent our Pareto argument do not meet this basic requirement. In our reply to Chang, we emphasized that none of the informal examples that Chang had offered us over the course of a year—including Chang's hybrid theory, which is essentially the same as Craswell's—in fact succeeded in providing a logically coherent construct that both accorded some decisive weight to a notion of fairness and avoided conflict with the Pareto principle (Kaplow and Shavell 2000, p. 246 and n. 24). Craswell's comment, in essence, offers an informal restatement of one of Chang's unsuccessful examples.

To illustrate the difficulty, suppose that there are three regimes, A, B, and C. Under a posited notion of fairness, A is perfectly fair, B is moderately unfair (say five individuals are treated somewhat unfairly), and C is significantly unfair (an additional 10 individuals are treated quite unfairly). Under a pure version of the notion of fairness, the regimes would be ranked A, best; B, second; and C, worst. But now suppose that the welfare of every individual in regime C is somewhat greater than it is in regime A (because some other aspect of the regime sufficiently benefits those treated unfairly in C). Under the hybrid approach, one is therefore compelled to hold that regime C is definitely morally superior to A. The problem, however, is that the same hybrid theory insists that regime A is definitely morally superior to regime B and that regime B is definitely morally superior to regime C.[23] And if one adheres to basic logic, it follows a fortiori from these two judgments that regime A is definitely morally superior to regime C. But a moment ago we noted

purely welfarist theory, as we explain (Kaplow and Shavell 2000, p. 241 n. 10). (Briefly, under a purely welfarist theory, it is sufficient to know every individual's welfare level to choose among policies, and Craswell's statement means that, for any configuration of individuals' welfare levels, there corresponds a unique social assessment that is independent of any nonwelfare factor, notably, any pertaining to a notion of fairness.) As we explain in note 24, however, the scheme that Craswell articulates in subsequent correspondence does not yield coherent choices among possible regimes; since it is thus a nontheory, it really cannot be classified under our welfarist-nonwelfarist dichotomy, which was meant to apply only to theories capable of yielding consistent normative choices.

23. We are assuming that there is no Pareto relationship between regimes A and B or between regimes B and C. It is obviously easy to construct such cases (simply have equal distributions in A and C but a somewhat unequal distribution in B such that at least one individual is better off than in A and C and another individual is worse off than in A and C).

that the hybrid theory necessarily deems regime C to be definitely morally superior to regime A, in order to avoid the Pareto conflict.

This failure of logical consistency—which is almost identical to one of the flaws we emphasized in our reply to Chang (Kaplow and Shavell 2000, pp. 243–45) and which Craswell does not address in his comment[24]—renders Craswell's hybrid theory a nontheory. One cannot determine what it favors even in principle. Once contradiction in normative analysis is embraced—something can be both morally superior and morally inferior, all depending on the order in which one chooses to think about the possibilities—it is really the case that anything goes.

Second, Craswell acknowledges that his proposed hybrid scheme is discontinuous in that it switches assessment criteria in response to small shifts in outcome. (This acknowledgment, as Craswell recognizes, is important because our proof shows that if an assessment method is continuous, it necessarily follows that it conflicts with the Pareto principle if weight is ever given to fairness, and this is so however much one might attempt to modify the notion of fairness, as long as there remains even a single case in which fairness receives any weight.) Craswell suggests that fairness proponents should not be bothered by such discontinuity, but we believe that once its meaning is fully grasped, this feature will be recognized to be wholly unacceptable.[25]

24. In private correspondence after we received Craswell's comment, we pressed this point, and he responded by offering a more precisely articulated (and subtly different) version of the hybrid theory, accompanied by careful analysis that itself demonstrated a similar sort of intransitivity. In particular, it had cases in which regime A is definitely morally superior to regime B, B definitely morally superior to C, but not having A definitely morally superior to C. (Specifically, moving from A to C may be immoral, even though it would be commanded to move from A to B if given that choice, and from B to C if given that choice.) In addition, this version had cases in which moving from A to B would be a definite moral improvement, and so would be moving from A to C, but B and C cannot be morally compared, rendering the theory conceptually indeterminate. (More precisely, whichever regime one "moves" to first, B or C, there one must stay; yet these moves are mere mental exercises, so it is as if whichever regime first comes to mind is on that account deemed to be the morally best regime.) For yet another example of inconsistency generated by Craswell's hybrid approach, see note 30.

