Economics and Philosophy, **21** (2005) 33–50
doi:10.1017/S0266267104000379

SYMPATHY, COMMITMENT, AND PREFERENCE

Daniel M. Hausman
University of Wisconsin-Madison

This paper will combine exegesis of what Sen takes preference, sympathy, and commitment to be with an alternative conceptualization that is directed toward Sen's own ends but is in some ways more conciliatory toward orthodox approaches.

1. TWO CONCEPTS OF PREFERENCE

The recent opportunity to reread Amartya Sen's many writings on preference and choice as a unit increased my appreciation of their depth and intricacy and of the wit and humanity of their author. It also made me realize that I had persistently misread them, mistakenly substituting my own notion of "preference" for Sen's. While still very much in Sen's camp in rejecting revealed preference theory and emphasizing the complexity, incompleteness, and context dependence of preference and the intellectual costs of supposing that all the factors influencing choice can be captured by a single notion of preference, I shall contest his view that economists should recognize multiple notions of preference. In my view, Sen's concerns are better served by embracing a single conception of preference and insisting on the need for analysis of the multiple factors that determine "preference" so conceived.

Like most economists and philosophers, Sen offers no explicit definition of "preference." Instead, he consistently emphasizes that economists have used the word "preference" to refer to many different things. Among these different concepts of preference, Sen believes that two are most important, and his discussion of preference, especially in his earlier works (Sen 1973, 1977) focuses on these. He writes:

> Certainly, there is no remarkable difficulty in simply defining preference as the underlying relation in terms of which individual choices can be explained . . . 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. (1973: 67)

One definition of preference is as "the underlying relation in terms of which individual choices can be explained," which Sen here identifies with "the binary representation of individual choice." Call this concept of preference "choice ranking." Second, there is what Sen labels "the usual sense" of preference. In this sense a person prefers x to y if and only if the person believes that he or she is better off with x than with y.[1] Let us call this sense of preference "expected advantage ranking." Unlike rankings, preferences are attitudes, which can be intense or weak, cool or emotional, but it will simplify this paper to identify preferences with the rankings they imply.

Sen has argued again and again that these two notions of preference – these two rankings in terms of choice and expected advantage – should not be conflated.

> the normal use of the word permits the identification of preference with the concept of being better off, and at the same time it is not quite unnatural to define "preferred" as "chosen". I have no strong views on the "correct" use of the word "preference", and I would be satisfied as long as both uses are not *simultaneously* made, attempting an empirical assertion by virtue of two definitions. (1977: 329)

As this last quotation makes clear, Sen has consistently avoided legislating the meanings of words (see, for example, Sen 1991a: 588; 1991b). Rather than arguing for some canonical usage for the word, "preference," his message has been that economists need to recognize that the term has many meanings. Yet Sen has substantive objections to the theory of revealed preference and hence to the use of the choice-ranking concept of preference (1973, 1993). That leaves expected advantage as the main concept of preference, and at least in his earlier works, when Sen uses the word "preference" without specifying its sense, one should probably interpret it as expected advantage. Sen often shies away from using the word, "preference," employing instead terms such as "desires" or "goals."

Although expected advantage seems to be Sen's default concept of preference, particularly in his earlier works, he never argues that this is what economists typically mean by "preference" or that it is what they

[1] In the quotation above, Sen takes expected advantage to be a necessary condition for preference, rather than necessary and sufficient, and one can read his words as making a claim about preference "in the usual sense" (whatever that may be) rather than as defining a concept of preference. But other comments make clear that Sen regards expected advantage as a competing *definition* of preference. He writes, for example, "Preference can be defined so 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" (1973: 73). See also Sen (1980: 442).

should mean by preference. He emphasizes that in everyday usage, people have "preferences" over alternatives that are irrelevant to their interests or even contrary to their advantage. Actual usage is multiply ambiguous, and Sen is not prepared to argue for any regimentation of usage. His reason for counseling awareness of ambiguity rather than proposing a cure is, I think, that he fears that a regimentation will encourage an overly simple and one-dimensional view of evaluation and choice.

2. THREE OTHER CONCEPTS OF PREFERENCE

People care about many things that do not bear on their own well-being, "interests", or advantage, and people may prefer to sacrifice their interests or well-being in order to accomplish something that matters to them more. Such "preferences," whose satisfaction does not serve self-interest, cannot be conceived of as rankings in terms of expected advantage. So far, the only alternative sense of preference we have considered is choice ranking. But the theory of revealed preference – the view that preference is the ranking of alternatives that is entailed by choice – rarely captures what economists mean by "preference," let alone ordinary usage. To give some sense of the difficulties, consider, for example, my preference for a state of affairs in which there are no wars on earth in the twenty-second century over a state of affairs in which there are wars (and not because nuclear warfare will have reduced the planet to a cinder). Whether there is war in the twenty-second century or not does not bear on my well-being or self-interest. So one cannot read this as my belief that the absence of war in the twenty-second century will be better for me. But my preference does not appear to be a choice ranking either, because I cannot choose whether there will be wars in the twenty-second century. One might argue that this preference is implicit in my choice of virtually any candidate over George Bush for president in 2004, but presumably many of his misguided supporters have the same preferences as I do with respect to future peace. There is obviously no one-to-one relationship between this preference and any choices that I face.

