PREFERENCE, RATIONAL CHOICE AND ARROW'S THEOREM*

IT seems intuitively plausible to expect of a consistent rational agent that if he preferred an alternative x to another alternative y and y to a third alternative z then he would still prefer x to y if z suddenly became unavailable or y to z if x suddenly became unavailable or x to z if y became unavailable. Similarly, if he was given a choice only between x and y and expressed a preference for x over y, we should expect that, if a third alternative z became available, x would still be preferred to y. As a consequence, we should expect that, if x is the most preferred alternative from a set S of alternatives, then x would be the most preferred alternative from any subset of S of which x is a member; that is, we should expect the following sentence to be true:

$$(\forall x)\{x \in S_1 \subset S_2 \rightarrow [x \in C(S_2) \rightarrow x \in C(S_1)]\}$$

where x ranges over alternatives, S_1 and S_2 are sets of alternatives, and $C(S)$ denotes the value of a function (called a "choice function") from S to the alternative(s) in S that is (are) preferred at least as much as any other alternative in S. To remain consistent with the literature on social choice theory, I shall follow A. K. Sen[1] in referring to this as "property α."

It is easy to see why property α is a fundamental assumption in virtually all the literature on rational preference and social choice. Consider the case in which an individual is asked to give a preference ordering over three political candidates A, B, and C. If he prefers A to B and B to C and A dies, then if no third candidate enters the race he should vote for B. If he, in fact, votes for C, then, it would seem, this must be because he has just changed his mind or

*I am grateful to Ellis Crasnow, David Gauthier, James Kahan, Sharon Labrot, Stephen Schiffer, Robert Schultz, and Bas van Fraassen for their comments on earlier versions of this paper.

[1] *Collective Choice and Social Welfare* (San Francisco: Holden-Day, 1970).

0022-362X/81/7812/0778$00.80

because the death of A has triggered some complex chain of events (for example, it was discovered that B killed A) that call for a major reappraisal or because his ordering over A, B, and C was misrecorded initially or because of some other such factor outside the domain of rational preference.

I shall, however, present a case in which property α is violated for none of these reasons, but rather for purely rational reasons. Such a counterexample should be of intrinsic interest, since property α seems such a minimal constraint to place upon rational preference and choice. Beyond this, however, is a point of specific interest; for the basic intuition that underlies property α is also the basic intuition behind one of the conditions necessary to prove Arrow's impossibility theorem. The connection between Arrow's theorem and my counterexample to property α will be discussed in section II of this paper.

I

Consider a game in which two players A and B, who are prohibited from communicating with each other, match coins against a bank. They may show heads, tails, or nothing. The payoffs, with A's shown first, are:

	B shows heads	B shows nothing	B shows tails
A shows heads	2,2	−1,−1	−1,−1
A shows nothing	−1,−1	1,1	−1,−1
A shows tails	−1,−1	−1,−1	2,2

Probability theory dictates that in this situation two ideally rational agents seeking to maximize their expected utility should settle upon a pair of strategies that will result in an undominated equilibrium outcome. In this game there are three equilibrium outcomes: (1) A and B showing heads, (2) A and B showing nothing, and (3) A and B showing tails. Outcome 2, however, is dominated by 1 and by 3.

Beyond the straightforward calculation, however, there is a certain epistemic complication pointed out by David Gauthier which

he calls "accessibility."[2] Although 2 is dominated by both 1 and 3, these two outcomes are *inaccessible.* Without communication neither *A* nor *B* can form any expectation about which of the two equally appealing outcomes the other will shoot for, and, without such an expectation, playing either heads or tails commits one to a 50 per cent chance of receiving 2 and a 50 per cent chance of receiving −1. Thus, playing either heads or tails becomes a bad gamble, and, since we are assuming *A* and *B* to be ideally rational, they should both realize this and show nothing in order to guarantee themselves a return of 1.

So *A* and *B* should both prefer showing nothing to showing either heads or tails, and, in addition, both should be indifferent between showing heads and showing tails. Let us represent these preference orderings as follows:

A	*B*
nothing	nothing
heads-tails	heads-tails

Now let us consider the same coin-matching game, but this time *A* is allowed to use only two of his original strategies: showing heads and showing nothing. The payoff matrix for this version of the game is:

	B shows heads	*B* shows nothing	*B* shows tails
A shows heads	2,2	−1,−1	−1,−1
A shows nothing	−1,−1	1,1	−1,−1

Since showing nothing is the rational choice in the set of alternatives {showing heads, showing nothing, showing tails}, property α requires that it be the rational choice in the set {showing heads, showing nothing}; but obviously this is false! In this second version of the game, *A* and *B* showing heads dominates *A* and *B* showing nothing, and, since *A* can't show tails (and *B* knows this), both *A* and *B* should expect the other to show heads. In other words, *A*

[2] "The Impossibility of Rational Egoism," this JOURNAL, LXXI (Aug. 15, 1974): 439-456, p. 448. Gauthier introduces the notion of accessibility in the context of the same coin-matching game that I have used.

and *B* showing heads is an *accessible* dominating equilibrium outcome. So, in this second version of the game, the preference orderings become representable as:

A	*B*
heads	heads
nothing	nothing
	tails

Note that not only has *A*'s ordering been changed, but *B*'s has also—and this is *just because* one of *A*'s alternatives has been removed.

