Economica (40), 1973

Behaviour and the Concept of Preference[1]

By AMARTYA SEN

I

Thirty-five years have passed since Paul Samuelson published in the house journal of the London School of Economics his pioneering contribution to the theory of "revealed preference".[2] The term was perhaps not altogether a fortunate one. Revelation conveys something rather dramatic, and the biblical association induced the late Sir Dennis Robertson to wonder whether "to some latter-day saint, in some new Patmos off the coast of Massachusetts, the final solution to all these mysteries had been revealed in a new apocalypse".[3] While the appropriateness of the terminology may be debatable, the approach of revealed preference has gradually taken hold of choice theory in general and of demand theory in particular.

My intention in this lecture is to examine the philosophy behind the approach of revealed preference and to raise some queries about its use, and then to go on to discuss the implications of these issues for normative economics. The crux of the question lies in the interpretation of underlying preference from observations of behaviour.

"The individual guinea-pig," wrote Paul Samuelson, "by his market behaviour, reveals his preference pattern—if there is such a consistent pattern."[4] If a collection of goods y could have been bought by a certain individual within his budget when he in fact was observed to buy another collection x, it is to be presumed that he has revealed a preference for x over y. The outside observer notices that this person *chose* x when y was available and infers that he *preferred* x to y. From the point of view of introspection of the person in question, the process runs from his preference to his choice, but from the point of view of the scientific observer the arrow runs in the opposite direction: choices are observed first and preferences are then presumed from these observations.

The consistency condition that Samuelson based his theory on, which has come to be known as the Weak Axiom of Revealed Preference, says that if a person reveals a preference—in the sense just defined—for x over y, then he must not also reveal a preference for y over x. That is,

[1] An inaugural lecture delivered at the London School of Economics on 1 February 1973.

[2] P. A. Samuelson, "A Note on the Pure Theory of Consumer's Behaviour", *Economica*, vol. 5 (1938). Also "A Note on the Pure Theory of Consumer's Behaviour: An Addendum," *Economica*, vol. 5 (1938).

[3] D. H. Robertson, *Utility and All That*, London, 1952, p. 19.

[4] P. A. Samuelson, "Consumption Theory in Terms of Revealed Preference," *Economica*, vol. 15 (1948).

if he chooses x when y is available, then he will not choose y in a situation in which x is also obtainable. Armed with this innocuous looking axiom, Samuelson proceeded to obtain analytically the standard results of the theory of consumer's behaviour with remarkable economy.[1] It also opened up the way for empirical studies of preferences based on observed market behaviour.[2]

The approach of revealed preference need not be confined to market choices only, and indeed it has been used in studying preferences revealed by non-market behaviour such as government decisions, choices of public bodies and political acts like voting. The exact mathematical structure of the problem differs substantially from case to case, and the formulation in the context of preferences revealed by political or bureaucratic decisions will differ from that in the context of consumer's choices. But there are common methodological elements, and I shall be concerned with them in this lecture.

II

Before I proceed to examine the status of the preference revealed by choice, I would like to comment on one very elementary issue that seems to me to have certainly clouded the interpretation of revealed preference theory. This concerns the somewhat surprising claim that has been frequently made that the theory of revealed preference "frees" demand theory from the concept of preference and *a fortiori* from the concept of utility on which traditional demand theory was based.

In his pioneering paper, Samuelson argued that his object was "to develop the theory of consumer's behaviour freed from any vestigial traces of the utility concept".[3] The exact content of the statement was not altogether clear, and in pushing forward the revealed preference approach in a classic paper, Little argued that one of his main aims was to demonstrate "that a theory of consumer's demand can be based solely on consistent behaviour",[4] adding that "the new formulation is scientifically more respectable [since] if an individual's behaviour is consistent, then it must be possible to explain that behaviour without reference to anything other than behaviour".[5] In a similar vein, Hicks stated that "the econometric theory of demand does study human beings, but only as entities having certain patterns of market behaviour; it makes no claim, no pretence, to be able to see inside their heads".[6]

[1] See Samuelson's articles, referred to earlier, and also his *Foundations of Economic Analysis*, Cambridge, Mass, 1947.

[2] For a recent survey of the analytical literature in this branch of economics, see A. Brown and A. Deaton, "Models of Consumer Behaviour: A Survey", *Economic Journal*, vol. 82 (1972).

[3] Samuelson, "A Note on the Pure Theory . . .", p. 71.

[4] I. M. D. Little, "A Reformulation of the Theory of Consumer's Behaviour," *Oxford Economic Papers*, vol. 1, (1949), p. 90.

[5] Little, *ibid.*, p. 97.

[6] J. R. Hicks, *A Revision of Demand Theory*, Oxford, 1956, p. 6. Hicks did not, however, fully subscribe to the revealed preference approach himself. See especially "The Measurement of Income," *Oxford Economic Papers*, New Series, vol 10, 1958.

On this interpretation the use of the word "preference" in revealed preference would appear to represent an elaborate pun. In saying that x is revealed preferred to y, it would not be asserted that x is preferred to y in the usual sense of the word "preferred". A redefinition of the expression "preference" is, of course, possible, but it is then legitimate to ask what does "consistency" of behaviour stand for and on what basis are the required consistency conditions chosen. The alleged inconsistency between (i) choosing x when y is available and (ii) choosing y when x is available, would seem to have something to do with the surmise about the person's preference underlying his choices.