One response is to link preference to *hypothetical* choice. To say that I prefer that there be no wars in the twenty-second century is to say that if I could, I would choose that state of affairs over one with wars. Indeed, many of those who claim to be defenders of revealed-preference theory implicitly reject the behaviorist revealed-preference theory defended by Paul Samuelson or Ian Little in the 1930s, 1940s and 1950s and (misleadingly, in my opinion) regard a hypothetical-choice view of preference as itself a formulation of revealed-preference theory. What is most important is that choice rankings not be confused with this third notion of preferences as the rankings implicit in hypothetical choices – not which theory is called "revealed-preference theory."

Although a hypothetical choice view of preference can cope with my preference that there not be wars in the next century, it is inadequate for many economic applications of the notion of preference. When modeling a strategic interaction among individuals as a game, theorists must assign utilities – indices of preference – to the outcomes of their interaction. But the "outcomes" which are the objects of preference are not alternatives among which a player can choose. (They are, in Sen's terminology, "comprehensive outcomes" (1997b: 745) – paths through games including their results). What people choose are strategies and these depend on both beliefs about what others will do and preferences as evaluations of outcomes, rather than preferences either in the sense of choice rankings or in the sense of rankings of hypothetical choices. For example, consider a traditional marriage proposal. This can be modeled as a game in which first Darcy proposes or not, and then Elizabeth accepts or refuses his proposal. Darcy prefers the outcome where he proposes and Elizabeth accepts to the outcome where he proposes and she refuses. But Darcy could not have a choice between these outcomes. The utility numbers in the game are not indices reflecting hypothetical choices.

What other notions of preference are there? I agree with Sen that there are many, and I shall discuss two more. One of these is especially important, and it has a claim to be *the* central notion of preference in economics and decision theory. Economists should, I shall argue, be encouraged to regiment their language and reserve the word, "preference," for this single usage. In contrast to Sen, I think there is a great deal to be said for *prescribing* how the word "preference" should be used in economics.[2]

This notion of "preference," which I believe should be the only concept of preference employed in economics and decision theory, is a technical notion that does not conform to ordinary usage of the word, "preference." But it derives from a traditional folk-psychological view of human action. This traditional view holds that human actions can be explained and predicted by the beliefs and "desires" of agents. "Desire" in this context is a catch-all including a diverse array of motivating factors – emotions of all sorts, aversions, appetites, feelings of obligation – basically any mental state that "pushes" an agent. So when one cold Friday night, a hungry student named Ellen takes a frozen pizza out of a refrigerator, unwraps it, puts it in a stove, and turns knobs on the stove, we folk psychologists explain Ellen's action by Ellen's beliefs – including especially her beliefs that turning the knobs will cause the stove to heat the pizza – and by her desire to eat hot pizza.

This sort of explanation is familiar, but not very satisfactory. Ellen might also like to eat frozen pizza, or she might also have a desire to

[2] I am here following Broome (1991) and many others both in suggesting that a single usage be prescribed and in the particular usage I favor.

reheat some left-over meatloaf. Or she might rather skip dinner and keep studying decision theory. What explains her action is not merely desiring to eat hot pizza (plus possessing the requisite beliefs), but desiring to do this as much or more than she wants to do any of the feasible alternatives.

One way to tighten up the folk psychological account of action is to replace the non-comparative catch-all notion of a "desire" with a comparative catch-all notion of "preference." One can then explain the little interaction between Ellen and the stove in terms of physical constraints, Ellen's beliefs about the outcomes of the alternative actions she can undertake that Friday night, and her catch-all ranking of those outcomes. One explains the pizza warming by showing that Ellen ranks its expected outcome at least as highly as any feasible alternative.

With a bit of mathematical dressing up, this is close to the "official" story of choice and preference in mainstream economics. An agent's preferences consist of his or her overall evaluation of the objects over which preferences are defined. This evaluation implies a ranking of these objects with respect to everything that matters to the agent: desirability, social norms, moral principles, habits – everything relevant to evaluation. Preferences thus imply all-things-considered rankings. In my view, all-thing-considered ranking rather than choice ranking is "the underlying relation in terms of which individual choices can be explained" (Sen 1973: 67). It should be the single correct usage of the term "preference" in economics and decision theory.