25. Craswell (2003, p. 258) seeks to cast doubt on our objection to discontinuity on the ground that the objection applies to nearly any theory involving individual rights. However, that an otherwise sound argument also applies to other theories is hardly a response to the argument. Furthermore, his claim is incorrect, for it has long been true that many philosophers (and even more so, their fellow travelers) who accept individual rights or other principles would allow them to be smoothly traded off in a manner that does not imply any discontinuity, both in cases of conflicts between competing rights and in cases of conflicts between rights and welfare. A prominent example is Ross (1930). Additionally, it is entirely familiar that the use of knife-edge distinctions in formulating

Craswell (2003, pp. 249–50) asks the reader to accept both of the following two features of his hybrid scheme:

- First, if anyone suffering from unfairness under a regime is in some (possibly unrelated and indirect) manner better off than under another, fair regime, the former regime's unfairness receives no weight whatsoever. Thus, if everyone is treated unfairly under regime A compared to regime B, and if each person thereby suffers in an amount that he regards as indifferent to losing $1,000, then the unfairness is entirely ignored if each individual has $1,000.00000000001 more under A than under B.
- Second, if the compensation instead leaves a single individual merely as well off, giving him exactly $1,000, so no longer is everyone strictly better off under regime A, then unfairness receives full weight and hence regime B would be favored.[26]

But if one indeed accepts the first feature (as is required under Craswell's scheme to avoid the Pareto conflict),[27] it seems ridiculous to accept the second. Rather, if unfairness dissolves entirely when individuals each re-

---

legal rules (see Craswell 2003, p. 258) hardly implies that the underlying principles, viewed from an ideal perspective, have radical discontinuities. (Consider setting an age floor for driving or voting.) And, as we emphasize in our book (see, for example, *FVW*, pp. 66–69, 76–77), similar points hold true for moral intuitions, including Craswell's, Chang's, or a reader's, that may seem to favor a discontinuous system of rights even though the true, ideal theory behind it would not. These ideas, which fully rationalize our intuitions yet lend no support to taking them as real indicators of the correct normative theory, are aspects of one of the main themes of our book (see Sec. 1.4, above), one not addressed by Craswell.

26. The example in the text, following Craswell, uses the weak Pareto principle, which is decisive only when every individual is strictly better off. If instead, one required only a single individual to be strictly better off (and the others merely at least as well off) or if one ignored fairness if everyone was equally well off, a slightly different illustration could be used to make the same point.

27. Craswell (2003, pp. 251–52) acknowledges that some fairness theorists would balk at his first premise, but he suggests that not all would and that they need not. We are much more skeptical, which is to say that it seems to us that most who endorse notions of fairness could not accept this premise. For example, Craswell (2003, p. 257) points out that under the required assumption nothing can ever be viewed as intrinsically evil—or, we would add, intrinsically wrong, unfair, or unjust. Craswell is not ruling out only the extreme view that intrinsic evil trumps all other considerations, including welfare, but also far more modest views that simply hold that certain acts, rules, or situations are viewed negatively from a normative point of view independent of the welfare levels of those involved and how these welfare levels may be influenced by aspects of the regime that have nothing to do with what makes something intrinsically evil. See also note 22 (suggesting related implications of his first premise that we suspect most fairness proponents would reject). Nevertheless, our argument only aims to show that, for any who do accept the first premise, the second premise is implausible.

ceive an additional $1,000.00000000001, then surely it dissolves substantially if one of them merely receives an additional $1,000. Put another way, if regime A is clearly and nontrivially normatively inferior to regime B if the one individual under A receives exactly $1,000 more than under B, then if he instead receives exactly $1,000 for sure and an additional one-billionth of 1 cent as well if lightning strikes 100 consecutive times at a given point within a predetermined 1-second interval, surely no coherent moral theory would reverse its verdict on account of such a difference. If individuals' well-being is decisively important when the amount individuals each receive is $1,000.00000000001, can it really make sense that consideration of individuals' well-being is radically less important or wholly unimportant when a single individual's receipt is ever so slightly less?[28]

We acknowledge that the reader may have some difficulty accepting the last set of points regarding incoherence and discontinuity. Why is it that, in order to avoid the Pareto conflict, Craswell (like Chang) offers modified theories that are so bizarre and incoherent? Why did he not suggest more plausible variations of notions of fairness instead? The answer lies in our proof, which demonstrates that under minimal assumptions of logical coherence and continuity, avoiding the Pareto conflict while still giving weight to notions of fairness is impossible. Hence, those seeking to avoid the proof's implications (the Pareto conflict) are forced to propose schemes having such unacceptable properties.

The proof itself appears in our original technical paper (Kaplow and Shavell 2001, p. 284) and in a prose version in our reply to Chang (Kaplow and Shavell 2000, pp. 240–41). One way to express the intuition underlying it is as follows: As long as a notion of fairness sometimes receives nontrivial weight, it follows that one would sometimes favor a fair regime over an unfair regime where the latter has greater welfare (but not by so much as to outweigh whatever weight was given

28. Craswell's hybrid theories can be described as ones involving waivers of rights, but this way of describing the theories does not evade the substance of our argument. If waiver is automatic (as under Craswell's theory) when an individual gains by $1,000.00000000001, then the degree of moral offense from a lack of waiver when the individual instead gains by $1,000 can hardly be significant. Indeed, it can hardly be worth an individual's—or a policy analyst's or government decision maker's—effort even to determine a regime's effects with such precision, if such were possible. We remind the reader that Craswell needs this discontinuity to avoid our proof of the conflict with the Pareto principle; hence, the obvious "fix" to his hybrid scheme, under which the degree of unfairness rises as the extent of compensation for the unfairness falls short of full compensation, is unavailable to him.

