ISSN 1045-6333

HARVARD

JOHN M. OLIN CENTER FOR LAW, ECONOMICS, AND BUSINESS

PARETO PRINCIPLE
AND COMPETING PRINCIPLES

Louis Kaplow

Discussion Paper No. 519

08/2005

Harvard Law School
Cambridge, MA 02138

JEL Classes A13, D63, H43

Pareto Principle and Competing Principles

Louis Kaplow[*]

Abstract

The Pareto principle, the seemingly incontrovertible dictum that if all individuals prefer some regime to another then so should society, may conflict with competing principles. Arrow's impossibility theorem and Sen's liberal paradox are two notable examples. Subsequent work indicates more broadly that the Pareto principle conflicts with all nonwelfarist principles. This essay surveys these results, including various extensions thereof, and offers perspectives on the conflict, drawing on classical and contemporary work in political economy and economic psychology.

Forthcoming, *The New Palgrave Dictionary of Economics, 2nd Edition*

[*] Harvard University and National Bureau of Economic Research. I am grateful to Steven Shavell for comments and the John M. Olin Center for Law, Economics, and Business at Harvard University for financial support..

Pareto Principle and Competing Principles

Louis Kaplow

The Pareto principle holds that if all individuals strictly prefer one state, regime, or policy to another, then that selection is deemed socially preferable as well. Because of the power of unanimous endorsement, the Pareto principle has understandably been important in normative economic analysis. Even though strict Pareto dominance is unlikely to prevail when society is deciding among plausible competing alternatives (for this would require that literally each of millions preferred the same outcome), the Pareto principle nevertheless offers important guidance. In particular, the principle may help in choosing among or ruling out various other evaluative notions; principles that turn out to conflict with the Pareto principle may accordingly be rejected. Alternatively, if some competing principles seem compelling, they may raise doubts about the ostensibly incontrovertible Pareto principle.

The first sections to follow review two well-established conflicts between the Pareto principle and certain competing principles: Arrow's (1951) impossibility theorem and Sen's (1970) liberal paradox. The succeeding section presents more recent work that establishes a general conflict between the Pareto principle and all nonwelfarist notions, whether they concern rights, justice, or other conceptions of fairness (apart from those pertaining only to the distribution of welfare itself). A final section examines classically-grounded strands of literature, on political economy and economic psychology, that help reconcile the tension between the seemingly unimpeachable Pareto principle and conflicting nonwelfarist principles, many of which have appeal to the public, policy-makers, and economists as well. (The Pareto principle is also important in normative economic analysis, notably with regard to the two fundamental theorems of welfare economics, a subject not considered in this essay.)

1. ARROW'S IMPOSSIBILITY THEOREM.

Perhaps the most famous instance of conflict between the Pareto principle and competing principles is Arrow's (1951) impossibility theorem. Arrow considered social choice procedures designed to generate a consistent social ordering (a complete and transitive ranking) from purely ordinal information about individuals' preferences. In one formulation of Arrow's theorem, the assumptions of universal domain (no restriction on individuals' preferences), independence of irrelevant alternatives (the social ordering of any two alternatives depends only on individuals' orderings of those two alternatives), nondictatorship (no one individual's preferences completely determine social preferences), and the Pareto principle imply that such a social ordering is impossible.

A large subsequent literature explores whether relaxing some of Arrow's assumptions modestly would make possible procedures that yield robust social orderings. Of particular relevance here are attempts to weaken the Pareto principle. As surveyed in Campbell and Kelly (2002), these efforts have been largely unsuccessful: Either there are frequent violations of the Pareto principle or a single individual will have substantial, even if not completely dictatorial, influence.

Nevertheless, Arrow's theorem does not rule out the class of standard, individualistic social welfare functions (SWF's), mappings from individuals' utilities to a measure of social welfare, that are fully consistent with the Pareto principle. Consider the discrete case, in which there are n individuals, $U_i(x)$ is the utility of the i^{th} individual, and x is a complete description of the pertinent state. Then we can define $W(U_1(x), \ldots, U_n(x))$ as an individualistic SWF (so called because it depends only on individuals' utilities). Assuming, as is standard, that W is increasing in each individual's utility, it follows that, for any set of individuals' utility functions $\{U_i(x)\}$, W provides a complete and transitive social ordering of all possible social states that is independent of irrelevant alternatives, nondictatorial, and satisfies the Pareto principle. The classical utilitarian criterion, $W = \sum U_i(x)$, is an example of such an SWF.