The links economists draw between preference and advantage and between preference and choice, which Sen takes to be alternative conceptions of preference, should, I think, instead be regarded as substantive claims about all-things-considered preferences. In some circumstances, it is a reasonable approximation to maintain that people's all-things-considered ranking of alternatives match their ranking of alternatives in terms of expected advantage. In that case, one can make inferences about expected advantage from preferences and inferences about preferences from expected advantage. But one need not follow Sen and take expected advantage to be "the usual sense" – or indeed any *sense* at all – of preference either in economics or ordinary life. Similarly, when George's all-things-considered ranking is limited to his ranking of the objects among which he is choosing, then, if George's choice tracks his preferences, his preferences will match the ranking that is implicit in his choices. When one can coherently describe a possibility of George choosing between two alternatives, his all-things-considered ranking of those two alternatives will coincide with how he would choose. But the relevant notion of preference is as an all-things-considered ranking, not as either a choice-ranking or a hypothetical choice ranking.

The reason why economists must employ a concept of preference as all-things-considered rankings rather than as choice or hypothetical-choice

rankings is that they need to relate the ranking of objects of choice to beliefs and to evaluations of things that are not objects of choice. If all that consumer choice theorists could say about why an agent purchased one thing rather than another was that the consumer preferred to make that purchase, rather than relating the action to prices, income, and the consumer's preferences over the commodity space, there would be no Nobel Prize in economics. If all that game theorists could say about why individuals play one strategy rather than another is that they prefer that strategy, game-theory texts could be very short indeed.

Moreover, all-things-considered preferences depend on *beliefs* as well as on motivating factors. Unlike primitive urges, people's preferences depend on their beliefs concerning the character and consequences of the objects of their preferences.[3] For example, my preference for drinking rather than discarding a glass of clear liquid in front of me depends on whether I believe it is water or gasoline. This means that preferences can be regarded as both as the *result* of deliberation and as an *input* into deliberation. Belief thus enters into choice in two ways. Belief and preferences can be inputs that influence choices, as they are when my desire for water and my belief that the liquid in front of me is water lead me to drink. Beliefs and preferences can also influence preferences, as, for example, they are when my preferences among flavors and textures and my aversion to early death coupled with my beliefs about the consequences of alternative diets determine my preferences among alternative foods. When the objects of preferences are the alternative choices – that is *actions* – themselves, then belief plays its full role in the deliberations that result in "final" preferences among the choices, which coincide with what I called the choice ranking.

The notion of preference as an all-things-considered ranking is close to the everyday concept of preference, but not quite the same, because in everyday usage, "preference," like "desire" or "inclination" is often contrasted with *duty* or *principle*. Everyday usage thus often treats preference as a ranking in terms of some, rather than all, relevant evaluative considerations. The distinction between preference and duty does not coincide with the distinction between self-interested motives and motives that are not self-interested. Self-interested choices can be governed by principles of prudence that restrain impulses, while desires to help others may have no connection to principles. So none of the four concepts of preference discussed above: choice ranking, expected advantage, hypothetical choice ranking or all-things-considered ranking

[3] Adapting some useful terminology that Sen introduced in a different context (1970: ch. 5), one might distinguish "basic" preferences, which are independent of beliefs, from "non-basic" preferences that depend on beliefs. It is very difficult to give examples of basic preferences.

match this fifth ordinary-language notion of preference. "Preference" as used by economists is a technical term within economics, and neither the case for taking all-things-considered ranking as the notion of preference economists ought to employ nor Sen's case for recognizing and distinguishing multiple notions rests on claims concerning ordinary language usage.

3. SYMPATHY, COMMITMENT, AND SEN'S TWO CONCEPTS OF PREFERENCE

Sen recognizes that when economists speak of preferences, they refer to motivations of all sorts, but he is skeptical of "the common tendency to make 'preference' (or a general-purpose 'utility function') an all-embracing depository of a person's feelings, values, priorities, choices, and a great many other essentially diverse objects" (1991a: 589). He sees this tendency as a conflation of different notions of preference – that is, as a failure to draw necessary distinctions.

> A person is given *one* preference ordering, and as and when the need arises this is supposed to reflect his interests, represent his welfare, summarize his idea of what should be done, and describe his actual choices and behavior. Can one preference ordering do all these things? A person thus described may be "rational" in the limited sense of revealing no inconsistencies in his choice behavior, but if he has no use for these distinctions between quite different concepts, he must be a bit of a fool. (1977: 335–36)

The last sentence quoted here addresses the question of whether *agents* need to distinguish between their interests and obligations or between their wants and the claims of others, but the question at issue is whether *economists* need to draw these distinctions. The argument for the claim that economists need to draw such distinctions rests on Sen's demonstration that choice rankings do not coincide with expected advantage and that these rankings do not exhaust the factors that influence evaluations and choices. For this reason, Sen suggests that a single all-things-considered ranking should be replaced with a variety of rankings of which choice rankings and rankings in terms of expected advantage are merely two prominent instances.

