CANADIAN JOURNAL OF PHILOSOPHY
Volume 20, Number 4, December 1990, pp. 469-486

Methodological Individualism and Marx: Some Remarks on Jon Elster, Game Theory, and Other Things

ROBERT PAUL WOLFF
University of Massachusetts/Amherst
Amherst, MA 01003
U.S.A.

In recent years, philosophers trained in the techniques and constrained by the style of what is known in the Anglo-American world as 'analytic philosophy' have in growing numbers undertaken to include within their methodological ambit the theories and insights of Karl Marx. Twelve years ago, Gerald Cohen startled the philosophical world with a tightly reasoned analytic reconstruction and defense of one of Marx's most influential and controversial teachings, historical materialism.[1] Seven years later, Cohen's friend and colleague, Jon Elster, produced what may fairly be considered the definitive analytic philosopher's encounter with the thought of Marx.[2]

I find Elster's book to be fundamentally a failure, despite its many virtues. It seems to me almost entirely to miss what is important in Marx's thought, frequently reducing it in the process to triviality, or, what is worse, to parody.[3] Now, stated thus baldly, my reaction might fairly be dismissed as a cheap shot, for Elster freely acknowledges the existence in Marx of depths and complexities which slip through his analytical filter. The problem, he thinks, is to find a way to translate Marx's rich, provocative, many-sided, but sometimes hopelessly metaphysically tainted theories, asides, *aperçus* insights, proposals, and

1 G.A. Cohen, *Karl Marx's Theory of History: A Defence* (Princeton: Princeton University Press 1978)

2 Jon Elster, *Making Sense of Marx* (Cambridge: Cambridge University Press 1985)

3 David Schweickart has written a splendid review of the book which exposes both the inadequacies of some of Elster's scholarship and also the deeper political significance of Elster's anti-Marxian 'Marxism.' See *Praxis International* 8 1988).

rhetorical flourishes into something that will withstand the critical scrutiny of a sympathetic modern analytic philosopher. Since I believe Elster has failed, it is incumbent upon me not only to gesture grandly in the direction in which Marx's superior wisdom seems to lie, but actually to state with some precision what Elster has missed, and how we might succeed in reclaiming it for our day. Beyond that, we must ask whether the models and forms of analysis which Elster takes from the modern theory of rational choice are fundamentally unsuited to the task of making sense of Marx.

Midway through the book, Elster writes the following words about his struggles with Marx's theory of ideology. They could as well have served as a general summary of the deeper philosophical purpose of the entire book:

> In my struggle with Marx's writings on ideologies, I have been constantly exasperated by their elusive, rhetorical character. In order to pin them down, I have insisted on the methodological individualism set out [in the Introduction], with results that may appear incongruous to some readers. Yet I fail to see any satisfactory alternative. A frictionless search for the "function" of ideologies or the "structural homologies" between thought and reality has brought this part of Marxism into deserved disrepute. To rescue it – and I strongly believe there is something here to be rescued – a dose of relentless positivism seems to be called for. (239)

In the opening pages of the book, Elster summarizes the doctrine of methodological individualism which he endorses. 'By this,' he says, 'I mean the doctrine that all social phenomena – their structure and their change – are in principle explicable in ways that only involve individuals – their properties, their goals, their beliefs and their actions. Methodological individualism thus conceived is a form of reductionism' (5).

Defenders of methodological individualism customarily ground their position on the ontological claim that only individuals are real, all else – corporations, institutions, states, societies – being in some way aggregates of individuals. Although Elster does invoke these considerations a page farther on,[4] they are not offered by him as the primary reason for his adoption of the individualist method. Rather, he says, the rationale stems from the fact that in scientific explanation, 'there is a need to reduce the time-span between explanans and explanandum – between cause and effect – as much as possible, in order to avoid spurious explanations (5).

---

4 Cf. where Elster writes: 'Methodological collectivism – as an end in itself – assumes that there are supra-individual entities that are prior to individuals in the explanatory order' (6). Although he does not say so explicitly, it is clear from the context that he rejects the appeal to such entities.

I share Elster's commitment to methodological individualism, but for the ontological reasons, which are, I think, more compelling and more constraining than those arising from considerations of the requirements of scientific explanation. I am prepared to assert not merely that an explanation in terms of individuals is better, simpler, or in other ways more desirable than an explanation in terms of such collective entities as states, classes, and institutions, but also that the inability to unpack such collectivist accounts into their individualist components of itself demonstrates that they can be no more than provisional sketches designed to guide us in promising directions.