to fairness). Now one can also imagine a regime like the latter one—just as unfair and for the same reason (say, unfair punishment is applied equally often and to the same extent), but where the overall level of social welfare is distributed the same as it is in the first regime, implying that the excess welfare relative to the first regime is shared pro rata. In that case, the modified second regime is viewed as inferior to the first regime (because the modified second regime is no better than the original second regime, which itself is inferior to the first regime according to the notion of fairness). Nevertheless, this modified second regime has higher total welfare with the surplus divided evenly, necessarily making everyone better off than under the first regime. To avoid this sort of conflict with the Pareto principle, Craswell (like Chang) must have the weight given to fairness change dramatically—in fact, at a literally infinite rate—in response to possibly tiny changes in the level or distribution of welfare. This explains why their circumvention schemes need to be discontinuous.[29]

In addition, the presence of such a discontinuity helps to understand why attempts to evade the reach of our proof also tend to display intransitivity (A definitely superior to B, B definitely superior to C, but A not superior to C). In essence, discontinuity in the present context means that qualitatively different evaluative principles are used depending on which comparisons are being made (under the Craswell/Chang hybrid, fairness ordinarily, but welfare alone when Pareto conflicts arise). But when the criteria used to compare A with B, to compare B with C, and to compare A with C can and sometimes do differ, it should not be surprising that inconsistent judgments are possible.[30]

29. It has been suggested to us that discontinuity is not problematic as long as the absolute amount sacrificed is not infinite; that is, there is no difficulty per se with the rate of change (in a sense, the price per unit) being infinite. But consider a simple example: A person enters a chocolate shop and asks to buy a pound of elegant truffles. The shopkeeper responds, "You misunderstand, sir. You have only $1,000 in your pocket, and the price is $1 trillion a pound." This suggestion about discontinuity is tantamount to saying, "O.K. then, I'll take $1,000 worth." The idea seems to be that this answer cannot be crazy if the customer pays only $1,000. But if indeed the $1 trillion price is ridiculous, then so is purchasing $1,000 worth, which in this case might amount to a few molecules (we have not performed the calculations). And since the person literally would purchase at an infinite price, even if one imagines that it is possible to taste such a minute morsel, this would not suffice to rescue the sanity of the approach because the customer must be willing to buy even if the price were $1 trillion per electron and so on ad infinitum.

30. Yet another manner in which a hybrid theory's discontinuity can generate inconsistency involves uncertainty. Suppose, for example, that a reform would otherwise be unfair to a large group of individuals, that each of them ex ante expects to gain from the

In sum, Craswell's suggestion that a hybrid theory might circumvent our Pareto argument, although superficially appealing, fails to respond to the thrust of our substantive claims—such as those about what the Pareto conflict reveals about the initial notion of fairness or about the implications of our symmetric-case demonstration—and his scheme does not, upon examination, constitute a logically coherent system of normative assessment. We are quite sympathetic to Craswell and others' intuitive conviction that there must be some way to circumvent our demonstrations regarding the Pareto principle. In fact, when writing *FVW*, we had substantially completed our detailed analysis before we realized the generality of the phenomenon that notions of fairness sometimes make everyone worse off. Had we been asked about the matter 5 years ago, we would have been inclined to agree with Craswell and others' suspicion that there must be some way around the claim for at least certain principles of fairness. But, after much reflection and subsequent formal analysis, we have been forced by the power of logic to change our minds. However counterintuitive at first (and second) glance, the deep conflict between notions of fairness and the Pareto principle is quite general, robust, and powerful in its implications for normative analysis.

3. DISTRIBUTIVE JUSTICE

Jeremy Waldron (2003) addresses matters of distributive justice. As we explain in Section 1.1, the welfare economic approach, which we endorse, incorporates general distributive concerns. However, because of the nature of our thesis, consideration of distributive justice is not central in our book. Our claim, after all, is that policy assessment should be based exclusively on a policy's effects on individuals' well-being and

reform, but that exactly one of them will in fact end up being the slightest bit worse off ex post (say 1 million people each have a one-in-a-million chance of being that person). Craswell's theory would favor the reform, viewed ex ante, because the unfairness is deemed to be waived by everyone. But if the decision maker (or moral analyst) had brief access to a crystal ball that revealed the identity of the loser, then if he happened to see the loser's name in the crystal ball an instant before reaching the mental conclusion in favor of the reform, the reform would have to be scrapped because that one loser could not be deemed to waive his right not to be subject to unfairness. This result must hold even though knowing the name of the person really is morally irrelevant information because it was already known that precisely one person would end up slightly worse off, and under the moral theory it is presumably immaterial whether that person, selected essentially randomly, happens to turn out to be Smith or Jones.

thus that no independent weight should be accorded to notions of fairness that are not concerned exclusively with individuals' well-being. Because principles of distributive justice often are (and, when not, can often be reformulated to be) based exclusively on individuals' well-being, the merits of our thesis do not bear very directly on what is the correct general distributive theory.[31] As we state early in *FVW* (and as Waldron quotes), "[W]e argue, in essence, that legal policy analysis should be guided by reference to *some* coherent way of aggregating individuals' well-being, in contrast to the view that policy analysis should be guided by notions of fairness and thus, at least in part, without regard to individuals' well-being" (*FVW*, p. 27; quoted in Waldron 2003, p. 287).