It is very important to note that it is *just* the removal of the alternative that brings on the reordering; for this fact separates the present case from that mentioned earlier in which a person who originally prefers three political candidates in order *A-B-C* winds up voting for *C* rather than *B*, who is still in the race for office in spite of the fact that he has been charged with the murder of *A*. In this case the reordering from *A-B-C* to *C-B* is *not* brought on *just because* the alternative of voting for *A* is removed, but rather because it is removed in a certain way, and obviously it is unreasonable to expect property α to hold regardless of what chain of events is triggered in the process of removing an alternative. In other words, *property α should be expected to hold all things being equal, not come what may.*

One might want to object to my counterexample on the grounds that the reordering brought about there is not *just* a result of removing a single alternative, that by removing an alternative one somehow changes the game. But this claim is simply not true. The rules that define the game and specify payoffs could be set out without any mention of which alternative strategies each player must have or, for that matter, without even specifying that every player have a coin (since a player can show nothing).

There is another possible objection which isn't very good but I'll mention it anyway. One might claim that I have done something illicit by considering preference relations over strategies rather than outcomes. If outcomes were being considered then there would be no problem at all, for receiving 2 would always be preferred to receiving 1 and 1 to receiving -1. This strikes me, though, as a very thin hair to try to split. On one reading it attempts a distinction between doing something and getting something, and one need hardly point out that more often than not what people are trying to *get* by their actions is the opportunity to *do* something. On

another reading it attempts a distinction between instrumentally good alternatives and intrinsically good alternatives, but one would certainly not want to claim that preference relations should be thought to obtain only between intrinsically good alternatives; for then one could not prefer (strictly speaking) getting $1000 to getting 1¢, since money is of only instrumental value. At any rate, one can easily devise a game in which the outcomes are directions to use certain strategies in a second game, in which case an outcome just *is* a strategy.

A related line of attack would insist upon a requirement that, whatever kind of alternatives can bear binary preference relations to each other, the only kind that decision theory should concern itself with are those with some kind of *determinate* value. Since the alternatives in my counterexample are strategies in a gamble, they do not have determinate values. In the first place, since the game in my counterexample is between two ideally rational players (and each knows that he is playing with an ideally rational player), their chosen strategies *do* have determinate outcomes (How else could *I* have known them?). Secondly, this requirement is too harsh anyway, since, as anyone who has tried to buy anything lately knows, not even money is of determinate value.

II

Basically, the strategy behind the Arrow impossibility theorem is to list some intuitively plausible constraints (or "conditions") that any method of arriving at social choices on the basis of individual preferences (which Arrow calls a "social welfare function"—or SWF) should satisfy and then show that these conditions are inconsistent and, thus, that no SWF can possibly meet them. Informally, three of the four conditions necessary in order to prove the Arrow theorem are:

(1) The SWF must supply a social ordering for every logically possible combination of individual preference orderings over any given set of alternatives (unrestricted domain).
(2) If everyone in the society prefers x to y, then the SWF must result in a social ordering of x over y (Pareto principle).
(3) The SWF cannot specify that the preferences of a single individual determines a social ordering on every issue regardless of the preferences of everyone else (nondictatorship).

The fourth condition is of special importance here; so it will be stated formally:

(4) Let $R_1, \ldots, R_n$ and $R_1', \ldots, R_n'$ be two sets of individual orderings, and let $C(S)$ and $C'(S)$ be the corresponding social choice functions. If, for all individuals i and all alternatives x and y in a given environment S, xR_iy if and only if $xR_i'y$, then $C(S)$ and $C'(S)$ are the same (independence of irrelevant alternatives).[3]

The notation 'xR_iy' here should be read "individual i prefers x at least as much as y." Informally, what is being required by this condition is that the social ordering derived from a given set (or "environment") S of individual orderings be unaffected by the existence of orderings over alternatives not in S.