Nevertheless, we need to ask a question that Elster seems never to think to ask, and having asked it, to stay for an answer. *Why* do serious, intelligent, clear- thinking social theorists like Marx – and like Emile Durkheim or Karl Mannheim – appeal to a language and style of explanation that seems, at least upon first examination, to violate the canons of individualist methodology to which one might otherwise imagine them to be committed? Let us grant, at least provisionally, that these men really had something authentic and important in view, that they were not misled by a faulty discourse or inadequate grasp of the tools of analysis, but rather were in the grip of an insight that they were unwilling to relinquish merely out of methodological piety. Rather than speaking dismissively and with a regrettable condescension of 'Marx's lack of intellectual discipline' (508), or of 'the omnipresent *bias of wishful thinking* in Marx's work' (438), it would be much more useful to take it as a working hypothesis that Marx really had his finger on something worth analysing, so that we might, by following along with him critically but generously, come upon understandings that otherwise might be denied us.

I begin with what I take to be the pivotal passage in Elster's book. A bit more than midway through the text, in a section entitled 'The conditions for collective action,' the following appears:

> The motivation to engage in collective action involves, centrally, the structure of the gains and losses associated with it for the individual.... The gains and losses associated with collective action must, for the present purposes, be measured in terms of expected utility. Hence they depend both on the individual's estimate of the likelihood of success and failure and on the degree of risk aversion.[5] For

5 By alluding to degrees of risk aversion, Elster implicitly invokes the assumption that utility is cardinally measureable, with no apparent awareness of the enormously powerful premises required for that assumption. There are even suggestions, as we shall see, of interpersonal utility comparisons. Here, as elsewhere, Elster uses what I should call the rhetoric of game theory with no attention to its logic.

> the time being I assume that the utility derives from the *material* gains and losses for the individual himself.... On these assumptions, then, the utility calculus of collective action is captured in three variables. The first is the *gain from cooperation*, defined as the difference between what accrues to the individual if all engage in the collective action and what accrues to him if none does. The second is the *free-rider gain*, that is the difference between what he gets if all but him engage in collective action and what he gets if everyone does so. Finally, there is the *loss from unilateralism* – the difference between what he gets if no one engages in collective action and what he gets (such as punishment or costs of engaging in useless individual action) if he is the only one or among the few to do so.
>
> Other things being equal, the probability of collective action increases with the first of these variables and decreases with the second and third. Frequently, however, they do not vary independently of one another.... In general, collective action will either be *individually unstable* (large free-rider gains), *individually inaccessible* (large losses from unilateralism) or both. Since nevertheless such action does occur, we must try to understand how these obstacles are overcome. (351-2)[6]

This passage perfectly captures the style and tone of Elster's analysis: superficially careful, precise, rigorous, apparently aware of the complexities of human motivation (in portions I have elided to save space, he recognizes the role of altruism, for example), a quantitative formalism lurking just below the surface. Clearly, Elster's language implies, if we insisted, he could put the whole thing into symbols, thereby removing the slightest vestige of subjective opinion from his analysis. And yet, the entire passage is utterly mad – a crackpot account that sounds as though it comes from Swift's account of the voyage of Lemuel Gulliver to Laputa, or from Anatol France's *Penguin Island*, or, worse still, from Robert Nozick's *Anarchy, State, and Utopia*.

Think for a moment about what Elster is saying. Collective action, according to him, is individually unstable, individually inaccessible. As he says several pages later, in the midst of a discussion of the rationality of collective action, 'for collective action to take place so many conditions must be fulfilled that it is a wonder it can occur at all' (361). But the most casual survey of history and society shows us that collective action is the norm in human affairs. In every human group one can think of, collective action dominates the waking hours – and even the sleep – of every one over the age of one and a half or two.

---

6 The notion of a 'difference' between two gains presupposes cardinal utility. Elster's formulation also makes sense only so long as there are no more than two strategies for each of two 'players.' Since, in general, there will be many strategies available to each of many players, the notions of gain from cooperation, free-rider gain, and loss from unilateralism are incompletely defined. As we shall see, Elster is mesmerized by the elementary pictures of the Prisoner's Dilemma, and forgets to ask whether these little sums and differences correspond to anything in the real world.

A little reflection will remind us that all of the productive activities of human beings are collective in character, even those of the fabled Robinson Crusoe.[7] All kinship interactions, sexual liaisons, all our activities of eating and warring, almost all religious activities and activities of artistic creation, reproduction, and appreciation, are collective in character. Voting, strikes, military campaigns, riots, cocktail parties, family vacations – all of these, on Elster's view, are so improbable that we can barely understand how they might, on rare occasions, actually happen. Clearly, there is something badly wrong with a theory of society that concludes that the norm is so abnormal that it is almost never likely to occur! Where does Elster go wrong?

Elster's first problem is that he never actually defines the phrase 'collective action,' despite the fact that the book is pretentiously quasi- formal, full of definitions, Game Theory jargon, allusions to payoff matrices and the like. Clearly, until we know with some precision what he means by the term, we cannot even begin to evaluate his claim that collective action is unlikely, albeit actual, nor can we determine in what sense, if any, Marx's explanations of collective explanation, or anyone else's, have violated the principles of methodological individualism.