Waldron, although applauding our acceptance of the importance of incorporating distributive justice in our normative framework,[32] nevertheless offers three objections to our treatment of distributive issues. First, he is bothered by our "idiosyncratic" or "odd" definition of fairness because it excludes many distributive principles (classifying them instead as welfarist when they are based exclusively on individuals' well-being).[33] We have some sympathy for this point but do not regard it as significant for a number of reasons.

Most important, we make our definition clear in our introduction,

31. Our thesis does imply, as the text suggests, that distributive theories should address themselves to individuals' well-being rather than other measures of individuals' situations. See *FVW*, pp. 29–30 n. 27.

32. We observe that we feel unworthy of any applause, for in *FVW* all we did was state the standard normative paradigm of welfare economics, which, as widely taught in graduate schools, takes distribution into account (although as we there acknowledge, see *FVW*, pp. 5, 28–29 and n. 26, this standard view of economists is not widely disseminated in legal scholarship).

33. See Waldron (2003, pp. 279, 282, for use of the quoted terms; and pp. 279–81 for the more general claim). In the course of making this point, Waldron also advances a more substantive objection, one that we find difficult to understand. See Waldron (2003, pp. 283–86). In essence, he states our thesis or claim—that policy assessment should be based exclusively on well-being—as a "premise" and finds our conclusion—that non-welfare-based principles, notions of fairness, should receive no weight—to be nearly a tautology, whereas we understand our claim as something that the body of our book was meant to demonstrate rather than as something to be taken as given. (As we note in Sec. 1.2, part of one of our arguments in support of our first main theme is a tautology, but that is another matter.) In this regard, Waldron does not here (or, to a very substantial extent, elsewhere in his comment) make reference to our extensive, detailed development of our three main themes in chapters 3–6. He also questions the significance of the Pareto conflict given that actual instances of conflict seem unlikely (Waldron 2003, pp. 284–85) and wonders what we would say about arguments such as those based on "desert" (p. 286). Hopefully, our summary in Sec. 1 (and, regarding the Pareto point, our further elaboration in Sec. 2) helps to clarify our enterprise in these respects.

in chapter 2, section B.1 (which is devoted entirely to our definition of fairness), and throughout the book. (We also offer a formal, mathematical statement of the definition, which should avoid any possibility of confusion. See *FVW*, pp. 39–40 n. 52.) In a related way, it seems unlikely that those who get past our title would misinterpret our thesis as a frontal assault on distributive justice, for we emphasize our contrary view of overall distributive matters in the text early in our introduction, in our conclusion, and in chapter 2, sections A.2 and A.3, which are devoted expressly to the subject.[34] We also consider distributive effects at a number of points throughout the body of the book. Indeed, the many quotations and citations in Waldron's comment testify if anything to an excess of concern on our part in communicating our meaning with regard to this issue.

Our definition of notions of fairness was not chosen arbitrarily, but rather for a very simple reason: "[W]e define notions of fairness as we do—to include all principles that give weight to factors independent of individuals' well-being but only such principles—because the substance of our argument depends precisely on this characteristic" (*FVW*, p. 44).[35] Thus, our definition neatly states the domain of our thesis and of our analysis. Furthermore, as we note, all of the leading notions of fairness that we address in the book are covered by this definition. We needed a simple term to refer to this crucial demarcation. Something like "notions of fairness other than those that are distributive in a purely welfarist manner" seemed far too clumsy (not to mention how it would have ruined our title), and we were unable to come up with anything better than "fairness" (which had the advantage over some alternatives in that it has no generally accepted canonical meaning).

Second, Waldron (2003, pp. 287–93) objects to our decision not to take up the substance of distributive justice in our book, having instead set it aside as a separate matter for inquiry.[36] Among other things, Wald-

34. Moreover, in the abstract of the earlier law review version (*FVW* 2001, p. 966) of our book, we also make clear that distributive concerns are not generally included in our critique of notions of fairness.

35. Waldron (2003, pp. 279–80) also quotes us (*FVW*, p. 39) to similar effect.

36. He also expresses the concern that our analysis makes unavailable the use of the language of fairness for addressing distributive questions. See, for example, Waldron (2003, p. 290). As we explain in Sec. 1.1, however, our argument concerns the proper criteria for policy assessment and not at all the language in which arguments relating to such criteria are expressed. Indeed, following the very quotation that Waldron offers to demonstrate the basis for his concern, we state, "And, as we elaborate in the next subsection, distribution can play an important role even under a system of evaluation that is concerned exclusively

ron (2003, p. 290) criticizes our "professed agnosticism" and our reliance on an "appeal to the division of academic labor." To a large extent, we plead guilty, but we fail to understand in what sense this claim amounts to a criticism.[37]

Distributive justice is a subject unto itself, the topic of countless books, indeed of many scholars' life work. As it stands, our book is quite long, and in advancing our thesis about notions of fairness, we are directly challenging central tenets held by a substantial majority of moral philosophers and legal scholars who have addressed our subject over the past century. Many warned us that we were taking on far too much already. Attempting as well to offer a comprehensive account of all important aspects of distributive justice in the same enterprise would have been imprudent.