The possibility of an SWF is restored by altering Arrow's framework to allow the domain of social choice procedures to consist of individuals' utilities rather than just their orderings. This approach entails interpersonal utility comparisons, which during the mid-twentieth century (and to an extent thereafter) were eschewed in welfare economics, following the argument of Robbins. As Robbins (1935, vii-x; 1938) himself clarified in his second edition and a subsequent essay, however, his argument was not that interpersonal comparisons should not be made – indeed, they were inevitable – but rather that they involve value judgments rather than scientifically verifiable statements. Much modern welfare economics has pursued analysis of SWF's that depend on individuals' utilities and not just orderings, presumably because of a belief that preference intensities matter and that interpersonal comparisons are required if distributive judgments are to be made.

2. SEN'S LIBERAL PARADOX.

In "The Impossibility of a Paretian Liberal," Sen considered whether the Pareto principle conflicts with a specific notion of liberalism, subsequently described by many (including, on occasion, Sen himself) as a species of libertarianism. His condition stipulates that that there exists certain choices about which the social ranking should reflect that of a particular individual, regardless of other considerations, including effects on the utility of others. This conception and Sen's analysis thereof is well illustrated by considering his much-discussed example. One individual, whom we shall call Prude, abhors erotic literature, and a second, Lewd, adores it. Both individuals' preferences, moreover, are assumed to be meddlesome in the following manner. Prude would be more upset by Lewd's reading a certain lascivious novel than reading it himself, and Lewd would get more pleasure from Prude's reading the novel than reading it herself. Therefore, as between just Prude reading the novel and just Lewd reading it, both prefer the former. However, Sen's liberal principle insists that the latter be the social choice: Prude's preference against his own reading of the book, ceteris paribus, dictates socially that Prude should not read the book, and likewise Lewd's desire that she read the book, ceteris paribus, dictates socially that Lewd should read it. Hence, the choice that Sen's liberal principle deems socially best is one that would be rejected under the Pareto principle.

Analytically, Sen's result can be understood by reference to the familiar concept of externalities. Lewd's reading the book involves a negative externality on Prude, whereas Prude's reading the book involves a positive externality on Lewd. (Compare the case in which Lewd moderately enjoys loud parties that greatly annoy his neighbor Prude, and Prude would rather not bother to replace his weed-ridden garden with flowers that would greatly delight his neighbor Lewd.) Failing to regulate externalities obviously may violate the Pareto criterion. Furthermore, in Sen's example, the two individuals – if left to themselves – would wish to enter a Coasian bargain under which Prude, rather than Lewd, reads the book (just as, in the variation, Lewd should agree to refrain from loud parties if Prude agrees to replace his weeds with flowers). Sen's principle implicitly prohibits both government regulation and private exchange in which individuals mutually relinquish their posited liberal rights. Preventing mutual waiver both by vote and by contract may hardly seem liberal, as argued by Gibbard (1974) and many others in a highly elaborated literature, surveyed by Suzumura (2005). Indeed, any notion that conflicts with the Pareto principle must embody an underlying opposition to freedom since a violation of the Pareto principle entails contravention of unanimous choice. Some of Sen's subsequent writing (e.g., 1992, 144-46) defends his original liberal principle on grounds of practicality and concern for governmental abuse of power. As will be explored in section 4, however, such Millian (1859) justifications for rights may be powerful but are not, at root, inconsistent with the Pareto principle.

3. CONFLICT BETWEEN PARETO PRINCIPLE AND ALL NONWELFARIST PRINCIPLES.

Sen showed that one particular formulation of a libertarian principle, which carries the implication that externalities of a sort may not be regulated, violates the Pareto principle. Subsequently, it has been asked more broadly which notions of right, justice, and fairness conflict with the Pareto principle. The answer, it turns out, is that essentially all such notions do, as long as they do not depend exclusively on individuals' utilities – that is, unless they are a reformulation of welfarism

To state the matter more precisely, we can contrast the individualistic SWF introduced previously, $W(U_1(x), \ldots, U_n(x))$, which by construction depends only on individuals' utilities, with the more generalized SWF, $Z(x)$ – which also may be written as $Z(U_1(x), \ldots, U_n(x), x)$. Under the latter, social welfare may depend on anything and, in particular, need not depend exclusively on how the pertinent state x affects individuals' utilities. For example, notions of merit or desert concern whether certain actions or attributes are rewarded, principles of corrective or retributive justice demand that specific norm violations be followed by compensation or punishment, and so forth. Under each of these nonwelfarist criteria, knowing each individual's utility in state x is insufficient information to form a social judgment.