The core of the argument, such as it is, can be found in Chapter 6, 'Classes.' Elster begins by defining 'class.' After reviewing in a useful and interesting fashion some of the disputes that have grown up around the term, he offers the following definition: *A class is a group of people who by virtue of what they possess are compelled to engage in the same activities if they want to make the best use of their endowments.*[8] To say that classes are *real*, he explains a bit later, is to say that 'under certain conditions they tend to crystallize into collective actors,' and this latter phrase – crystallizing into collective actors – is explained as meaning that they 'achieve class consciousness' (344).

Thus far, not much light has been shed. In particular, we want to know what Elster understands by class consciousness. His rather bizarre answer is this: 'I define (positive) class consciousness as *the ability to overcome the free-rider problem in realizing class interests*' (347).

This is, to put it mildly, not what Marx and other social theorists have seemed to have in mind when they used the term 'class consciousness,'

---

7 Marx labelled the efforts by Vulgar Economists to read economic laws out of the imagined experiences of isolated producers 'Robinsonades' (*Robinsonaden*). See *Marx-Engels Werke* (Berlin: Dietz Verlag 1962 B. 23, 90. see also S.S. Prawer, *Karl Marx and World Literature* (Oxford: The University Press 1976), 273 ff.

8 Elster, 331. Needless to say, this sounds more like a neo-classical than a Marxian definition, but so be it.

but let us deal with Elster in his own terms, for the moment, and see whether we can figure out what he is saying. There are two difficulties with the definition he offers of class consciousness: the first is that he does not tell us what *the* free-rider problem is, and the second is that he does not explain what one would have thought would be for him the extremely problematic notion of class interests.

Consider first the so-called free-rider problem. There are actually two free-rider problems, not one. The first is a problem for those who want to get a group of people to act together in pursuit of some social, political, cultural, economic, religious, or other goal. The second is a theoretical problem of explanation for rational choice theorists. Elster, like most rational choice theorists, confuses the two.

The practical free-rider problem is that sometimes, when we are trying to get a group of people to pursue a goal, it is hard to get everyone to pitch in and do his or her part, because some individuals may figure that it can't make any noticeable difference if they slack off. Especially when the action involves considerable effort, or cost, or danger, or when the connection between the action and the end isn't very clear, this sort of thinking may pose a serious threat to the success of the effort. My admittedly limited experience suggests that relatively rarely can the problem be traced to deliberately selfish calculations in which individuals literally figure out that their dominant strategy is nonparticipation. Notice that this is a practical problem which only rises to the level of concern when large numbers of people are slackers. A strike, an election, a riot, even a family picnic, can survive *some* level of free-riding, and experience tends to teach us when that level is likely to be exceeded, and when it is not. To cite one actual example from my own recent experience: when an organization I run conducted a telephone campaign to get out a vote and raise money, we were told by the marketing firm doing the calling that a 50% rate of pledge fulfillment was a reasonable expectation. Now, in fact, we experienced only a 40% fulfillment rate, which created some financial problems for us. But a 50% rate, which rational choice theorists do not even deign to discuss, would in real world terms have been quite satisfactory.

The theoretical free-rider problem is this: if we make some very powerful, very restrictive assumptions about the utility functions of a group of individuals and about the canons of choice to which they conform their decisions – assumptions which amount, roughly, to the premises that individuals make choices solely on the basis of expected benefits to themselves, very narrowly construed, that they impute identical preference structures to all other individuals, that there is insufficient information or communication or enforcement procedures to affect individual choices, and that individuals choose so as to maximize expected benefits – then we can deduce that there are certain sorts of

actions, requiring the active participation of large numbers of people, which will not occur. Strikes will be called, but no one will show up on the picket line. An election will be held, but no one will vote. The command will be given to charge the enemy emplacement at the top of the hill, but no one will move. A leader will cry, 'to the barricades,' but no one will budge. Now, the fact is that strikes, elections, infantry charges, and street rebellions *do* occur. So the theory of rational choice has a problem. Clearly, some of the premises of Rational Choice Theory are wrong. (Note: this is *not* a problem for the strike leaders, party bosses, Second Lieutenants, or revolutionaries. The group efforts they are trying to promote are, ex hypothesi, occurring. The problem is for the theorists, who must confront the fact that their theories entail conclusions which are disconfirmed by the facts.)

A little later, I shall address directly the question of just which premises of rational choice theory ought to be called into question. But first, we must try to decode the second phrase in Elster's definition: class interests. What are class interests, according to Elster, and what does it mean to 'realize' class interests.