Preferring x to y is inconsistent with preferring y to x, but if it is asserted that choice has nothing to do with preference, then choosing x rather than y in one case and y rather than x in another need not necessarily be at all inconsistent. What makes them look inconsistent is precisely the peep into the head of the consumer, the avoidance of which is alleged to be the aim of the revealed preference approach.

It could, however, be argued that what was at issue was not really whether the axiom of revealed preference represented a requirement of consistency, but whether as a hypothesis it was empirically verified. This line would not take one very far either. Consider the simplest situation of one consumer facing two divisible commodities—the case that figures on blackboards in every Economics Department in the world, and would have, I imagine, adorned the magnificent glass doors of the St. Clement's Building but for the greater deference shown by our architects to the even more classic demand-and-supply intersection. Even in this rudimentary case, the set of possible choice situations for any individual is infinite—indeed uncountable. To check whether the Weak Axiom holds for the entire field of all market choices, we have to observe the person's choices under infinitely many price–income configurations. In contrast, the number of actual choices that can be studied is extremely limited. Not only is the ratio of observations to potential choices equal to zero, but moreover the absolute number of cases investigated is also fairly small. Comparisons have to be made within a fairly short time to avoid taste change, but the time elapsed must also be sufficiently long so that the mutton purchased last time is not still in the larder, making the choices non-comparable. With durable goods the problem is quite vicious. The actual number of tests carried out have, not surprisingly, been very small. Faith in the axioms of revealed preference arises, therefore, not from empirical verification, but from the intuitive reasonableness of these axioms interpreted precisely in terms of preference. In fact, the concept of taste change is itself a preference-based notion, and the whole framework of revealed preference analysis of behaviour is steeped with implicit ideas about preference and psychology.

I would, therefore, argue that the claim of explaining "behaviour without reference to anything other than behaviour"[1] is pure rhetoric,

[1] Little, *op. cit.*, p. 97.

and if the theory of revealed preference makes sense it does so not because no psychological assumptions are used but because the psychological assumptions used are sensibly chosen. The use of the word preference in revealed preference must indeed be taken to be more than a pun.

Indeed, the psychological assumptions involved have been discussed explicitly or by implication in all the major contributions to revealed preference theory. There have also been discussions about "the transition to welfare economics" from revealed preference theory, and even Ian Little has argued that among the possible routes for this transition is the view "that a person is, on the whole, likely to be happier the more he is able to have what he would choose".[1] Samuelson had in any case put less emphasis on sticking exclusively to observed behaviour, and his statement, which I quoted earlier, that "the individual guinea-pig, by his market behaviour, reveals his preference pattern",[2] makes the fundamental assumption of revealed preference theory explicit. The rationale of the revealed preference approach lies in this assumption of revelation and not in doing away with the notion of underlying preferences, despite occasional noises to the contrary. So we would be justified in examining the philosophical foundations of the revealed preference approach precisely in terms of the assumption of revelation. This is what I shall now go on to do.

III

I shall take up a relatively minor question first. The Weak Axiom of Revealed Preference is a condition of consistency of two choices only. If x is revealed preferred to y, then y should not be revealed preferred to x. Perhaps because of this concentration on the consistency between any *two* choices and no more, the Weak Axiom has appeared to many to be a condition of what Hicks calls "two-term consistency". And it has appeared as if the other well-known requirement of consistency, *viz*, transitivity, lay outside its scope. Transitivity is a simple condition to state: if x is regarded as at least as good as y, and y at least as good as z, then x should be regarded as at least as good as z. In the case of preference, it implies that if x is preferred to y and y preferred to z, then it should also be the case that x is preferred to z. Since this condition involves at least three choices and since the Weak Axiom involves a requirement of consistency only over *pairs* of choices, it might look as if the Weak Axiom could not possibly imply transitivity. This has indeed been taken to be so in much of the literature on the subject, and additional conditions for transitivity have been sought. In a very limited sense this point about transitivity is indeed correct. But it can be shown that the limited sense in which this is true ignores precisely the methodological point concerning the *interpretation* of revealed preference theory which I discussed a few minutes ago.

[1] Little, *ibid.*, p. 98.
[2] Samuelson, "Consumption Theory . . .", p. 243.

The philosophical issue involved is, therefore, worth discussing in the light of the logical problems raised by revealed preference theory. Consider a case in which we find a consumer choosing x and rejecting y, and another in which he is found to choose y and reject z. So he has revealed a preference for x over y and also for y over z. Of course, even under the assumption of transitivity of the underlying preference, the person is not obliged to *reveal* a preference for x over z since such a choice may not in fact arise in his uneventful life. But suppose we could offer this person choices over *any* combination of alternatives and could thus ensure that he had to choose between x and z. Then clearly it would be required by transitivity that he must choose x and reject z. Is this guaranteed by the Weak Axiom? The answer is: clearly yes.

To understand why this is so, imagine the contrary and suppose that he did choose z instead of x. We could then offer him the choice over the set of three alternatives, x, y and z. What could this man now choose? If he chose x, which would involve rejecting z, this would violate the Weak Axiom since he had earlier rejected x and chosen z. If he chose y, which would mean that he would be rejecting x, this would also violate the Weak Axiom since he had rejected y and chosen x earlier. Finally, if he chose z, which would imply a rejection of y, he would again be running counter to the Weak Axiom since earlier he had chosen y rejecting z. So no matter what he chose out of this set of three alternatives (x, y, z), he must violate the Weak Axiom. He is in this impasse only because he chose z and rejected x after having revealed a preference for x over y and for y over z. To be able to choose in a manner consistent with the Weak Axiom of Revealed Preference, he would have to choose x faced with a choice between the two.