Kaplow and Shavell (2001) prove that if an SWF is not individualistic, then it violates the Pareto principle, if one makes a certain continuity assumption. The assumption is not that the

SWF is continuous in all respects. (It is allowed, for example, that infinitesimal violation of some right might cause a discrete reduction in social welfare.) Rather, it is assumed that there exists some good that, if all individuals are given more of it, ceteris paribus (e.g., holding rights violations constant), all will have a higher utility and, moreover, the value of the SWF changes continuously as the amount of that good is changed.

The proof is roughly as follows. First, if the SWF does not depend only on individuals' utilities, there must exist two states that are evaluated differently despite everyone's utilities being the same. That is, the nonwelfarist SWF is supposed, in at least one instance, to rank states differently on account of a nonwelfare difference. Now, taking whichever of the two states ranks lower, we can increase slightly everyone's allotment of the aforementioned good. By continuity, if that increase is sufficiently small, the lower-ranking state must still be ranked lower. However, since all individuals had equal levels of utility in the two initial states, every individual in the modified state now has greater utility, making it Pareto preferred despite the fact that the posited nonwelfarist SWF ranks it lower. Hence, the Pareto principle is violated.

One way to understand the conflict between the Pareto principle and all nonwelfarist principles is to reflect on the fact that a nonwelfarist SWF by definition gives some weight in some instances to a factor independent of its effect on individuals' utilities. We can compare a state that is preferred on account of this nonutility factor to a state that is otherwise identical except that all individuals are slightly better off with respect to some commodity. In other words, a nonwelfarist SWF, by its nature, sometimes sacrifices welfare, and nothing in logic rules out the possibility that the welfare sacrifice is borne pro rata.

Subsequent work has generalized and extended this theorem. Campbell and Kelly's (2002) survey notes that the proof in Kaplow and Shavell (2001) does not require the SWF to be a function, rather than a binary relation; that this relation need not be fully transitive, only acyclic; and that only lower continuity is required. In a different vein, Suzumura (2005) derives a sort of converse, namely, given Pareto indifference (if everyone is indifferent then society is indifferent – a principle implied by welfarism), social choice must respect the weak Pareto principle (the version defined at the outset of this entry) as well as the strong Pareto principle (if everyone weakly prefers one alternative and at least one individual strictly prefers it, then it is socially preferred). This theorem requires two additional assumptions: positive responsiveness of the social decision to individual preferences and that, ceteris paribus, any utility level for an individual can be reached by adjusting the amount of a particular divisible good received by that individual.

Kaplow and Shavell (2002) also offer a complementary demonstration of the conflict between all nonwelfarist principles and the Pareto principle. If one restricts attention to symmetric settings – those in which all individuals are identically situated – then any nonwelfarist principle conflicts with the Pareto principle in every instance in which its ranking differs from a purely welfarist one. Because everyone is affected identically, it must be that, whenever any amount of aggregate welfare is sacrificed, each and every individual's welfare is sacrificed. The significance of this result is that many traditions favor assessing principles for guiding society in hypothetical situations that, because they are designed to create an impartial perspective, have a symmetric character. Consider, for example, the original position of Rawls (1971) – with important prior formulations thereof by Harsanyi (1953) and others – in which individuals are taken to have no knowledge of their own characteristics. Likewise, the injunctions of the Golden Rule and, relatedly, of Kant's (1785) categorical imperative demand, in essence, that one examine rules as if both positive and negative consequences were borne symmetrically by all. Since, as noted, all choices in symmetric settings involve strict Pareto rankings (except in cases in which all are indifferent), admitting a nonwelfarist principle entails the view that the socially preferred state is systematically one in which everyone is worse off.

4. PERSPECTIVES ON THE CONFLICT.

The Pareto criterion is a bedrock principle. Yet it conflicts with all nonwelfarist principles – whether they pertain to rights, justice, or fairness – and some of these principles have apparent appeal. How may this tension be reconciled? That the Pareto principle should be seen as paramount is suggested by the rhetorical question: To whom is one doing right, providing justice, or being fair if every possible beneficiary is thereby made worse off? Additionally, as Sidgwick (1907) and others have queried, if something like utility does not underlie rights and related concepts, by what criterion is the proper list of rights determined in the first instance and how in principle should the inevitable conflicts between different rights be resolved? A possible reconciliation is suggested by lines of thinking that trace their roots to prominent political economists of a prior era (among others), as more recently elaborated in Kaplow and Shavell (2002).