An interest is a goal or end or aim or purpose that a purposive agent sets for itself (or, alternatively, that may, on some theory, be imputed to the agent despite the agent's unawareness of it. Elster understands how important this addendum is, and has some intelligent things to say about it. Since my disagreements with him do not turn on this aspect of the subject, I shall ignore it here.) Any methodological individualist – such as myself – will presumably need to define, or explain, class interests in terms of the interests of individual persons. It is thus extremely puzzling that Elster does not directly address himself to this task, apparently considering it self-evident what a *class* interest might be.

Our best evidence of what Elster has in mind – and the real indication, I think, of the real use he wishes to make of rational choice theory – appears in the section of Chapter 6 entitled 'The rationality of collective action.' Here, at some length, is the passage:

> On first principles, one should seek for micro-foundations for collective action. To explain the collective action simply in terms of the benefits for the group is to beg all sorts of questions, and in particular the question why collective action fails to take place even when it would greatly benefit the agents. The individual-level explanations should be constructed according to the following heuristic principle: first assume that behavior is both rational and self- interested; if this does not work, assume at least rationality; only if this is unsuccessful too should one assume that individual participation in collective action is irrational. ...
>
> The basic problem confronting any group of people trying to organize themselves is that of the Prisoner's Dilemma. In its simplest form it is a strategic game between any given individual and "everyone else." To each of these actors, two

strategies are available: to engage in the collective action or to abstain. For any pair of strategies chosen by the actors, there is a well- defined payoff (in expected material welfare) to each of them. In the matrix below the first number in each cell represents "my" payoff and the second the pay-off to each of the individuals included in "everyone else."

Table 6.1

		Everyone else: Engage	Everyone else: Abstain
I	Engage	b,b	e,f
	Abstain	c,d	a,a

Here b-a represents the gain from cooperation.... Similarly c-b represents the free-rider gain and a-e the loss from unilateralism. Clearly, whatever everyone else does, it is in my interest to abstain. If all others engage in collective action, I can get the free-rider benefit by abstaining and if everyone else abstains I can avoid the loss from unilateralism by abstaining too. Since the reasoning applies to each agent, in the place of "I," all will decide to abstain and no collective action will be forthcoming.

In one sense the logic is compelling. If (i) the game is played only once, (ii) the actors are motivated solely by the payoff in the matrix and (iii) they behave rationally, collective action *must* fail. By contraposition, we might look into the possibilities for collective action if the interaction is repeated several times; if the payoffs that motivate the actors differ from the material reward structure; and if the behavior is less than fully rational. It turns out that under all these conditions, collective action does become possible. The three cases correspond to what is referred to earlier as rationality-cum-selfishness; rationality simpliciter; and irrationality. (359-60)

In light of these remarks, it would appear that by 'realizing class interests' Elster means moving from the sub-optimal equilibrium of the Prisoner's Dilemma Game to the Pareto-preferred outcome of mutual trust and cooperation. Implicitly (but *only* implicitly), collective action is then action which achieves (or aims at? who is doing the aiming?) the Pareto-preferred outcome.

This simply won't do. Indeed, it won't do for so many different reasons that it is a bit hard to know where to begin the critique. For purposes of organization, if no other, let me start with the most interior criticisms – those which accept Elster's framework of analysis – and then proceed to call that framework itself into question.

Let us begin, where Elster does, with the much-discussed, much-misunderstood Prisoner's Dilemma. From a Game Theoretic point of

view, the little payoff matrix which he introduces into the text is a complete mess. Here are some of the problems:

1. The matrix purports to represent payoffs in expected material welfare resulting from the four possible pairs of strategy choices. Elster forgets to tell us how the players rank these outcomes, so the matrix, as it stands, doesn't define a Prisoner's Dilemma game. Furthermore, since interpersonal comparisons of utility are, I presume, not being posited, the use of the same letters (a and b) for payoffs to both players is extremely misleading. To make the matrix represent a Prisoner's Dilemma, we must assume that the players have the following preference structure (I pass over the not insignificant fact that Elster fails to distinguish between the rank ordering of quantities of material welfare and the rank ordering of preferences): for the player identified as 'I,' c›b›a›e, and for the player identified as 'everyone else,' f›b›a›d.

2. A Prisoner's Dilemma is a two-person game with no communication. It is assumed that this is a situation of choice under uncertainty, which means that the outcomes are well-defined, the strategy options are well-defined, and the players' preferences are well-defined, but the players have no way of estimating the probability that other players will select particular strategies. In the present case, this means that neither player can make a reasoned estimate of the probability that the other players will choose to coordinate on a policy of mutual engagement in collective action. But although this may very well model a laboratory situation in which subjects recruited from a university campus are run through little artificial games, it completely fails to model the actual situation of a platoon, a union local, a family, or an electorate. Note: this is *not* to say that such real-life groups act 'irrationally.' Quite to the contrary – it would be wildly irrational for a group of voters, workers, or soldiers to ignore what they know about one another, what they remember of their past interactions, and what they have communicated to one another. The Prisoner's Dilemma, mesmerizing though it may be, simply is not a model of group action.