It is also the case that dividing analysis of nonwelfarist notions of fairness that are not concerned with global distributive questions and analysis within a welfarist framework of purely distributive questions makes a good deal of sense as a logical matter since the subjects are largely separable.[38] In any event, *FVW* does offer a number of remarks concerning distributive matters.[39] And it so happens that we have addressed issues of income distribution in some of our other work.[40] Finally, one of us (L.K.) is currently writing a series of articles and possibly two books on distributive justice and government policy.

Third, Waldron (2003, pp. 293–99) argues that we fail to appreciate how situational distributive judgments (that is, those concerned with who should pay or be paid in a given legal dispute) entailed by the notions of fairness that we address might be provisionally helpful—pending construction of a full distributive theory—because they may constitute fragments of, or constraints on, an overall theory of distributive justice within a welfarist framework. Because *FVW* does not

---

with individuals' well-being. Moreover, the criticisms of notions of fairness that we offer are not criticisms of the language that analysts use or of the need to make value judgments in assessing legal policy" (*FVW*, p. 28).

37. Waldron's (2003, p. 290) complaint about our "indifference" and some of his other statements, however, do not seem apt, as amply demonstrated by his extensive quotations from our book indicating the contrary.

38. For a qualification, see note 31.

39. See especially chap. 2, secs. A.2 and A.3; chap. 3, secs. C.2 and D.2; chap. 4, sec. D.1; and chap. 8, secs. C and D.4.b. Some other pertinent comments can be located using our index.

40. Most obviously, Kaplow and Shavell (1994). Interested readers can also see our other papers cited therein as well as a substantial portion of Kaplow's writing on taxation.

purport to address what distributive theory is most compelling or how that determination should be made, we also do not see this claim as a criticism of our work. Nevertheless, drawing on our analysis, we do offer two observations regarding Waldron's argument.

One is that we discuss explicitly the possibility that some fairness principles that we criticize may nevertheless serve as proxy indicators of overall distributive concerns. We make this point especially with regard to notions of fairness in tort law.[41] However, we find that the notions are a poor proxy.[42] Also, examining a proxy notion is generally not a promising way of achieving a refined understanding of whatever concept underlies it.[43]

Additionally, we suspect that Waldron is overly optimistic about any direct role that situational notions of fairness could have in constructing components of an overall theory of distributive justice. After all, these notions are not welfarist, so if our thesis is accepted (and Waldron does not challenge it in advancing his claim), it follows that the notions would tend to push a distributive theory in normatively inappropriate directions. More broadly, we have trouble understanding how Waldron thinks that a nonwelfarist principle can serve as a component of or constraint on a social welfare assessment that is not supposed to give weight to any factor other than welfare. Finally, given the many defects that we identify in these fairness principles (see Sections 1.2 and 1.3), one should be rather reluctant to take intuitions related to such notions (see Section 1.4) as reliable guides in formulating a sound distributive theory.[44]

41. See, for example, *FVW*, p. 96 n. 20, p. 122 and n. 93, and p. 138 n. 127. For example, "We also observe that some notions of fairness might be viewed as providing proxy tools for identifying opportunities to improve the distribution of wealth, rather than as independent evaluative principles potentially opposed to it. For example, a principle requiring that victims be compensated might have this feature if victims are typically poorer than injurers" (*FVW*, p. 122).

42. See, for example, *FVW*, p. 122 and n. 93.

43. Furthermore, as implied by our discussion in *FVW*, chap. 2, sec. A.3, notions of fairness—even if they were illuminating proxies for aspects of distributive justice—are unlikely to be helpful in making actual policy choices in legal (and many other) settings because distributive issues are usually best addressed more directly, through the tax and transfer system. See also Kaplow and Shavell (1994).

44. We note that, in this section of Waldron's comment, he makes no reference to our general arguments about the defects of these notions of fairness or to our detailed analysis thereof in chaps. 3–6.

4. PREFERENCES AND WELL-BEING

Lewis Kornhauser (2003) raises some issues involving our broad notion of well-being, in particular with regard to our including, among all the intangibles that people may value, the possible tastes individuals might have regarding notions of fairness themselves.[45] He insists that our approach is a "strategy," one "comparable to the legal strategy of confession and avoidance" Kornhauser (2003, p. 307). Yet this characterization is inapt. To begin, we fail to understand what we are confessing to or avoiding.[46] Nor should our definition be viewed in this manner when it is in fact standard (although often implicit) in economics not to discriminate among different possible sources of well-being in the assessment of individuals' welfare. In addition, as explained in Section 1.1 and as is clear from our book (for example, *FVW*, chap. 2, sec. A.1; and chap. 8, secs. B.3 and B.4), we adopt the position that we do because we believe that one should be neutral regarding individuals' well-being rather than arbitrarily privilege some aspects over others.