A single distinction between choice rankings and rankings in terms of expected advantage is not nearly fine-grained enough to explain why people may be cutthroats at work, devoted parents at home, liberals at the voting booth, racists at the club, public-spirited at one moment, pious at another, principled before lunch, and utterly selfish afterwards. To make even a first stab at accounting for some reasonable portion of human behavior, economists should, Sen maintains, at the very least, also distinguish between what he calls "sympathy" and "commitment." Sympathy obtains when "the concern for others directly affects one's own

welfare" (1977: 326). Commitment is in contrast non-egoistic (1977: 326–67) and contrary to preference (at least in the sense of expected advantage) (1977: 327). In Sen's words, "If knowledge of torture of others makes you sick, it is a case of sympathy; if it does not make you feel personally worse off, but you think it is wrong and you are ready to do something to stop it, it is a case of commitment" (1977: 326).

On Sen's account of sympathy, helping someone from a simple desire to do so, when the agent neither anticipates nor achieves any (personal) benefit, is not a case of sympathy. Unless I expect to benefit from doing X, my doing X is not a case of sympathy. This is clear and unambiguous, and Sen even cautions the reader not to place too much weight on the particular words chosen. Yet I for one previously misread Sen and identified sympathy with altruistic motivation. One reason is that I also misunderstood what Sen meant by preference. If one interprets preference as all-things-considered ranking, then altruistic motivation accords with preference. But if one interprets preference as expected advantage, then altruistic motivation, unlike sympathy, is typically counterpreferential.

The way to think about sympathy is to recognize that among the many sources of George's expected advantage are states of affairs involving other things than George. For example, when somebody tramples George's roses, the harm done to the roses diminishes George's welfare. If roses were people, this would be a case of sympathy. If somebody instead tramples George's friend, any consequent lessening of George's welfare counts as sympathy. Sympathy is the way in which benefits and harms to other people register within self-interested preferences.

"Commitment" then covers some motivations other than expected advantage. Like altruism, commitment is counterpreferential if preference is identified with expected advantage. It is apparently not counter-preferential if preferences are all-things-considered rankings, because all-things-considered rankings already reflect all those factors that constitute any "commitments" of the agent. The examples Sen gives of commitment suggest that commitment involves adherence to principle, and readers (or at least this reader) have assimilated Sen's distinction between sympathy and commitment to the everyday contrast between action motivated by altruistic concerns and action motivated by adherence to principle.[4] This assimilation is mistaken, both because sympathy must be motivated by an expected benefit to the chooser and because Sen never explicitly restricts what he takes to be "commitment" to cases where principles govern choice. So I am not sure, for example, whether he would regard a sacrifice of my welfare motivated by a simple desire to help some particular person as an instance of commitment, or whether such action involves neither sympathy nor commitment. It is as faithful to Sen's characterization of

[4] Some philosophers have read more carefully. See for example Anderson (2001: 22–3).

commitment to take *any* non-self-interested motivation as an instance of "commitment" as it is to identify commitment and acting from principle.

Recognizing the existence of sympathy in Sen's sense enables one to square behavior that helps others with a model of individuals as governed by a concern for their own advantage. Commitment adds a recognition that individual choice is not always governed by a concern for one's own advantage – that choice is sometimes "counterpreferential."[5] When Sen speaks of "counterpreferential" choice in "Rational Fools" (1977: 328), it seems that he means non-self-interested choice: choice that is counterpreferential only in the expected-advantage sense of "preference." Altruistic motives would in this sense be counterpreferential. The fact that choices may be counterpreferential in this sense blocks the inference from choice to welfare and thereby complicates welfare economics.

In later works, Sen argues that commitment can involve counter-preferential choice in a stronger sense. In the 1980s, Sen refined his views of sympathy and commitment, distinguishing three theses whose conjunction leads to the identification of people's preferences with not just expected advantage, but with an exclusive concern with themselves. These are: (1) "self-centered welfare" (the person's welfare does not depend on benefits or harms to others); (2) "self-welfare goal" (a person's preferences among alternatives depend only on their upshot for the person's welfare); and (3) "self-goal choice" (a person's choices depend only on his or her own goals) (1985a: 213–14; 1987a: 80). Sen then writes:

> Sympathy does, specifically, violate self-centered welfare, but commitment does not have such a unique interpretation. It can, of course, reflect the denial of self-welfare goal, and indeed it is perhaps plausible to interpret in this way the example of a person acting to remove the misery of others from which he does not suffer himself. But commitment can also involve violation of self-goal choice, since the departure may possibly arise from self-imposed restrictions on the pursuit of one's own goals (in favor of, say, following particular rules of conduct). (1985a: 214)

When commitment conflicts only with self-welfare goal, then a person's choice is still determined by his or her goal (which I here identify with preference). It is just that the person's preferences are not always addressed to his or her own advantage. But when commitment conflicts with "self-goal choice," the agent's self-imposed "restrictions on the pursuit of [his] own goals" leads to a choice of an option that does not best fulfill the agent's "goals." If one can identify "goals"

[5] Non-self-interested motivations might lead to the same choices that expected advantage implies. Yet one would nevertheless have a case of commitment if the person would still have chosen the same action even if expected advantage lay with some other alternative (1977: 327).

and "preferences," this is a stronger conception of counterpreferential choice.