3. A Prisoner's Dilemma is a game defined by a two-by-two matrix, which means that it is a game in which each player has only two strategies. It is actually very difficult for those unfamiliar with Game Theory to grasp just how reductively, absurdly, uninterestingly simple a game must be in order to offer only two strategies to each player. By way of example, consider the following silly little game, which I invented to make my point. There are two players, A and B, who start with a pile of four matchsticks. A move consists of removing either one or two matchsticks from the pile, and players move alternately,

who goes first being decided by a coin toss. The player who removes the last remaining matchsticks loses.

The longest the game can last is four moves – two for A and two for B. The shortest is two moves – one apiece. Not a very interesting game, certainly not as interesting as the game of getting eighty million people to vote, or the game of getting three thousand workers to strike, or even the game of getting eleven soldiers to charge a hill defended by a machine gun nest. And yet in this little game, A has twelve strategies, B has twelve strategies, and the payoff matrix is a twelve by twelve matrix with one hundred forty-four boxes (leaving to one side Nature's choice of heads or tails, which would require a third dimension to represent). The degree of simplification and abstraction needed to construe a situation as capable of being modelled by a two-by-two matrix is such that there is almost certain to be no interesting connection between the model and any social, political, or economic reality.[9]

4. Elster puts on a fine show of formalist rigor with his equating of the quantity (b-a) to the gain from cooperation, and so forth. These quasi-quantitative formulae have any meaning at all *only* if we are talking about a two-person game in which each player has only two strategies. If, as must certainly be the case in the real world, there are many players, each with many strategies, then the meaning of 'gain from cooperation' or 'loss from unilateralism' loses all precise meaning, and becomes a metaphor without a referrent.

To see that this is so, consider a very slightly more realistic game, called Strike. The game has three players: A, who is the strike leader, and B and C, who are the followers. The game has four moves: A goes first, and either calls a strike, or doesn't. B who goes second, must go along with A if A doesn't call a strike, but may choose either to join or not join a strike if A calls one. C goes third, and has the same options as B, but with the difference that C knows what B has done. Finally, A goes again, and can either affirm or cancel a strike in light of what B and C have done, assuming that A has called a strike on the first move.

---

9 The game is much simpler if A automatically goes first. Then, A has 3 strategies, B has 4, and the matrix is 3 x 4. For those who are curious, here are the strategies available to the two players in this simpler game. A's strategies are (1) Take 2. If B takes 1, take 1. (2) Take 1. If B takes 1, take 1. If B takes 2, take 1. (3) Take 1. If B takes 1, take 2. If B takes 2, take 1. B's strategies are (1) If A takes 2, take 2. If A takes 1, take 1. If A then takes 1, take 1. (2) If A takes 2, take 2. If A takes 1, take 2. (3) If A takes 2, take 1. If A takes 1, take 1. If A then takes 1, take 1. (4) If A takes 2, take 1. If A takes 1, take 2.

In this little game, A turns out to have 17 strategies, B has 2, and C has 4. A payoff matrix would therefore have to be a 3-dimensional array, 17 x 2 x 4, with 136 boxes, in each of which would be entered a triad of numbers or letters representing A's, B's, and C's evaluations of the particular one of the nine possible outcomes arrived at by the intersection of the three strategies corresponding to the row, column, and depth (or whatever) intersecting at that box. There is nothing in this game quite so simple as the gain from cooperation, the free-rider gain, or the loss from unilateralism, even assuming one could define the appropriate cardinal measures of utility.

5. Elster completely confuses his own formalism by describing the second of each pair of payoff entries as representing the payoff 'to each of the individuals included in "everyone else."' But this simply won't do! If all the others are independent players, then this is an n-person game, not a two-person game. And if 'everyone else' really is a group acting *as* a group, with two strategies, then all those people have, ex hypothesi, *solved* the problem of acting collectively, in which case what the odd person out does is of relatively little interest!

To see how confused Elster's account actually is, consider this passage, appearing immediately after the definition of class consciousness as the ability to overcome the free-rider problem. The problem, he says, is that 'the individual can reap a greater reward if he abstains from the action to get the benefits without the cost. This generates a conflict between the interest of the individual class member and that of the class as a whole. (347)[10] But what can the phrase 'that of the class as a whole' possibly mean for Elster? If the logic of free-riding leads everyone to defect, then the problem is not a conflict between the interest of the individual and the interest of the class as a whole. On Elster's own view, the problem is a sub-optimal outcome for each individual. The conflict is between what the individual wants and what she gets. The notion of class interest does not enter.