45. In passing, Kornhauser makes various other claims and characterizations, but we will confine our attention to his two main conclusions. We note, however, that many of his statements are incomplete or otherwise misleading with regard to what we actually say in our book. For example, Kornhauser (2003, p. 304) opens his comment with the statement, "In *Fairness versus Welfare*, Louis Kaplow and Steven Shavell reformulate an earlier claim asserted in the legal literature that judges ought to maximize wealth." As is apparent from the many quotations in Waldron's comment (see, for example, Waldron 2003, p. 291), this characterization is inaccurate. At the outset of our introduction, we emphasize that our thesis is emphatically not wealth maximization (see *FVW*, p. 5), and when we discuss the matter directly, we state that wealth maximization "is not a well-defined concept" and "[m]ore importantly, and more obviously, . . . wealth would not constitute a measure of social welfare under welfare economics because wealth is not defined in terms of individuals' well-being" (*FVW*, pp. 35–36). In addition, Kornhauser's (2003, pp. 305–8) summary of our argument is both incomplete and somewhat inaccurate, as should be clear if one compares it with our summary in Sec. 1 above or the summaries in *FVW* (for example, the introduction and chap. 2). Likewise, he follows this summary with the statement, "Three of these four claims are controversial. A large literature in ethics debates the merits of welfarism" (Kornhauser 2003, p. 308). It is as if our book reflected no awareness of such debates, whereas in fact that the book is entirely devoted to engaging in that debate and, in the process, cites and discusses in detail literally hundreds of books and articles by philosophers holding views opposed to ours. As one last example, his conclusion (pp. 325–28) addresses a different subject from that in the rest of his paper, that is, the relationship between the work of policy analysts and public officials. In so doing, he cites only one of our many sections addressed directly to these matters (ignoring, for example, *FVW*, chap. 5, sec. C.2; chap. 7, sec. B; and chap. 8, sec. A), and what he says about that one section (chap. 7, sec. C) does not closely reflect what is actually presented therein.

46. Our discussion below of Kornhauser's apparent substantial misunderstanding of our book may provide the explanation.

Moreover, as we emphasize in our book[47] (as others have noticed),[48] whether one adopts our view of individuals' well-being or some other is largely irrelevant to our enterprise, for neither our thesis nor the arguments supporting it depend on the definition of well-being that is chosen. We suspect that this is apparent from our summary in Section 1, and nowhere in Kornhauser's comment does he explain otherwise. Thus, neither of his two main conclusions really matter for our purposes. Nevertheless, we now address each in turn.

First, Kornhauser (2003, pp. 310–16) argues that taking an encompassing view of well-being, specifically, one that includes any tastes individuals might have for notions of fairness, "does not obviously resolve conflicts between rights and Pareto optimality" or, as he puts it in his conclusion, "between morality and efficiency" (Kornhauser 2003, pp. 316, 325). We would prefer to phrase the point as involving a failure to resolve the conflict between fairness and welfare. But how could he imagine that we would think that it would? If a mere definition, first stated in the introduction of our book, resolved the basic conflict, why would we have written what we did? And why the "versus" in our title *Fairness versus Welfare*? As explained, we chose our definition of well-being for concreteness (as our analysis does not depend on it) and because we thought it the most compelling, not because we thought that our definition would eliminate the need to offer substantive arguments in establishing our thesis.

The actual content of this section of Kornhauser's paper consists of an extended discussion of an example of Amartya Sen's, featuring a demonstration of the point that, indeed, an encompassing definition of well-being that includes tastes for fairness (for a certain notion of rights, in the particular example) does not necessarily dissolve the underlying conflict. As stated, we are hardly bothered by this and fail to see the relevance of the entire discussion. Part of our problem in making sense of how this section relates to our arguments is that at no point in it does Kornhauser cite anything in our book. In particular, he does not mention any of our points directed to rights, the existence of our chapter 5 on legal procedure, which contains a specific section (C.1) on the subject

47. See *FVW*, pp. 23–24 n. 14, and p. 409. These statements appear, respectively, at the close of our section "Individuals' Well-Being" at the outset of our book and on the first page of our later section "Preferences and Individuals' Well-Being." Since these are the two discussions in our book most pertinent to Kornhauser's comment, it is surprising that he ignores this central point.

48. See Craswell (2003, pp. 272–73).

of tastes regarding procedures (a subject on which Kornhauser merely speculates as to what we might think), or even of our specific discussion of the very example of Sen's that Kornhauser considers.[49]

Second, Kornhauser (2003, pp. 316–23) criticizes us for equating raw preferences (Jill likes chocolate, I prefer vanilla, and there's no room for dispute in matters of raw taste) with judgments (vanilla is more healthful or eating chocolate is morally preferable because it is a worthy political statement against exploitation of peasants involved in the harvesting of vanilla beans). Kornhauser (2003, p. 325) summarizes his claim by stating that we "conflate[] judgments that individuals make with their preferences, which are understood as a favorable attitude, such as a desire." Throughout the section, he offers characterizations of our views on the matter that range from admitted speculation to fairly confident assertions, and he explains why he finds our alleged views incorrect, notably, for failing to appreciate that judgments, unlike raw tastes, may reflect matters of fact or analysis about which individuals may be mistaken.