In what sense of "preference" can one identify preferences with what Sen here calls "goals"? It seems to be quite a wide sense. For example, Sen writes of goals as "including moral objectives" (1987a: 81). So the "self-imposed restrictions on the pursuit of one's own goals (in favor of, say, following particular rules of conduct)" are not moral objectives. But goals must not be all encompassing, either. If goals are all-things-considered preferences, how could counter-preferential choice ever be rational?

4. MANY CONCEPTS OF PREFERENCE OR JUST ONE?

Sen's underlying concern, I think, is to push economists toward a more nuanced view of rationality and rational choice (see particularly his 1985b and the introduction to his 2002). He seeks to avoid a view of rational decision-makers as rational fools who carry around some single ranking of all the objects of choice and, within the constraints of feasibility, simply choose from the top of the ranking. Economists should instead recognize that there are many different ways in which alternatives can be evaluated and hence many different conceptions of preference. Though these need not always be distinguished from one another, any acceptable theory of rational choice will in Sen's view have to make room for many notions of preference.

Those who, like me, believe that economists need only a single notion of preference as an all-things-considered ranking might disagree with Sen in two very different ways. One possibility is that they believe that people *are* rational fools. As a general view of people, this is a silly view, but it is not obviously silly to maintain that viewing people as rational fools is a reasonable approximation with respect to the phenomena of concern to economists. That, however, is not my view. I am no more in sympathy with modeling people as rational fools than Sen is.

A second possibility is that those who believe that preferences are all-things-considered rankings agree with Sen concerning human motivational complexity but disagree with him concerning the best strategy for dealing with this complexity. Rather than capturing this complexity by means of multiple concepts of preference, one might capture it by a nuanced account of the many factors that influence preferences. People prefer some things because of their expected benefits, others because of emotional reactions toward other people, others because of adherence to social norms or to moral principles, others out of mere habit. Depending on the context and the objectives, the account one might offer of the factors that influence preferences might be very simple or extremely complicated. Many economists would go on to argue for a division of labor, whereby

psychologists and philosophers study the deliberative processes that give rise to an all-things-considered ranking, and economists investigate the consequences of the choices that arise from those rankings coupled with constraints and individual expectations.

Though I think that it is better to locate the complexity in the account of what determines preferences, I reject a division of labor that denies that economists should be concerned with preference formation (Hausman forthcoming). To attribute to people a consistently articulated all-things-considered ranking of alternatives is a strong idealization, which in many contexts may be extreme and unreasonable. Without thinking about where this ranking comes from, economists will not understand when such an idealization is sensible and when it is not. Nor will they understand how changing beliefs and circumstances will influence both preferences and action. As Sen has shown – though in other terminology – the task of understanding how agents construct their all-things-considered preferences cannot reasonably be left out of economics. In my view, the same considerations that drive Sen to insist on the multiplicity of notions of preference justify instead identifying preferences with all-things-considered rankings and distinguishing sharply between preferences on the one hand and, on the other hand, expected advantage, the ranking implicit in choice, or any other substantive dimension of evaluation.

In addition to the expected advantage and choice construals of preferences, Sen notes that economists have taken preferences to refer to "mental satisfaction," "desires," and "values" (1997a: 303).[6] Because these are often taken to go hand-in-hand, economists have not felt it necessary to draw distinctions between these very different things, and in some contexts that may be a perfectly sensible policy. But, as Sen goes on to argue, "the eschewal of these distinctions in the characterization of persons amounts to seeing people as 'rational fools', as undiscriminating and gross thinkers, who choose one all-purpose preference order . . . A theory of human behaviour – even on economic matters – calls for much more structure and many more distinctions" (1997a: 304).

Sen is right to maintain that "A theory of human behaviour – even on economic matters – calls for much more structure and many more distinctions." But it doesn't follow that it needs multiple notions of *preference*. On the contrary, it seems to me that Sen's concern to draw distinctions can be accommodated at least as well by distinguishing sharply *between* preference (as all-things-considered ranking) and other things. Rather than taking expected advantage to be one concept of preference, one can distinguish *between* preference and expected advantage. Instead of taking

[6] And utility, which is frequently taken to be an index of preference, has an even wider range of meanings. See Sen (1987b: 5ff.) and Sen (1991b).

one concept of preference to refer to "mental satisfactions," one can distinguish between "mental satisfactions" and the extent to which preferences are satisfied.[7] A theory purporting to explain or predict the ways in which agents evaluate states of affairs will need to make distinctions between values and mere tastes, but these can be seen as distinctions among the factors that are responsible for a person's preferences, not as different conceptions of preference.