It is precisely here that Elster's failure to explain the notion of class interest leads him into confusion. It is clear that no aggregative function, such as average utility or total utility or a weighted average of utilities, is going to do the trick, even if we allow interpersonal comparisons of cardinal utility. In fact, this problem helps us to see why

---

10 Elster then gives, as a supposed example of this, a passage from the *Communist Manifesto*, but in fact the passage quoted does not describe an example of the free rider problem at all. It alludes to the fact that competition for jobs in the labor market interferes with the 'organization of the proletarians into a class,' an entirely different matter.

the Prisoner's Dilemma has such an appeal for him. In that little game, there are only two players and no pre-play communication – hence, there are no aggregation problems and no possibilities of side-payments, etc. In the n-person case, however, very sticky theoretical puzzles arise which powerfully resist plausible analysis by the models of Game Theory.

We could continue to give instances of Elster's misuse, or lack of understanding, of the terminology and formalism of Game Theory[11] but it is more useful to try to locate the source of the inadequacy of his methodology. The real problem, I suggest, is an incorrect notion of the self which is engaged in action, collective or otherwise, and a consequent inability to understand how individuals conceive of their situation, or formulate the goals of their action.

Elster is quite right that we should, on principle, seek micro-foundations, even though, as he points out, one must avoid 'premature reductionism' because 'collective action may simply be too complex for individual-level explanations to be feasible at the current stage' (359). He is also correct, in my judgment, in asserting that we ought to begin by assuming that behavior is rational, although not in the sense of being calculatively maximizing behavior. Rather, we should assume that behavior is rational in the sense of being purposive, goal-oriented, guided by considerations of instrumentality – that the individuals whose behavior we wish to understand could, at least in principle, give a coherent account of why they are acting as they are, by reference to what they seek to achieve and how they expect what they are doing to advance their goals.

The problem starts with Elster's inclusion of the assumption that behavior is 'self- interested.' To see what is wrong with this assumption, let us return to the Prisoner's Dilemma. Formally speaking, a Prisoner's

---

11 See, for example, the misuse of the term 'constant-sum game' on 373, and the completely garbled term 'variable-sum game,' which has no meaning in the formal development of Game Theory. A two-person strictly competitive game is correctly describable as 'zero-sum' or 'constant-sum' under certain extremely powerful assumptions about the preference structures of the players – who, incidentally, can be classes of individuals only if one can give meaning to the notion of the preference structure of a class! Games which are not constant-sum can only be described as *not constant-sum*. The concept of a sum of payoffs is undefined for such games (because such a sum would involve interpersonal comparisons of utility). Hence, they cannot be said to be variable sum. Since it is precisely the notion of *class* interests rather than *individual* interests which Elster is trying to elucidate, it is especially misleading to throw these terms around with no awareness that their use begs precisely the questions about the nature of collective interests which are at issue. This is the way in which the unrigorous use of formalism conceals rather than dispels confusion.

Dilemma is a two-person game, with two strategies for each player, in which the players prefer the outcomes in the pattern indicated above. *Any* two person game, with two strategies per player, with preferences for payoffs conforming to this pattern is a Prisoner's Dilemma. But there is a little story that gives this formal structure its name, and that little story contains more information than ends up being encoded in the payoff matrix. It is that extra information that is the source of the difficulty. The story, as everyone knows, concerns a pair of criminals who are nabbed in a robbery, held separately in jail cells, and presented by the District Attorney with a set of threats and promises concerning the jail sentences they will receive if each of them does or does not turn state's evidence.

The outcomes resulting from the criminals' various strategy choices are expressed in the number of years they may receive as sentences. The unspoken assumption which underlies the game is that each criminal ranks outcomes solely according to the length of *his* sentence (or hers, but the little story is always told about two men), preferring a shorter to a longer sentence. It is thus assumed that neither criminal is willing to serve as little as another minute in jail even in order to keep his buddy from going to the gas chamber. It is these assumptions that allow us to translate the story into a payoff matrix.[12]

But rationality does not require that individuals rank outcomes in this way. Indeed, even *self*-interest, broadly enough construed, does not imply the utility functions assumed to be operative in the Prisoner's dilemma. To see how things might be different, consider another game, which consists of two violinists playing the Bach Double Concerto. Let us suppose that each violinist can choose between playing as fast as possible, or playing *a tempo*. There are four possible outcomes: if both play as fast as they can, the result is a musical fiasco, but a personal standoff. Neither is humiliated by having been shown up as incapable of presto playing. If both play *a tempo*, the result is beautiful music. If one plays fast and the other plays *a tempo*, the result is personal triumph for the first and humiliation for the second, and, of course, musical disaster.