The effect of this requirement is twofold. First, it rules out the interjection of nonfeasible alternatives into lists of feasible ones in order to determine interpersonal comparisons of *strengths* of preference. For example, given that a choice is to be made between two alternatives x and y in a society of two individuals A and B and that A prefers x to y and B prefers y to x, we would think it reasonable to say that the society as a whole is indifferent between x and y. If, however, we were to interject the set of nonfeasible alternatives {+\$1000, +\$1, −\$1, −\$1000} into the set of feasible ones $\{x, y\}$ and A's and B's orderings were then representable as follows:

A	B
+\$1000	y
+\$1	+\$1000
x	+\$1
y	−\$1
−\$1	−\$1000
−\$1000	x

then it would seem reasonable to say that y should be the social choice, since B *strongly* prefers y to x and A *just barely* prefers x to y. But the independence condition rules out this change of mind from our original finding that society should be indifferent between x and y.[4] Since the choice—in fact—involves only x and y, it

[3] Kenneth J. Arrow, *Social Choice and Individual Values*, 2nd ed. (New Haven, Conn.: Yale, 1963), p. 27. Parenthetical page references hereafter will be to this book. Arrow's original proof used five conditions. I am following Sen in using only four. The numbers I have assigned to the four conditions do not correspond to those used by Arrow.

[4] This problem with the independence condition is noted by R. Duncan Luce and Howard Raiffa in *Games and Decisions* (New York: Wiley, 1957), p. 341.

should, according to the independence condition, be independent of alternatives that cannot—in fact—be chosen. The essential point of this requirement is, as Arrow puts it, to reinforce the principle that "Only observable differences can be used as a basis for explanation" (109). He continues,

> Given the set of alternatives available for society to choose among, it could be expected that, ideally, one could observe all preferences among the available alternatives, but there would be no way to observe preferences among alternatives not feasible for society (110).

The second effect of this condition is to enforce a certain conception of rationality which holds "that the choice to be made from any set alternatives can be determined by the choices made between pairs of alternatives" (20). Stated alternatively,

> Knowing the social choices made in pairwise comparisons . . . determines the entire social ordering and therewith the social choice function $C(S)$ for all possible environments (28).

The usual course pursued by those who would take Arrow to task on the independence condition is to attack the condition because of its first effect, arguing either that Arrow's presupposed constraints on observability are too harsh and that empirically respectable methods of determining preference strength do exist[5] or—as John Harsanyi does—that the whole business of firmly shackling explanation to observables is "a result of uncritical acceptance of a seriously mistaken—and by now completely superseded—philosophical doctrine, that of logical positivism."[6] The case made by either of these arguments is quite strong. As for the first, although it may be difficult—in practice—to ascertain strengths of preferences by interjecting nonfeasible alternatives, it is not difficult to suppose that—in principle—such information can be reliably obtained, and, at least in those cases when it can be obtained, it should be used—or, at least, not disregarded as a matter of principle. The power of Harsanyi's line of argument is, I think, fairly obvious, given the general turn away from positivism in the last few decades, and he has convinced many people working in the field of social decision theory to see things his way.

Whatever the strengths or weaknesses of these attacks, however, mine is an attack on the other front. What is called into question

[5] Several suggestions about how strengths of preferences can be calculated and amalgamated within various sets of constraints on observability are discussed by Luce and Raiffa, pp. 345-353.

[6] "Bayesian Decision Theory, Rule Utilitarianism, and Arrow's Impossibility Theorem," *Theory and Decision*, XI, 3 (September 1979): 289-317, p. 302.

by my counterexample to property α is the whole conception of rational preference formation which it and the independence condition are intended to capture. In particular, I am arguing that, in general, when a fully rational agent arrives at a choice from a set S of alternatives, he is not necessarily asserting anything at all about what his second, third, or fourth choices would be if his first choice became nonfeasible, and, likewise, that when he gives his pairwise comparisons over all the two-member subsets of S, he is not necessarily asserting anything at all about his choice from S. Given, then, that this conception of rationality seems to be missing something at the level of individual choice, it is not at all clear why we should try to impose it at the level of social choice by requiring the condition of the independence of irrelevant alternatives.

It should be pointed out that arguments against the independence condition like mine were, to some extent, anticipated by Arrow; for he realized that

> . . . the model of rational choices as built up from pair-wise comparisons does not seem to suit well the case of rational behavior in the . . . game situation . . . . The precise shape of a formulation of rationality which takes . . . [game-theoretic considerations] into account or the consequences of such a reformulation on the theory of choice in general or the theory of social choice in particular cannot be foreseen; but it is at least a possibility, to which attention should be drawn, that the paradox discussed below [Arrow's theorem] might be resolved by such a broader concept of rationality (20/21).

Although the point of divergence between the game-theoretic conception of rational choice and Arrow's conception which I have pointed out is not the same as the one Arrow noticed,[7] I think it fair to say that what I have done here is to press the basic issue with which Arrow was concerned.

TAL SCRIVEN

University of Southern California
California Polytechnic State University at San Luis Obispo

[7] The point of dissimilarity in question which Arrow noticed was that in a game situation the environment from which an agent must choose contains an infinite number of alternatives, since any environment that contains more than one strategy also contains all the possible randomizations over those strategies. Thus, a choice could not be produced by pairwise comparisons, for there are infinitely many of them to make. See Arrow, p. 20.