Sen's concern with the complexities of rational deliberation and choice can in this way be accommodated by those who take preference to be all-things-considered ranking. That concern consequently provides no strong reason for distinguishing many conceptions of preference. On the other hand, the fact that one *can* accommodate Sen's concerns with a single all-things-considered conception of preference is not by itself an argument for making do with this one conception of preference rather than recognizing many conceptions. Why then am I making a fuss? Is the issue just semantics? Does it matter whether one takes "preference" to be multiply ambiguous or whether instead one takes "preference" to be all-things-considered ranking and gives other names to what Sen takes to be other conceptions of preference?

There are four reasons why I think it matters. First, to regard choice rankings, expected advantage rankings, hypothetical choice rankings, "mental satisfaction," "values," "tastes," "all motivational considerations other than principle," and "all-things-considered rankings" as eight different conceptions of preference is an invitation to perpetuate the confusions that Sen has so justly criticized. To mark the distinctions between these different things, economists should use different words. The justification for retaining the word, "preference," for the last of these eight concepts is that it matches the "official" notion of preference in basic presentations of the theory of rational choice as well as economic practice, especially in game theory and expected utility theory.

Second, treating choice rankings, expected advantage rankings, and so forth as alternative conceptions of preference makes it more difficult to pose questions concerning what things determine preferences. The concept of all-things-considered rankings is the most suitable concept of preference, precisely because it does not settle *a priori* what those "things" are. That way, economists can separate the use of the word "preference" from substantive views about what preferences depend on.

[7] A preference is satisfied or not in the same sense that a requirement is satisfied or not, by things being as they are preferred or required to be. *If* an agent knows that some preference is satisfied (which need not be the case, even if the preference is in fact satisfied), then the agent *may* feel satisfied. But there is no other connection between preference and "mental satisfaction".

Third, several of the supposed conceptions of "preference" fly in the face of everyday understandings of the word.[8] I'm not saying that everyday usage is determinative, and indeed I believe that taking preferences to be all-things-considered rankings is not fully in accord with everyday usage. But conforming roughly to everyday usage helps avoid misunderstandings. Taking preferences to imply all-things-considered rankings modestly extends the everyday notion to serve the purposes of economists and decision theorists.

Finally, I shall argue in the next section that only the conception of preference as all-things-considered ranking permits game theory and expected utility theory to serve their predictive and explanatory roles.

5. GAME THEORY AND COUNTERPREFERENTIAL CHOICE

In certain contexts, Sen's "rational fools," whose behavior is unthinkingly governed by a single all-things-considered ranking, appear very foolish indeed. The foolishness is glaring in the case of finitely iterated prisoners' dilemmas or centipede games. In laboratory circumstances designed to implement these games, people frequently manage to cooperate and to do much better than the rational fools whose strategy choices are studied by game theorists. How should one make sense of this fact?

	C	D
C	($4, $4)	($0, $5)
D	($5, $0)	($1, $1)

FIGURE 1.

Following Sen's lead, I shall focus on the one-shot prisoners' dilemma, which has the advantage of being much simpler to present and to analyze than iterated prisoners' dilemmas or centipede games. Consider the interaction or "game form" shown in Figure 1, where the first number indicates the dollar payoff to the row player from each strategy combination and the second the dollar payoff to the column player.

[8] This claim requires qualification, because many people in fact believe that people's preferences are always dictated by their self-interest, and many hold the psychological hedonist view that whatever people do, they do because of the pleasure they expect. But, as Sen himself emphasizes, these views of preference are false, and they depend on well-known philosophical mistakes. Following ordinary usage, when that usage is confused and mistaken, is not a virtue.

https://doi.org/10.1017/S0266267104000379 Published online by Cambridge University Press

Assume that the game form is common knowledge. Each player earns a larger monetary payoff by playing strategy D "defect" than by playing strategy C ("cooperate"), regardless of what the other player does. Note that Figure 1 does *not* depict a game. By definition, one does not have a game until preferences are assigned to the outcomes.

If, in addition, each player confronting the situation shown in Figure 1 cares only about his or her own financial payoff, and this is common knowledge, then one has the prisoner's dilemma game shown in Figure 2.

	C	D
C	(3, 3)	(1, 4)
D	(4, 1)	(2, 2)

FIGURE 2.

Figure 2 is the normal form of a game of complete information.[9] The numbers are indices of preference with larger numbers indicating more preferred outcomes. That means that the alternative strategies, the outcomes, and the player's preferences (ordinal utilities) are common knowledge. D is a strictly dominant strategy for both players – that is, each does better playing D rather than C, regardless of what the other does. The reason why two players who both play C do better than two who play D is that each benefits from the *other* player choosing C. Since this is a one-shot game, in which there is no role for reputation or reciprocation, each player harms himself (in terms of his or her own preference ranking) by playing C.