---

12 Formally speaking, all of this amounts to stipulating that each player's utility function is inversely monotonically related to sentence length, or, alternatively, that each player's utility function is a lexicographic function which first minimizes that player's sentence, and only then responds to other variables. Without assumptions like these, we can construct an outcome matrix which specifies what each player gets for each pair of strategy choices, but we have no way of translating that outcome matrix into a payoff matrix. The point of the phrase 'as little as another minute' is that with only ordinal rankings (and no interpersonal comparisons of utility), one cannot say anything about how much one is giving up in relation to how much one is inflicting on the other player.

Now imagine two pairs of players playing this game. The first consists of David and Igor Oistrakh, who in fact produced a transcendently beautiful recording of the composition, and the second consists of Beverly Sumac and myself at Mrs. Zacharias's annual parlor recital for her violin pupils in the late spring of 1946.

Beverly and I, I will simply report, treated the event as a Prisoner's Dilemma (in more senses than one!), with the result that we raced as fast as we could to the end of the piece. Since she was on her way to a career as a concert performer, and I, to put it mildly, was not, she won the race. The result was not music. The Oistrakhs, on the other hand, had a different preference structure. They were engaged in a collective activity – the making of beautiful music. For them, the joint playing of the concerto *a tempo* was the most preferred outcome, a madcap presto performance, we may imagine, came next, and the two other outcomes were treated as indifferently worst. The result was a coordination game in which the purely game theoretic aspects of the coordination were trivially easy, inasmuch as each could only lose by defecting from the strategy of playing *a tempo* (I leave to one side the non-game theoretic aspects of the effort to play the Bach Double, which were also easy for them, but would be distressingly difficult for me).

Note that there is no question here of altruism, or self-denial. Indeed, although Igor was David's son, they could just as well have been mortal enemies, so far as the problem of coordination was concerned. Each is engaged in goal-oriented, purposive, self-interested action – i.e., each is acting rationally. But since the goal that each pursues is the mutual and collective making of beautiful music, they coordinate rather than frustrate one another.

The same point can be made with regard to the free-rider problem. It must puzzle Elster how a large symphony orchestra can ever play the Brahms Second. After all, we can imagine him reasoning, no one but a Toscanini – and certainly not Seije Ozawa – will notice if a single second violinist puts soap on his bow and only pretends to play. But then, by parity of reasoning, the entire second violin section will soap up, and the orchestra will fall flat. No doubt some orchestral musicians, long in the tooth and cynical besides, might reason this way. But most orchestra players have, as their first preference, to get paid for a first-class performance *in which they participate*. Professional violinists do not begin to exhibit negative marginal utility for playing the violin until long past the limits of a Brahms symphony (I leave to one side the question of a Mahler symphony).

It is entirely possible – indeed, it is, I suggest, usually actual – that men and women will have, in this sense of the term, collective goals in the pursuit of which they engage in collective action. Nothing in

the having of such goals requires us to posit any entities save individual persons, nor is the pursuit of such goals in any sense irrational. Indeed, the pursuit of such goals is not even altruistic or non-self-interested. David Oistrakh does not play *a tempo* out of a selfless putting of his son's interests ahead of his own. He plays *a tempo* because he aims at a goal – the collective creation of beautiful music – which cannot be reached in any other way.

But, Elster will reply, romantic sentimentality to one side, a worker does not engage in strikes because he has adopted it as his goal to raise everyone's wages in a just and equal manner through joint action. If he did, there would *be* no free-rider gain. He engages in strikes in order to raise *his* wages. The raising of everyone else's wages may be a necessary means to his end – one which he will therefore support. But on the assumption that there are costs associated with committing oneself to a strike, he will ever be on the lookout for a way of obtaining the benefits – higher wages for himself – without incurring the costs.

Now, the plain fact of the matter is that this objection is, as a universal generalization about the behavior of workers – or other groups of people – just plain false. Most human behavior, it seems to me – the mean-spirited, ugly, cruel, unjust behavior included – is motivated by what may be called the pursuit of social, or collective ends. The reason for this is that human personality is formed by the internalization of social norms and roles, and by the identification of self as a member of familial, religious, geographic, political, military, social, or cultural groups, so that most people, most of the time, understand themselves and their situations in terms of the groups in which they are most securely imbedded. The conception of self on which rational choice theory bases its assumptions about individual motivation and choice are not only historically and culturally quite specific. Even in those cultures, and at those times, when individuals learn – culturally – to exhibit what Elster calls rational behavior, they exhibit it in very severely constrained ways and with regard to very narrowly limited ranges of options.