The relationship between the Pareto principle and other seemingly appealing principles can be understood by reference to what are known as two-level moral theories. (Act versus rule utilitarianism comes to mind, although that somewhat problematic distinction is subtly different from the one under consideration.) As suggested by Hume (1751), Mill (1861), and Sidgwick (1907), one can envision a first-level principle (such as utility) that provides our ideal assessment of states (corresponding to an SWF) and also numerous second-level principles (e.g., that one should keep promises, tell the truth, not kill others) that are to used as guides by individuals in their everyday conduct. Subsequent prominent statements of this view include Harrod (1936), Rawls (1955), and, most extensively, Hare (1981).

Put in a more explicit optimizing framework, the first-level principle serves as the objective function and possible second-level principles constitute the universe of feasible policies. This feasible set is assumed to be constrained by limits of human nature and human institutions. Accordingly, the optimal scheme – taken here to consist of the optimal subset of second-level principles – will be only second best. The aforementioned limits render any attempt at direct implementation of the first-best criterion – commanding that everyone in their individual or institutional capacity act always so as to maximize social welfare – inferior to employment of second-best principles that, inevitably, deviate from the first-best criterion (welfare) in some instances. Two sets of rationales for this conception of the social maximization problem have been offered.

The first sort of justification is based on decisionmaking costs, complexity, limited information, limited self-control (e.g., myopia), and so forth. Such considerations imply that all manner of behavior, including some types that have no interpersonal effects, should be guided by rules. Moreover, given the nature of the problems that such rules are designed to address, it is inevitable that the rules will not require performance of a complete social welfare calculus and hence will sometimes command behavior that differs from the first-best outcome. This conflict hardly makes the first-best principle any less of an ideal, just one that is not perfectly achievable in practice.

Second, the nature of human motivation, particularly the problem of cabining self-interest, provides another reason that sensible individual and institutional commands sometimes deviate from a pure concern for individuals' utilities and thus offers another account of the conflict between the Pareto principle (viewed here as an aspect of the first-level social objective) and alluring nonwelfarist principles (understood as second-level rules). Emphasized by Hume, Mill, and Sidgwick, and also by Smith (1790) and Darwin (1874), this strand of thinking is rooted in what may be called moral psychology. As a consequence of biological and social evolution, human emotions may help to channel behavior in a positive fashion. Opportunism – whether through cheating, theft, or aggression – may be constrained by the prospect of guilt feelings or social disapprobation. Cooperation may be encouraged by anticipated positive internal sentiments or praise by others. Two familiar examples are the retributive urge, the prospect of which may deter aggression, and the desire for social approval, which may inhibit opportunism and encourage constructive collaboration. Given the limitations of biological evolution (limits on altruism as well as the tendency of evolved mechanisms to be specialized), constraints on social inculcation (including the fact that much is directed at young children), and the factors mentioned with regard to the first rationale for second-level rules, it is unsurprising that the resulting precepts sometimes deviate from the first best. Once again, this gap does not call into question the supremacy of the first-best ideal as a matter of principle. (Interestingly, however, this second explanation suggests that emotional force will be associated with moral criteria – various notions of what is right, just, or fair – that conflict with the Pareto principle, which helps explain why our intuitions may be in tension with pure welfarism in some settings.)

Both of these enduring strands of thought that help to reconcile the conflict between the Pareto principle and nonwelfarist notions are related to the more recent upsurge of interest at the intersection of economics and psychology, often under the rubric of behavioral economics. Just as Tversky and Kahneman (1974) have stimulated research on heuristics and biases in a range of economic settings, Baron (1993) and others have documented similar phenomena – such as overgeneralization – in individuals' moral thinking. Likewise, many researchers, including Frank (1988) – following intervening provocative statements by Darwin (1874) and Wilson (1975) – have reinvigorated Smith's interest in human emotions as forces that guide human behavior, although not always in an ideal manner.