All of this would be clear and uncontroversial if it were clear and uncontroversial what the utility numbers, which represent "preferences," meant. I maintain that they represent all-things-considered rankings. Somewhat tendentiously, I shall call this "the orthodox interpretation" of game theory. If utilities represent all-things-considered rankings, then anyone playing C when faced with the game form in Figure 1 is either irrational or is not playing a prisoner's dilemma.[10] By taking utilities to

[9] Since both players have strongly dominant strategies, they do not in fact need to know the payoffs to the other player in order to arrive at their strategy choices. But their interaction will not, strictly speaking, be a prisoners' dilemma if their information is not perfect.

[10] In this I am following Ken Binmore (1994), who argues this point at length. Unlike Binmore, who takes this view to support a revealed preference interpretation of preference, I take this point to be linked to a conception of preference as all-things-considered ranking and to be inconsistent with revealed preference theory. As argued in Hausman (2000) and very briefly above in section 2, "preferences" in game theory are not choice rankings.

represent all-things-considered rankings, game theorists are able to predict and explain strategy choices in terms of facts about games, including rankings of their outcomes. "Outcomes" must be understood in this context to be what Sen calls "comprehensive" as opposed to "culmination" outcomes (1997b: 745), since features of the play – that is, of the path through the game tree – may matter to the players in addition to the characteristics of the culmination.

Although this is one way to understand the utility payoffs in game theory, it is not the only way; and it has some disadvantages. In particular, it appears to limit game theory proper to an examination of how strategy choices depend on game forms, beliefs, and all-things-considered rankings of comprehensive outcomes. That means that game theory has nothing to say about how individuals construct their all-things-considered rankings of comprehensive outcomes. When faced with the fact that experimenters find high rates of cooperation in interactions that appear to have the structure of prisoner's dilemmas, all the game theorist can say is that the subjects are irrational or, more plausibly, that they are not playing a prisoner's dilemma game. If the subjects are rational, then, in terms of their own *all*-things-considered preferences, D cannot be a strongly dominant strategy. But the game theorist has nothing to say about why their preferences are like this. The task of figuring out how individuals think about their strategic interactions and how they decide how to rank comprehensive outcomes (which may depend on reasons they have for preferring particular moves or strategies as well as on preferences for the culmination) is ruled out of game theory. The task resides instead in a sort of limbo. It is not governed by any economic theory, but it is not studied by any other discipline either.

If one takes preference to be expected advantage or indeed anything short of all-things-considered rankings, then rational strategy choice need no longer be *determined* jointly by beliefs, knowledge of the game form, and preferences over outcomes. For example, suppose one interpreted the utilities in Figure 2 as reflecting expected advantage. Given this interpretation of preference, counterpreferential rational choice is clearly possible. The fact that D is a dominant strategy in terms of expected advantage does not imply that rational individuals will play D. Game theorists would be unable to deduce strategy choices from knowledge of games. This is a serious loss. On the other hand, if one takes utilities to be indices of expected advantage, one opens a space in which to discuss the thought processes that may lead individuals to make choices that are not governed by their expected advantage. So there may be a significant gain here, too.

Sen does not adhere to an expected advantage interpretation of preference in his discussion of game theory. Indeed he explicitly recognizes that altruists whose preferences conflict with expected advantage can find

themselves in a prisoner's dilemma, too.[11] Furthermore, as we saw at the end of section 3, Sen does not limit the possibility of counterpreferential rational choice to cases where individuals make choices that do not serve their expected advantage. He entertains the possibility of (rational) counterpreferential choice, even when preferences or goals include "moral objectives". Yet preferences cannot be all-things-considered rankings, because then rational choice would have to follow preference. What does he have in mind?

Sen writes:

> The language of game theory . . . makes it . . . very tempting to think that whatever a person may appear to be maximizing, on a simple interpretation, must be that person's goal . . . . There is a genuine ambiguity here when the instrumental value of certain *social* rules are accepted for the *general* pursuit of *individual* goals. If reciprocity is not taken to be intrinsically important, but instrumentally so, and that recognition is given expression in actual reciprocal behaviour, for achieving each person's own goals better, it is hard to argue that the person's "real goal" is to follow reciprocity rather than their respective actual goals. (1987a: 86)

I interpret these remarks as follows: Suppose individuals cooperate when facing the strategic situation in Figure 1, and suppose this cooperation is rational. If preferences are all-things-considered rankings, then the individuals are not playing a prisoner's dilemma. But if instead preferences are rankings that are influenced only by what an individual values *intrinsically*, then the individuals can be playing a prisoner's dilemma game. When the reciprocity the players show is instrumental to pursuit of what they value intrinsically, "it is hard to argue that the person's 'real goal' is to follow reciprocity rather than their respective actual goals."