Further, if he chose *both* x and z in a choice between the two, there must be inconsistency also. In a choice over (x, y, z), he could not choose z since he had chosen y rejecting z in a choice between the two. For the same reason he could not choose y since he had revealed a preference for x over y. So he would have to choose only x in the choice over x, y, z, rejecting z. But then he could not choose z in the presence of x in the choice over that pair in view of the Weak Axiom of Revealed Preference and this is a contradiction.

The Weak Axiom not only guarantees two-term consistency, it also prevents the violation of transitivity. The fact that the Axiom applies to two choices at a time does not rule out its repeated use to get the result of transitivity.

Why is it then that people have looked for stronger conditions than the Weak Axiom to get transitivity or similar properties? For example, Houthakker has proposed a condition, the so-called Strong Axiom of Revealed Preference, which demands more than the Weak Axiom of Samuelson to get us towards transitivity.[1] Similar conditions have been

[1] H. S. Houthakker, "Revealed Preference and the Utility Function", *Economica*, vol. 17 (1950). The Strong Axiom guarantees a property that Houthakker called semi-transitivity.

proposed by Ville, von Neumann and Morgenstern, and others.[1] Hicks, who noted that the Weak Axiom did make things fine for transitivity in a world of two goods only, proceeded to argue that "three-term inconsistency is only ruled out in the two-goods case by the special properties of that case".[2] But the simple argument we examined a few minutes ago assumed nothing about there being only two goods. What explains this mystery?

The clue lies in the fact that in the revealed preference literature, it has been customary to assume, usually implicitly, that the Weak Axiom holds only for those choices that can be observed in the market and not necessarily for other choices.[3] And given divisible commodities, the market can never offer the man under observation the choice, say, between x, y and z only. If these three baskets of goods were available then so should be an infinite number of other baskets that would cost no more at given market prices. This is how in the theory of consumer's behaviour, the man can get away satisfying the Weak Axiom over all the cases in which his behaviour can be observed in the market and nevertheless harbour an intransitive preference relation.

The moment this is recognized the question arises: why this distinction between those choices in which the person's behaviour can be observed in the market and other choices in which it cannot be? Presumably, the argument lies in the fact that if market choices are the only observable choices, then the Weak Axiom can be verified only for those choices and not for others that cannot be observed in the market. But as we saw earlier, the Weak Axiom cannot be verified anyway even for market choices and the case for its use lies not in verification but in its intuitive plausibility given the preference-based interpretation of choice. And there is no reason whatsoever to expect that the Weak Axiom is more plausible for "budget triangles" thrown up by market choice situations than for other choices that cannot be observed in the market; at any rate I have not seen any argument that has been put forward justifying such a dichotomy. The distinction lies only in the verification question and that, as we have seen, is really a red herring.

Treated as an *axiom* in the light of which consumer's choices are analysed and interpreted, rather than as a *hypothesis* which is up for verification, there is no case for restricting the scope of the Weak Axiom arbitrarily to budget sets only, and in the absence of this invidious distinction, transitivity follows directly from the Weak Axiom of Revealed Preference. If a consumer has chosen x rejecting y in one case, chosen y rejecting z in another, and chosen z rejecting x in a third case, then he has not only violated transitivity, he must violate the Weak

[1] J. Ville, "Sur les conditions d'existence d'une ophélimité totale et d'un indice du niveau des prix", *Annales de l'Université de Lyon*, vol. 9 (1946); J. von Neumann and O. Morgenstern, *Theory of Games and Economic Behaviour*, Princeton, 1944.
[2] Hicks, *op. cit.*, p. 110.
[3] *Cf.* D. Gale, "A Note on Revealed Preference", *Economica*, vol. 27 (1960).

Axiom of Revealed Preference as well. No matter what he chooses given the choice over x, y and z, he must run counter to the Weak Axiom, as demonstrated. The fact that he cannot be observed in a choice over (x, y, z) makes no real difference since *no matter what he chooses* he must logically violate the Weak Axiom.

In this sense, an observed violation of the Strong Axiom will logically imply a violation of the Weak Axiom as well. A number of other distinct axioms that have been proposed in the literature can also be shown to be equivalent once the arbitrary restrictions are removed.[1] (If the domain of the choice function includes all pairs and triples, then these apparently different axioms turn out to be logically equivalent.)

IV

I would now like to turn to the fundamental assumption of the revealed preference approach, viz, that people do reveal their underlying preferences through their actual choices. Is this a reasonable presumption? If a person chose x when y was available, it would seem reasonable to argue that he did not really regard y to be better than x. There is, of course, the problem that a person's choices may not be made after much thinking or after systematic comparisons of alternatives. I am inclined to believe that the chair on which you are currently sitting in this room was not chosen entirely thoughtlessly, but I am not totally persuaded that you in fact did choose the particular chair you have chosen through a careful calculation of the pros and cons of sitting in each possible chair that was vacant when you came in. Even some important decisions in life seem to be taken on the basis of incomplete thinking about the possible courses of action, and the hypothesis of revealed preference, as a psychological generalization, may not be altogether convincing. These questions are well-known as also are the difficulties arising from open or hidden persuasion involved in advertisements and propaganda, which frequently mess up not only one's attitude towards the alternatives available but also towards the act of choice itself. These problems are important, but I shall not go into them any further, partly because they have been much discussed elsewhere, but also because I have no competence whatever to throw light on the psychological issues underlying these problems. Instead I shall try to discuss one and a half other issues which seem to me to be also important. The half issue should perhaps come first.