The conscious regulation of conduct by calculations of self-interest is so rare in human history that the greatest sociologists of the classical period – Marx, Weber, Sombart, and the rest – devoted endless efforts to explaining its appearance at a particular historical moment in the evolution of western European society. So unusual is such conduct even in capitalist societies that when we encounter an individual who allows rational calculation to regulate more than a narrowly circumscribed sphere of economic decisions, we are likely to consider him or her seriously pathologically deranged. Paul Goodman captured the crackpot quality of calculative instrumental rationality run amok in the

character of the mercantile capitalist Eliphaz in his wonderfully satirical burlesque of capitalism and modern educational theories, *Empire City*. What makes Eliphaz so wacky and non-human is precisely that he resists the natural tendency to engage in collective action.

Oddly enough, Elster knows that something is wrong with his attempts to explain collective action by appeals to iterated Prisoner's Dilemma games and such like ephemera. After actually canvassing the possibility that workers' utility functions are influenced by 'externalities' (economists' jargon for whatever doesn't fit the Procrustean bed of economic reasoning), he writes:

> Workers no less than capitalists might engage in collective action because they find it selfishly rational. I find it hard to reconcile this idea with the extensive literature on working-class culture, but on the other hand the elusiveness and subtlety of these problems of individual motivation should make us wary of dismissing it out of hand.(363)

It is hard to know what to say to an author who finds the most commonplace sorts of human motivation 'elusive and subtle,' and yet thinks that something as esoteric as egoistic maximization of expected utility is so transparent that it can simply be taken, unexplained, as a datum of explanation.

Let me offer an analysis of class consciousness and collective action as an alternative to Elster's. This analysis will necessarily be brief, but it is taken from a somewhat longer essay published more than twenty years ago, and those who are interested can consult the fuller version.[13]

If we adopt Ralph Barton Perry's useful definition of a value as *any object of any interest*, then we can say that Rational Choice Theory is designed to analyse the choices of individuals who pursue *egoistic values*, or, equivalently, who aim at objects, events, or states of affairs in which they take an *egoistic interest*. An egoistic interest is an interest which relates solely to the subjective state of the individual him- or herself. The goods and services flowing from economic activity, for example, are assumed by economists to be enjoyable by the solitary consumer, and to be valued for that reason. Needless to say, the goods cannot be produced by the individual independently, but what the consumer values is the consumption of the goods and services, to which end the economic system is merely a highly efficient means.

Egoism as a theory of the nature of value is not to be confused with the assumption that individuals act selfishly. One can hold that all value

---

13 See Chapter Five, 'Community,' in Robert P. Wolff, *The Poverty of Liberalism* (Boston: Beacon Press 1968).

is egoistic and yet prescribe, or claim that men and women do in fact practice, altruism. The simple altruist believes that all value is egoistic. He or she simply seeks to maximize someone else's value.

There is, however, another class of values, or of valued states and experiences, which we may call *social* values. These are states of affairs whose realization depends essentially (not merely instrumentally) upon a reciprocal relation between another's experience and my own. The example most familiar from the literature of political theory is the master/slave relationship described by Hegel. Those who wish to be masters – as opposed to those who merely desire the private satisfactions that were once provided by servants, and now more and more are provided by labor-saving mechanical devices – those, that is to say, who want the experience of mastery, require sentient, purposeful, servants who are subservient, and who conceive of themselves as subservient. It is impossible to describe mastery adequately without making reference to the servant's awareness of his subservience. Thus, if one sought to be a master, and by a peculiar accident succeeded only in bending to one's will a flagellant who, for his own religious ends, was using the relationship as a means to achieving a saving humiliation, one would entirely fail to achieve one's goal.

Somewhat more to the point, those who pursue democracy for its own sake seek to bring into existence a state of affairs in which free and equal men and women engage in rational discourse for the purpose of choosing, and then realizing, jointly arrived at ends. For them, the process of free deliberation is itself valuable, over and above the ends which may thereby be attained.

I suggest that collective consciousness is that state of affairs in which all or most of the members of a group take an interest in, or aim at, the same social value, and know that the others are doing so. Collective action is then the cooperative action of a group of people in pursuit of the actualization of some social value. Class consciousness, in particular, is the pursuit, by all or most of the members of an economic class (however defined) of the state of affairs in which the members of the group achieve economic well-being and political power, *and are mutually aware of having done so through their cooperative and collective efforts*. That mutual awareness is a part of what is aimed at, and hence the value that each seeks to actualize is inseparable from it.

Thus understood, collective action is neither mysterious nor methodologically suspect. What needs explaining – what Marx undertook to explain in the context of mid-nineteenth century European politics and economy – is how, why, and under what constraints a group of individuals come to take an interest in particular *social* values.

The Prisoner's Dilemma and the Free-Rider Problem are inappropriate analytic tools for understanding class consciousness because both

of them assume that in their preference structures, individuals pursue only egositic values. That assumption, which underlies all classical and neo-classical economic theory, is in fact so restrictive and incompatible with the common place reality of human experience that it provides no firm basis at all for an explanation of what we commonly understand as collective action.

Received April, 1990