Sen is right about the importance of modeling the intricate thought processes individuals go through when faced with strategic problems like the one shown in Figure 1, and his suggestion that those who decide to cooperate may still in some sense "prefer" the outcomes where they defect to the outcomes where they cooperate is plausible.

[11] Altruists whose choices are governed entirely by the payoffs for the other player, who face the game form in Figure 1 would not be playing a prisoner's dilemma, but suppose the payoffs were as follows:

	C	D
C	($4, $4)	($5, $0)
D	($0, $5)	($1, $1)

With these payoffs D would be a dominant strategy for such altruists.

https://doi.org/10.1017/S0266267104000379 Published online by Cambridge University Press

When one asks game theorists why so many individuals facing the situation shown in Figure 1 cooperate and thus turn out not to be playing a prisoner's dilemma, they should have something better to say than, "That's not our department. Go talk to the psychologists." But it does not follow that economists should reject the conception of preferences as all-things-considered rankings and redefine the notion of a game so that strategy choices are no longer deducible from normal forms such as the one shown in Figure 2. The costs of doing that are too high. Preserving the determinate relationship between games and strategy choices provides a decisive reason to take preferences to be all-things-considered ranking.

A better way to meet Sen's concerns is to argue that the study of games needs to include the study of how games are constituted as well as the study of strategy choices and equilibria. To argue that economists should seek an explicit theory of how games are constituted, which would include an account of how individuals who are interacting strategically construct their beliefs and preferences, does not require that one break the connection between dominant strategies and rational choices. The way to fight Sen's battle – and I see myself as his ally, not his opponent – is to argue for an enlargement of economists' concerns from games themselves to the process of modeling of strategic circumstances as games, rather than to argue for a reconceptualization of the concept of a game, which is what rejecting the concept of preferences as all-things-considered rankings would require.

6. CONCLUSIONS

This paper calls for a reformulation of Sen's invaluable critique of the ways in which economists conceive of preferences and of the impoverishment of the theory of rational choice that results. While endorsing his criticisms and supporting his call for a more nuanced view of the psychology of rational decision-making, I maintain that it is better to criticize economists for making false claims about what determines preferences, conceived of as all-things-considered rankings, than to criticize them for conflating different notions of preference. The more nuanced theory of rational choice that Sen rightly looks forward to should in my view make room for many evaluative concepts besides preferences (again conceived of as all-things-considered rankings) rather than making room for many concepts of preference. Theories about rational thought in complex strategic situations are, as Sen argues, badly needed. But they are, I have argued, better supplied by maintaining the view of preferences as all-things-considered rankings and supplementing game theory with theories of the processes by which actors transform strategic situations into games, than by adopting other notions

of preferences and weakening the links between facts about games and conclusions about which strategies rational players choose.*

REFERENCES

Anderson, E. 2001. Unstrapping the straitjacket of "preference": a comment on Amartya Sen's contributions to philosophy and economics. *Economics and Philosophy* 17:21–38
Binmore, K. 1994. *Playing fair*. MIT Press
Broome, J. 1991. Utility. *Economics and Philosophy* 7:1–12
Hausman, D. 2000. Revealed preference, belief, and game theory. *Economics and Philosophy* 16:99–116
Hausman, D. (forthcoming). Consequentialism and preference formation in economics and game theory. *Philosophy* (supplement)
Sen, A. 1970. *Collective choice and social welfare*. Holden-Day
Sen, A. 1973. Behaviour and the concept of preference. *Economica* 40:241–59; Rpt. and cited from Sen (1982: 54–73)
Sen, A. 1977. Rational fools: A critique of the behavioural foundations of economic theory. *Philosophy and Public Affairs* 6:317–44
Sen, A. 1980. Description as choice. *Oxford Economic Papers* 32:353–69; rpt. and cited from Sen 1982 (432–49)
Sen, A. 1982. *Choice, welfare, and measurement*. Blackwell
Sen, A. 1985a. Goals, commitment, and identity; rpt. and cited from Sen 2002 (206–24)
Sen, A. 1985b. Rationality and uncertainty. *Theory and Decision* 18:109–27
Sen, A. 1987a. *On ethics and economics*. Blackwell
Sen, A. 1987b. *The standard of living*. Cambridge University Press
Sen, A. 1991a. Opportunities and freedoms (from the Arrow Lectures) in Sen 2002 (583–622)
Sen, A. 1991b. Utility: ideas and terminology. *Economics and Philosophy* 7:277–84
Sen, A. 1993. Internal consistency of choice. *Econometrica* 61:495–521
Sen, A. 1997a. Individual preference as the basis of social choice; rpt. and cited from Sen 2002 (300–24)
Sen, A. 1997b. Maximization and the act of choice. *Econometrica* 65:745–79
Sen, A. 2002. *Rationality and freedom*. Harvard University Press

* I am grateful to Harry Brighouse, Michael McPherson, the participants in the Workshop at the University of St. Gallen, and especially to Geoffrey Brennan for their comments on earlier drafts.