The foregoing discussion suggests that, in regulating individuals' behavior, various normative criteria that conflict with the Pareto principle may nevertheless usefully advance welfare and thus, at root, be consistent with the underlying force for that principle. These nonwelfarist notions may also be relevant to the promotion of welfare for other, related reasons. As argued at length by Bentham (1822-1823) in his constitutional writings and Mill (1859) in *On*

Liberty, second-best rules obviously may play an important role in constraining government officials. In addition, since many of the nonwelfarist criteria exist because of their relationship with the promotion of welfare, they may be useful proxy standards in some settings. Finally, due to the affective aspect of many nonwelfarist principles, a complete welfarist account would incorporate them because they are in part constitutive of individuals' utilities. Note that, in each instance, because the relevance of nonwelfarist criteria lies in the advancement of welfare, there is no conceptual inconsistency with the ultimate motivation for the Pareto principle even though the nonwelfarist second-level rules on their face deviate from the posited first-level ideal.

In sum, a complete understanding of the relationship between the Pareto principle and other, possibly competing normative principles involves many dimensions. Formal analysis of these principles reveals the existence of an underlying, logical conflict. Examination of literatures in other fields of economics and in other disciplines, however, suggests a fundamental harmony.

BIBLIOGRAPHY

Arrow, K.J. 1951. *Social Choice and Individual Values*. New York: Wiley.

Baron, J. 1993. *Morality and Rational Choice*. Boston: Kluwer Academic Publishers.

Bentham, J. 1822-23 [1990]. *Securities Against Misrule and Other Constitutional Writings for Tripoli and Greece*. P. Schofield, ed. Oxford: Oxford University Press.

Campbell, D.E., and Kelly, J.S. 2002. Impossibility Theorems in the Arrovian Framework, in *Handbook of Social Choice and Welfare, vol. 1*. K.J. Arrow, A.K. Sen, and K. Suzumura, eds. Amsterdam: Elsevier Science, 35-94.

Darwin, C. 1874 [1998]. *The Descent of Man; and Selection in Relation to Sex*. Second ed. Amherst, NY: Prometheus Books.

Frank, R.H. 1988. *Passions within Reason*. New York: W.W. Norton & Co.

Gibbard, A. 1974. A Pareto Consistent Libertarian Claim. *Journal of Economic Theory* 7: 388-410.

Hare, R.M. 1981. *Moral Thinking: Its Levels, Method, and Point*. Oxford: Oxford University Press.

Harrod, R.F. 1936. Utilitarianism Revised. *Mind* 45: 137-156.

Harsanyi, J.C. 1953. Cardinal Utility in Welfare Economics and in the Theory of Risk-Taking. *Journal of Political Economy* 61: 434-435.

Hume, D. 1751 [1998]. *An Enquiry Concerning the Principles of Morals*. T.L. Beauchamp, ed. Oxford: Oxford University Press.

Kant, I. 1785 [1997]. *Groundwork of the Metaphysics of Morals*. M. Gregor, trans. and ed. Cambridge: Cambridge University Press.

Kaplow, L. and Shavell, S. 2001. Any Non-welfarist Method of Policy Assessment Violates the Pareto Principle. *Journal of Political Economy* 109: 281-86.

Kaplow, L. and Shavell, S. 2002. *Fairness versus Welfare*. Cambridge, Mass.: Harvard Univ. Press.

Mill, J.S. 1859. *On Liberty*. London: J.W. Parker.

Mill, J.S. 1861 [1998]. *Utilitarianism*. R. Crisp, ed. Oxford: Oxford University Press.

Rawls, J. 1955. Two Concepts of Rules. *Philosophical Review* 64: 3-32.

Rawls, J. 1971. *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.

Robbins, L. 1935. *An Essay on the Nature and Significance of Economic Science*. Second ed. London: Macmillan.

Robbins, L. 1938. Interpersonal Comparisons of Utility: A Comment. *Economic Journal* 48: 635-41.

Sen, A.K. 1970. The Impossibility of a Paretian Liberal. *Journal of Political Economy* 78: 152-157.

Sen, A.K. 1992. Minimal Liberty. *Economica* 59: 139-159.

Sidgwick, H. 1907 [1981]. *The Methods of Ethics*. Seventh ed. Indianapolis: Hackett Publishing Company.

Smith, A. 1790 [1976]. *The Theory of the Moral Sentiments*. Sixth ed. Oxford: Oxford University Press.

Suzumura, K. 2005. Welfarism, Individual Rights, and Procedural Fairness, in *Handbook of Social Choice and Welfare, vol. 2*. K.J. Arrow, A.K. Sen, and K. Suzumura, eds. Amsterdam: Elsevier Science (forthcoming).

Tversky, A. and Kahneman, D. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 195: 1124-1131.

Wilson, E.O. 1975. *Sociobiology: The New Synthesis*. Cambridge, Mass.: Harvard University Press.