The logical property of connectedness (or completeness as it is sometimes called) of binary relations is an important characteristic to examine in the context of evaluating the fundamental assumption of revealed preference. Connectedness of preference requires that between any two alternatives x and y, the person in question either prefers x to y,

[1] See K. J. Arrow, "Rational Choice Functions and Orderings", *Economica*, vol. 26 (1959); and A. K. Sen, "Choice Functions and Revealed Preference", *Review of Economic Studies*, vol. 38 (1971).

or prefers y to x, or is indifferent between x and y. The approach of revealed preference makes considerable use of connectedness. If a person chooses x rather than y, it is presumed that he regards x to be at least as good as y, and not that may be he has no clue about what to choose and has chosen x because he had to choose something.

The point can be illustrated with a variation of the classic story of Buridan's ass. This ass, as we all know, could not make up its mind between two haystacks; it liked both very much but could not decide which one was better. Being unable to choose, this dithering animal died ultimately of starvation. The dilemma of the ass is easy to understand, but it is possible that the animal would have agreed that it would have been better off by choosing either of the haystacks rather than nothing at all. Not choosing anything is also a choice, and in this case this really meant the choice of starvation. On the other hand, had it chosen either of the haystacks, it would have been presumed that the ass regarded that haystack to be at least as good as the other, which in this version of the story was not the case. The ass was in a real dilemma *vis-à-vis* the revealed preference approach.

The traditional interpretation of the story is that the ass was indifferent between the two haystacks. That indifference may be a cause for dithering has often been stated. For example, Ian Little prefaced his closely reasoned attack on the concept of indifference by posing the rather thoughtful question: "How long must a person dither before he is pronounced indifferent?"[1] But in fact there is hardly any real cause for dithering if one is *really* indifferent, since the loss from choosing one alternative rather than another is exactly zero. The person can choose either alternative and regret nothing in either case. This, however, is not the case if the preference relation is unconnected over this pair, i.e. if the chooser can neither say that he prefers x to y, nor y to x, nor that he is indifferent between the two.

If Buridan's ass was indifferent, choosing either haystack would have been quite legitimate and would not have misled the observer armed with revealed preference theory provided the observer chose a version of the theory that permitted indifference.[2] The real dilemma would arise if the ass had an unconnected preference. Choosing either haystack would have appeared to reveal a view that that haystack was no worse than the other, but this view the ass was unable to subscribe to since it could not decide what its preference should be. By choosing either haystack it would have given a wrong signal to the revealed preference theorist since this would have implied that he regarded the chosen haystack to be at least as good as the other. There is very little doubt that Buridan's ass died for the cause of revealed preference, though—alas—he was not entirely successful since non-choice leading to starva-

[1] Little, *op. cit.*, p. 92.

[2] See Arrow, *op. cit.*; Sen, *op. cit.*,; H. Herzberger, "Ordinal Choice v. Rationality", *Econometrica*, forthcoming; C. R. Plott, "Path Independence, Rationality and Social Choice", *Econometrica*, forthcoming.

tion would have looked like the chosen alternative, at any rate from the point of view of mechanical use of the fundamental assumption of revealed preference. There was no way the ass could have rescued that assumption given its unconnected preference.

But what if all these problems are ruled out? That is, if the person has a connected preference relation, takes his decisions deliberately after considering all alternatives, and is not swayed to and fro by the lure of advertisements. Obviously none of the problems discussed in the last few minutes will then arise. Will the life of the revealed preference theorist, then, be uncomplicated? I fear that it will not, and there is, it seems to me, a difficulty in some sense more fundamental than all the ones discussed so far. This problem I would like to go into now.

The difficulty is seen most easily in terms of a well-known game, viz, "the Prisoners' Dilemma",[1] which has cropped up frequently in economics in other contexts. The story goes something like this. Two prisoners are known to be guilty of a very serious crime, but there is not enough evidence to convict them. There is, however, sufficient evidence to convict them of a minor crime. The District Attorney—it is an American story—separates the two and tells each that they will be given the option to confess if they wish to. If both of them do confess, they will be convicted of the major crime on each other's evidence, but in view of the good behaviour shown in squealing, the District Attorney will ask for a penalty of 10 years each rather than the full penalty of 20 years. If neither confesses, each will be convicted only of the minor crime and get 2 years. If one confesses and the other does not, then the one who does confess will go free and the other will go to prison for 20 years.

What should the prisoners do? It is not doubted by the game theorist that any self-respecting prisoner will begin by drawing a pay-off matrix to facilitate rational choice. The table of pay-offs will look something like this. (The first number in each slot is the sentence of prisoner 1 and the second of prisoner 2. The numbers are negative to remind us that the prisoners dislike going to prison.)

		Prisoner 2	
		Confess	Not Confess
Prisoner 1	Confess	−10, −10	0, −20
	Not Confess	−20, 0	−2, −2

Each prisoner sees that it is definitely in his interest to confess no matter what the other does. If the other confesses, then by confessing himself this prisoner reduces his own sentence from twenty years to ten.

[1] See R. D. Luce and H. Raiffa, *Games and Decisions*, New York, 1958; also A. Rapoport, *Two Person Game Theory*, Michigan, 1966.