Kornhauser's discussion, however, is disconnected from what we actually say. Like the preceding section of his comment, it contains not a single reference to our book. The sections that he does not mention include chapter 8, section B, which is entitled "Preferences and Individuals' Well-Being," which includes a subsection B.1 on precisely how to address the possibility that individuals' preferences might be mistaken and a subsection B.4, which is entitled "Tastes for Notions of Fairness." By contrast, in a footnote in an earlier section of his comment, he does cite a relevant passage, stating that "Kaplow and Shavell *would* permit the policy maker to revise the extended preferences of an individual to accord with the preferences she would have given true beliefs and sufficient deliberation" (Kornhauser 2003, p. 310 n. 12, emphasis added; citing *FVW*, pp. 23–24). This accurate representation clearly contradicts the characterizations of us that Kornhauser offers, without citation, in the section itself.

As noted above, our views on this subject are ultimately irrelevant to our thesis and our main themes and arguments. Nevertheless, a brief summary of our position might be helpful. We do not in fact believe that all tastes, preferences, aspects of individuals' well-being—ranging

49. See *FVW*, p. 54 n. 75. (Lest the reader be misled, this is one of our "long" footnotes, and it is readily located, both because it is in our main, fairly short section that first discusses our argument about the Pareto conflict, to which Sen's article pertains, and because it is listed in at least three places in our index, under Pareto principle, rights, and Sen.)

from those concerning material satisfaction to relationships with others, aesthetics, or fairness—are identical in every possible respect. Many types differ in ways that are important for some purposes and yet also have other features in common. Thus, as Kornhauser's footnote statement acknowledges, we do indeed think that a raw taste is different from, say, a belief about states of the world, and we expressly discuss the potential relevance to legal policy analysis, noting, for example, that imperfect information may affect the assessment of legal rules related to safety regulation.[50] More generally, we view all possible tastes as the same in the sense that what is relevant for normative analysis is what actually matters to individuals, but tastes may differ, for example, regarding ease of measurement, the likelihood that they may change over time or in reaction to policy changes, variation across individuals, and the possible importance of errors in individuals' assessments.

Regarding tastes involving fairness in particular, we expressly raise the possibility of mistaken preferences and suggest that analysts should be willing to look behind such preferences (see, for example, *FVW*, chap. 8, sec. B.4, esp. pp. 433–34). To take a concrete example, suppose that shortly after a heinous crime, many individuals, in the heat of the moment, desire that the suspect be lynched (or rushed to trial without procedural safeguards) out of a sense that the gross immorality of the act demands, as a matter of justice, an extremely swift response. It is obvious that such tastes could involve mistakes of various sorts; indeed, it may even be likely that the same individuals would, once the tension eases, deeply regret a prior hasty action and accordingly feel much worse off. Thus, we would not consider the initial expression of individuals' tastes for fairness or justice to be indistinguishable from an ordinary expression of a preference for chocolate or vanilla. (For various reasons, we also believe that Kornhauser may overdraw the distinctions among various types of tastes,[51] but the foregoing should be sufficient to indicate

50. See *FVW*, chap. 8, sec. B.1. In that section, we also note the case in which individuals may be misinformed about their own preferences.

51. First, although we are unsure of Kornhauser's position in this regard, what he calls preferences involving judgments do seem to also have an irreducible aspect of raw taste, making his dichotomous categorization misleading. This is obvious with his examples involving health but seems equally applicable to tastes regarding notions of fairness. Thus, contemplation may affect one's tastes, but ultimately, if one does really care about fairness in some manner, there must exist an underlying basis for the motivation. Perhaps a person would feel guilty if he thought he was supporting an unfair institution. Without such an underlying desire, a view about fairness would not really constitute a component of well-being rather than an opinion on policy or ethics. Second, his analogy between matters of

that our views—as reflected in what we actually wrote—are substantially different from what Kornhauser imagines them to be and criticizes.)

5. CONCLUSION

In *FVW*, we argue that policy assessment should be based exclusively on well-being and hence that no weight should be accorded to independent notions of fairness. We advance our thesis by developing three themes, both in general and abstract terms and in detailed analysis in a range of legal policy contexts and with regard to numerous leading notions of fairness.

Our first theme concerns how pursuing notions of fairness often involves perverse sacrifices of human welfare, including the possibility that everyone will be made worse off. Craswell's claim that this latter possibility can be circumvented by resorting to hybrid notions of fairness does not really address most of our argument and in any event is unsuccessful. Waldron's objections regarding distributive justice suggest the need for more work on the subject but do not call into question any of our analysis. And Kornhauser's comments on preferences and well-being are essentially irrelevant to our arguments in this regard and also reflect misunderstandings of our views. Our second important theme, concerning additional deficiencies in notions of fairness, especially with regard to their lack of rationale, and our third theme, which reconciles the intuitive appeal of notions of fairness with our claim that they should receive no independent weight in policy assessment, are not questioned by these commentators.