If the other does not confess, then by confessing he himself goes free rather than getting a two year sentence. So each prisoner feels that no matter what the other does it is always better for him to confess. So both of them do confess guided by rational self-interest, and each goes to prison for ten years. If, however, neither had confessed, both would have been in prison only for two years each. Rational choice would seem to cost each person eight additional years in prison.

This game has been much discussed in the literature of resource allocation as an illustration of the failure of individualistic decision taking and as a justification of a collective contract. It has an obvious bearing on the theory of optimum savings, on taxation theory, on allocation decisions involving externalities and public goods, and on a number of related issues.[1] Through a collective contract the group of individuals can do better than what they will do under individualistic action. The distinction has something to do with Rousseau's contrast between "the general will" and "the will of all", and with the necessity of a "social contract" to achieve what the general will wills.[2] In the particular story of the prisoners' dilemma, the general will can be interpreted to be the rule of non-confession which is beneficial for both, and the vehicle for achieving this will be a mutual non-confession treaty. If such a social contract can be accepted and enforced, both prisoners will be better off. So far so good. But what if no such contract can be arrived at? Are the prisoners doomed to suffer a heavy penalty constrained by their own rational choice calculus?

It is possible to argue that this is precisely the type of situation in which moral rules of behaviour have traditionally played an important part. Situations of the type of the prisoners' dilemma occur in many ways in our lives and some of the traditional rules of good behaviour take the form of demanding suspension of calculations geared to individual rationality. In different periods of history in different social situations in response to different types of problems particular rules of behaviour have been proposed which have in common the analytical property of trying to generate the results of a social contract without there being any such formal contract. Behavioural rules to handle problems of interdependence, arising in specific social and economic formations, can be seen in such diverse approaches as Christian or Buddhist ethics on the one hand and the philosophy of the Chinese "cultural revolution" on the other. I shall have a bit more to say on this presently, but the implication of all this for the theory of revealed preference should be first spelt out.

[1] See W. J. Baumol, *Welfare Economics and the Theory of the State*, Cambridge, Mass., 1952; A. K. Sen, "On Optimizing the Rate of Saving", *Economic Journal*, vol. 71 (1961); S. A. Marglin, "The Social Rate of Discount and the Optimal Rate of Investment", *Quarterly Journal of Economics*, vol. 77 (1963); A. K. Sen, "A Game Theoretic Analysis of Theories of Collectivism in Allocation", in T. Majumdar (ed.), *Growth and Choice*, Bombay, 1969.

[2] W. G. Runciman and A. K. Sen, "Games, Justice and the General Will", *Mind*, vol. 74 (1965); J. Rawls, *A Theory of Justice*, Cambridge, Mass., 1971.

Suppose each prisoner in the dilemma acts not on the basis of the rational calculations outlined earlier but proceeds to follow the dictum of not letting the other person down irrespective of the consequences for himself. Then neither person will confess and they will both get off lightly. Now, consider the job of the observer trying to guess the preferences that have been revealed by the choice of non-confession. There is, of course, an element of uncertainty in the exercise of choice that the prisoners face, for neither of them knows what the other is up to. It should be clear, however, that if there is anything in the assumption of revealed preference as it stands, it must be presumed that each prisoner prefers at least one of the possible outcomes resulting from his non-confession to what would have happened had he confessed, given other things. That is, either he prefers the consequence of his not confessing given the other prisoner's non-confession, or the consequence of his not confessing given the other prisoner's confession. But in fact neither happens to be true. The prisoner does not prefer to go to prison for twenty years rather than for ten; nor does he prefer a sentence of two years to being free. His choice has not revealed his preference in the manner postulated.

At this stage a couple of warnings may be worth stating since the point that is being made can be easily misunderstood. The prisoners' non-confession will be quite easy to put within the framework of revealed preference if it were the case that they had so much concern for the sufferings of each other that they would choose non-confession on grounds of joint welfare of the two. There is indeed nothing extraordinary in assuming that a person may prefer that both should go to prison for two years each rather than that the other should suffer twenty years while he himself goes free. The problem arises precisely because that is *not* being assumed. Each is assumed to be self-centred and interested basically only in his own prison term, and the choice of non-confession follows *not* from calculations based on this welfare function, but from following a moral code of behaviour suspending the rational calculus. The preference is no different in this case from that in the earlier example, but behaviour is. And it is this difference that is inimical to the revealed preference approach to the study of human behaviour.

A second point to note is that the entire problem under discussion can be easily translated into the case in which each person does worry about the other's welfare as well and is not concerned only with his own welfare. The numbers in the pay-off matrix can be interpreted simply as welfare indices of the two persons and each person's welfare index can incorporate concern for the other. The prisoners' dilemma type of problem can arise even when there is concern for each other.

Third, no special importance should be attached to the specific story of the prisoners in terms of which this particular analytical problem is expounded. The interest in the prisoners' dilemma lies not in the fiction which gives the problem its colour, but in the existence of a strictly

dominant strategy for each person which together produce a strictly inferior outcome for all. One feature of the prisoners' dilemma is, in fact, particularly misleading. This concerns the complete symmetry of the positions of the two players. Some suggestions for the resolution of the dilemma within the framework of rational choice make considerable use of this particular feature,[1] but even with asymmetrical prison sentences as long as the orderings of the penalties are the same we can get exactly the same dilemma and the same implications for revealed preference theory.

V

The concentration on the contractual side of the prisoners' dilemma has perhaps tended to obscure the important implications of this type of situation for the relation between choice and preference. If the prisoners agree to a non-confession treaty and if that treaty can be enforced the prisoners will indeed get off the hook, but such a contract may be difficult to devise and conceivably impossible to enforce under certain circumstances. When it comes to the use of this type of model in economics in interpreting problems of resource allocation, one can distinguish between those situations in which a contract may be easy to operate and those cases in which it will be far from easy to do so.

I am concerned here with cases in which a contractual solution is not possible. This corresponds to the case in which the prisoners are not bound by any contract but nevertheless decide not to confess. The essence of the problem is that if both prisoners behave *as if* they are maximizing a different welfare function from the one that they actually have, they will both end up being better off even in terms of their *actual* welfare function. To take the extreme case, if both prisoners try to maximize the welfare of the other, neither will confess in the case outlined since non-confession will be a superior strategy no matter what is assumed about the other person's action. The result of each trying to maximize the welfare of the other will, therefore, lead to a better situation for each in terms of his own welfare as well. It is not necessary that the prisoners in fact have this much concern—or indeed any concern—for the other, but if they behave *as if* they have this concern, they will end up being better off in terms of their real preference. This is where the revealed preference approach goes off the rails altogether. The behaviour pattern that will make each better off in terms of their real preferences is not at all the behaviour pattern that will *reveal* those real preferences. Choices that reveal individual preferences may be quite inefficient for achieving welfare of the group.

I would argue that the philosophy of the revealed preference approach essentially underestimates the fact that man is a social animal and his

[1] See Rapoport, *op. cit.*; and J. W. Watkins, "Self-Interest and Morality in the Light of the Prisoners' Dilemma", paper read at the Bristol Conference on "Practical Reason", September 1972, to be published in a volume edited by Professor S. Körner.

choices are not rigidly bound to his own preferences only. I do not find it difficult to believe that birds and bees and dogs and cats do reveal their preferences by their choice; it is with human beings that the proposition is not particularly persuasive. An act of choice for this social animal is, in a fundamental sense, always a social act. He may be only dimly aware of the immense problems of interdependence that characterize a society, of which the problem under discussion is only one. But his behaviour is something more than a mere translation of his personal preferences.

VI

In economic analysis individual preferences seem to enter in two different roles: preferences come in as determinants of behaviour and they also come in as the basis of welfare judgments. For example, in the theory of general equilibrium the behaviour of individuals is assumed to be determined by their respective preference orderings, and problems of existence, uniqueness and stability of an equilibrium are studied in the context of such a framework. At the same time, the optimality of an equilibrium, i.e. whether the market can lead to a position which yields maximal social welfare in some sense, is also examined in terms of preference with the convention that a preferred position involves a higher level of welfare of that individual.[1] This dual link between choice and preference on the one hand and preference and welfare on the other is crucial to the normative aspects of general equilibrium theory. All the important results in this field depend on this relationship between behaviour and welfare through the intermediary of preference.

The question that is relevant in this context is whether such heavy weight can be put on the slender shoulders of the concept of preference. Certainly, there is no remarkable difficulty in simply defining preference as the underlying relation in terms of which individual choices can be explained; provided choices satisfy certain elementary axioms, the underlying relation will be binary, and with some additional assumptions it will be an ordering with the property of transitivity.[2] In this mathematical operation preference will simply be the binary representation of individual choice. The difficulty arises in interpreting preference thus defined as preference in the usual sense with the property that if a person prefers x to y then he must regard himself to be better off with x than with y. As illustrated with the example of the prisoners' dilemma, the behaviour of human beings may involve a great deal more than

[1] See J. R. Hicks, *Value and Capital*, Oxford, 1939; P. A. Samuelson, *Foundations of Economic Analysis*, 1947; G. Debreu, *Theory of Value*, New York, 1959; K. J. Arrow and F. H. Hahn, *General Competitive Analysis*, San Francisco and Edinburgh, 1971.

[2] The respective conditions for binariness, transitivity of strict preference and full transitivity are presented in Sen, "Choice Functions . . .". For the conditions that guarantees a numerical representation of the individual welfare function based on their preference relation, see Debreu, *op. cit.*; M. K. Richter, "Revealed Preference Theory", *Econometrica*, vol. 34 (1966); Arrow and Hahn, *op. cit.*

maximizing gains in terms of one's preferences and the complex interrelationships in a society may generate mores and rules of behaviour that will drive a wedge between behaviour and welfare. People's behaviour may still correspond to some consistent *as if* preference but a numerical representation of the *as if* preference cannot be interpreted as individual welfare. In particular, basing normative criteria, e.g. Pareto optimality, on these *as if* preferences poses immense difficulties.

To look at the positive side of the issue, the possibilities of affecting human behaviour through means other than economic incentives may be a great deal more substantial than is typically assumed in the economic literature. The rigid correspondence between choice, preference and welfare assumed in traditional economic theory makes the analysis simpler but also rules out important avenues of social and economic change. An example may make the point clearer.