Given these commentaries, the presence of other critiques of our work both in print and forthcoming, and perhaps most importantly the almost

---

fairness and of health is incomplete in an important respect. Suppose that I feel good because I did what I view to be the morally right thing. Even if the best analysis by moral theorists reveals that I am mistaken, this does not deny my feeling, and if I never had explained to me the error of my ways, I would never in fact feel regret. But if I mistakenly eat poison, even if I never come to understand what hit me, I will nonetheless suffer. One must keep in mind, as we emphasize throughout *FVW*, for example, pp. 11–12, 21–23; chap. 5, sec. C.1; and chap. 8, sec. B.4, that when fairness is viewed as a taste rather than as an evaluative principle to be given weight in policy assessment, the question of its importance is entirely an empirical one, concerning what (and how strong) individuals' tastes actually are at various points in time and under various conditions, not the normative question of what their tastes for fairness might (or should) be if only they were in fact more reflective. (Compare: Maybe I should like abstract art and in fact would do so if I had the proper training, but if I have not, being forced to hear a 3-hour lecture on an obscure modern artist will in fact bore me to tears rather than engender feelings of ecstasy.)

unshakable belief we all have in our own instincts and intuitions that seem to support notions of fairness, we find it useful to echo the conclusion of *FVW* by reminding the reader of what seems minimally necessary, in light of our analysis, for a proposed notion of fairness to be taken seriously.[52]

1. The notion of fairness must be stated with some precision and in a manner that is complete (unlike virtually all the leading notions of fairness that we consider).

2. It must be explained how the notion can make sense given that the consequences of pursuing it may well run counter to the notion's underlying motivations.

3. The manner in which the notion reduces individuals' well-being, including the possibility of its reducing everyone's well-being, needs to be clearly identified.

4. An explicit rationale for according weight to the notion must be offered.

5. Contrary explanations for the notion's seeming appeal have to be ruled out; this is especially so regarding the possibility (which appears to be realized with respect to every leading notion of fairness that we examine) that the notion's appeal may lie in its correspondence to social norms that themselves are best understood functionally, as serving to promote individuals' well-being.

In our book, we developed each of these points in great detail, with tremendous attention to the views of the many scholars over the ages who advance notions of fairness. For those who advocate that analysts should nevertheless be guided by notions of fairness, it would seem incumbent on them to offer a direct response, addressing each of these points with regard to whatever notion of fairness is being offered.[53]

52. Brief, often familiar, and typically emotionally charged hit-and-run counterexamples to our position (many of which we examined, some in depth, in our book) are, we would argue, insufficient to raise a serious challenge in light of our arguments and, in many instances, the serious, thoughtful analyses of many scholars who have preceded us. (We note that the commentators in the present symposium do not resort to this common tactic.)

53. Furthermore, we suggest that an attack on our overall position would be more plausible if it concentrates on one of the many leading notions of fairness that we consider in *FVW*. After all, we take on a variety of fairness principles, and those that we examine are among the most prominent ones developed over the ages, by the likes of Aristotle and Kant as well as legions of contemporary moral philosophers (not obscure notions that have received limited support). As we believe we have demonstrated, notions that have long seemed persuasive can be shown to be fundamentally deficient on many grounds if the notions are scrutinized with sufficient care.

REFERENCES

Chang, Howard F. 2000a. A Liberal Theory of Social Welfare: Fairness, Utility, and the Pareto Principle. *Yale Law Journal* 110: 173–235.

———. 2000b. The Possibility of a Fair Paretian. *Yale Law Journal* 110: 251–58.

Craswell, Richard. 2003. Kaplow and Shavell on the Substance of Fairness. *Journal of Legal Studies* 32:245–75.

Kaplow, Louis, and Steven Shavell. 1994. Why the Legal System Is Less Efficient Than the Income Tax in Redistributing Income. *Journal of Legal Studies* 23: 667–81.

———. 1999. The Conflict between Notions of Fairness and the Pareto Principle. *American Law and Economics Review* 1: 63–77.

———. 2000. Notions of Fairness versus the Pareto Principle: On the Role of Logical Consistency. *Yale Law Journal* 110:237–49.

———. 2001. Any Non-welfarist Method of Policy Assessment Violates the Pareto Principle. *Journal of Political Economy* 109: 281–86.

———. 2002. *Fairness versus Welfare*. Cambridge, Mass.: Harvard University Press (*FVW*). (Version previously published in *Harvard Law Review* 114 (2001): 961–1388.)

Kornhauser, Lewis A. 2003. Preferences, Well-Being, and Morality in Social Decisions. *Journal of Legal Studies* 32:303–29.

Rawls, John. 1980. Kantian Constructivism in Moral Theory. *Journal of Philosophy* 77:515–72.

Ross, W. D. 1930. *The Right and the Good*. Oxford: Clarendon Press.

Waldron, Jeremy. 2003. Locating Distribution. *Journal of Legal Studies* 32: 277–302.