Suppose it is the case that there are strong environmental reasons for using glass bottles for distributing soft drinks (rather than single-use steel cans) and for persuading the customers to return the bottles to the shops from where they buy these drinks (rather than disposing of them in the dustbin). For a relatively rich country the financial incentives offered for returning the bottles may not be adequate if the consumers neither worry about the environment nor are thrilled by receiving back small change. The environment affects the life of all, true enough, but from the point of view of any one individual the harm that he can do to the environment by adding his bottles to those of others will be exceedingly tiny. Being generally interested in the environment but also being lazy about returning bottles, this person may be best off if the others return bottles but not he, next best if all return bottles, next best if none does, and worst of all if he alone returns bottles while others do not. If others feel in a symmetrical way we shall then be in a prisoners' dilemma type situation in which people will not return bottles but at the same time all would have preferred that all of them should return bottles rather than none. To tackle this problem, suppose now that people are persuaded that non-return is a highly irresponsible behaviour, and while the individuals in question continue to have exactly the same view of their welfare, they fall prey to ethical persuasion, political propaganda, or moral rhetoric. The welfare functions and the preference relations are still exactly the same and all that changes is behaviour. The result is good for the environment but sad for the theory of revealed preference.

I am not, of course, arguing that a change in the sense of responsibility is the *only* way of solving this problem. Penalizing non-return and highly rewarding return of bottles are other methods of doing this, as indeed will occur to any economist. In this particular case, these methods can also be used quite easily (since the problem of checking is not serious with the return of bottles), even though any system of payments and rewards also involve other issues like income distribution. The real difficulty arises when the checking of people's actions is not easy.

Examples of these cases vary from such simple acts as littering the streets to such complex behaviour as paying one's taxes.

VII

To avoid a possible misunderstanding, I would like to distinguish clearly between four possible cases all of which involve the same choice (e.g. the use and reuse of glass bottles) but the underlying preferences have different interpretations:

(1) The person simply prefers using glass bottles rather than steel cans from a purely self-regarding point of view, e.g. because he likes glass, or (perhaps somewhat incredibly) he believes the impact on environment of his using single-use steel cans (*given* the choices of others) will hurt him significantly.

(2) The person is worried about the welfare of others as well and his own welfare function includes concern for other people's welfare,[1] and he reuses glass bottles because he takes the hurt on others as hurt on himself.

(3) The person's concern for other people's welfare reflected in his notion of his own welfare would not be sufficient to prevent him from using single-use steel cans if he could do it on the sly, but he is afraid of the social stigma of being seen to do the "wrong" thing, or afraid of others emulating him in doing the "wrong" thing and thereby his getting hit indirectly.

(4) The person can do the "wrong" thing on the sly without being noticed and he feels that if he did that he personally would be better off (even after taking note of whatever weight he might wish to put on the welfare of the others), but he feels that he would be acting socially irresponsibly if he did proceed to do it, and therefore does not do so.

I am primarily concerned with case (4), even though case (3) would also pose some problem for revealed preference theory (and the normative aspects of general equilibrium) since preferences are not usually defined on the space of stigmas and such things, and identical commodity choices will involve quite different welfare levels depending on the reaction of others. But case (3) can be, in principle, taken care of through a suitable redefinition of the domain of choice. Case (4) poses a more serious difficulty and it is with this case that I am concerned.

It is, of course, perfectly possible to argue that actions based on considerations of social responsibility as opposed to one's own welfare do reflect one's "ultimate" preferences, and in a certain sense this is undoubtedly so. The question is whether the identification of welfare with preference (in the sense of the former being a numerical representation of the latter) will survive under this interpetation. The problem arises from the dual link-up between choice and preference on the one hand and preference and welfare on the other. Preference can be quite

[1] *Cf.* A. K. Sen, "Labour Allocation in a Cooperative Enterprise", *Review of Economic Studies*, vol. 33 (1966).

reasonably defined in such a way as to maintain one or the other, but the issue is whether *both* can be maintained through some definition of preference, and it is this dual role that I am trying to question here.[1]

With what frequency do problems of the kind of case (4) arise? I do not know the answer to this question. It seems clear, however, that they arise often enough to be worried about their implications for traditional economic analysis. Moral considerations involving the question "if I do not do it, how can I morally want others to do it?", do affect the behaviour of people. The "others" involved may be members of narrowly defined groups or classes, or widely defined societies, but such considerations do have a role in influencing choice.[2]

What harm would there be, it might be asked, in identifying welfare with what is revealed by a person's choices, even if that is not what he would claim to be his welfare as he himself sees it? Apart from the danger of being misled by the confusing use of words, like "preference" or "welfare", which have some specific meanings as used in normal communication, there are also some difficulties for normative economics in basing optimality criteria (e.g. Pareto optimality) on *as if* preferences. There is a distinction from the point of view of social judgment between the relevance of a choice made under a moral sense of social responsibility and that made under a straightforward pursuit of one's welfare (including any pleasure one takes in the happiness of others). The identification of welfare with *as if* preferences blurs this distinction and withholds relevant information from the analysis of social welfare and collective choice.

VIII

An interesting illustration of the problem of the relation between preference, choice and social responsibility can be seen in the recent Chinese debates on the use of financial incentives in the allocation of labour in communal agriculture. During the so-called Great Leap Forward in 1958–60 the Chinese tried to reduce drastically the use of work rewards and raised very substantially the proportion of income distributed in the communes on other criteria such as the size of the family. In the absence of what the Chinese called "socialist consciousness", a system of this kind produces precisely the prisoners' dilemma type of problem. Each may prefer that others should work hard, but

[1] The problem discussed here should not be confused with the important but different problem of strategic reasons for "not revealing preference" (see, for example, T. Majumdar, *The Measurement of Utility*, London, 1958; R. Farquharson, *Theory of Voting*, Oxford, 1970). The latter is a problem of establishing correspondence between rankings of the outcome space and those of the strategy space. Our problem arises in the ranking of the outcome space itself.

[2] I have tried to argue elsewhere that there are advantages in viewing moral judgments not as one other ordering of actions or outcomes but as an ordering (or a quasi-ordering) of orderings of actions or outcomes. "Choice, Orderings and Morality", paper read at the Bristol Conference on "Practical Reason", September 1972, to be published in a volume edited by Professor S. Körner.

given the actions of others may prefer to take it easy oneself, even though given the choice between all working hard and none doing so people may prefer the former. A social contract of sincere efforts by all is easy to think of but difficult to enforce, given the difficulties of supervision of the intensity of work.

At the end of the Leap Forward period this experiment was abandoned, or drastically cut, and it was generally thought that the experiment was premature. The use of financial incentives was again expanded. How much of the difficulties of the Leap Forward period arose from this attempt at dissociating work from material incentives is not known clearly, but it certainly did not make things any easier.

After the end of the Leap Forward period there have been several further attempts to move away from material incentives. Meanwhile the Chinese also tried out a programme of reorientation of behaviour patterns. The well-known "cultural revolution" put particular emphasis (as the so-called "Sixteen Points" explained) on "an education to develop morally, intellectually and physically and to become labourers with socialist consciousness."[1] The relation of all this to the problem of work motivation is, of course, very close, and I have tried to discuss it elsewhere.[2] Briefly, this can take one of two forms, viz, either (i) a reorientation of the individual welfare functions of the people involved, or (ii) a different basis of behaviour emphasizing social responsibility whether or not individual welfare functions are themselves revised. In practice it was probably a mixture of both.

How successful the Chinese experimentation has been in the reorientation of behaviour patterns, it is difficult to assess fully at this stage. What is, anyway, important for our purposes is to note the relevance of this experiment on work motivation in China to the problem of the relation between choice, preference and welfare.

IX

I should perhaps end with a critical observation on what tends to count as hard information in economics. Much of the empirical work on preference patterns seems to be based on the conviction that behaviour is the only source of information on a person's preferences. That behaviour is a major source of information on preference can hardly be doubted, but the belief that it is the only basis of surmising about people's preferences seems extremely questionable. While this makes a great deal of sense for studying preferences of animals, since direct communication is ruled out (unless one is Dr. Dolittle), for human beings surely information need not be restricted to distant observations of choices made. There is, of course, something of a

[1] For a penetrating analysis of work motivation in China, see C. Riskin, "Maoism and Motivation: A Discussion of Work Motivation in China", mimeographed, 1972, forthcoming in the *Bulletin of Concerned Asian Scholars*.

[2] A. K. Sen, *On Economic Inequality*, Oxford, 1973, chapter 4.

problem in interpreting answers to questions as correct and in taking the stated preference to be the actual preference, and there are well-known limitations of the questionnaire method. But then there are problems, as we have seen, with the interpretation of behaviour as well. The idea that behaviour is the one real source of information is extremely limiting for empirical work and is not easy to justify in terms of the methodological requirements of our discipline.

There is an old story about one behaviourist meeting another, and the first behaviourist asks the second: "I see you are very well. How am I?" The thrust of the revealed preference approach has been to undermine thinking as a method of self-knowledge and talking as a method of knowing about others. In this, I think, we have been prone, on the one hand, to overstate the difficulties of introspection and communication, and on the other, to underestimate the problems of studying preferences revealed by observed behaviour.

X

Perhaps I should now gather together the main themes that I have tried to develop in this lecture. First, I have tried to argue that the interest of revealed preference theory lies in the skilful use of the assumption that behaviour reveals preference and not, despite claims to the contrary, in explaining "behaviour without reference to anything other than behaviour".

Second, if revealed preference is interpreted in this light, some of the additional axioms of revealed preference theory can be seen to be redundant for the purpose for which they are used. For example, the Weak Axiom of Revealed Preference can be seen to be quite strong and certainly sufficient for transitivity without requiring a stronger axiom.

Third, the fundamental assumption about the revelation of preference can be criticized from many points of view, including the possibility that behaviour may not be based on systematic comparison of alternatives. More interestingly, the person in question may not have a connected preference pattern and in terms of observation it is difficult to distinguish such incompleteness from indifference.

Fourth, even if all these problems are ruled out, there remains a fundamental question of the relation between preference and behaviour arising from a problem of interdependence of different people's choices which discredits individualistic rational calculus. The problem was illustrated in terms of the game of the prisoners' dilemma. The usual analysis of the prisoners' dilemma has tended to concentrate on the possibility of a collective contract, but in many problems such a contract cannot be devised or enforced. Even in the absence of a contract, the parties involved will be together better off following rules of behaviour that require abstention from the rational calculus which is precisely the basis of the revealed preference theory. People may be induced by social codes of behaviour to act *as if* they have different preferences from what they really have. This type of departure may also be stable for those

codes since such behaviour will justify itself in terms of results from the point of view of the group as a whole.

Finally, this problem has an important bearing on normative problems of resource allocation formulated in terms of the dual link between choice and preference and between preference and welfare. The type of behaviour in question drives a wedge between choice and welfare, and this is of relevance to general equilibrium theory as well as to other aspects of normative economics. Preference can be defined in such a way as to preserve its correspondence with choice, or defined so as to keep it in line with welfare as seen by the person in question, but it is not in general possible to guarantee both simultaneously. Something has to give at one place or the other.

The London School of Economics.