CANADIAN JOURNAL OF PHILOSOPHY
Volume XIII, Number 1, March 1983

CRITICAL NOTICE

Jon Elster, *Ulysses and the Sirens: Studies in Rationality and Irrationality*. (Cambridge: Cambridge University Press 1979). Pp. ix + 193.

Philosophers will neglect this book at their peril. And not philosophers alone – all who are concerned with rationality and rational behaviour, and all who examine explanation in the biological and social sciences, will need to ponder the issues and arguments that it raises. The task is not easy. Apart from the difficulty of many of the issues, Elster's manner of presentation is not always as clear or straightforward as might be desired. Not all parts of this book illuminate. But those that do, illuminate greatly.

The four chapters are largely self-contained essays. All concern rationality, and are strung together on what we might dub a Plotinian thread – from the light of perfect rationality we move through imperfect and problematic rationality to the deep darkness of the irrational. But the unity that this metaphor suggests may mislead; each chapter approaches its subject from a quite different perspective and with very different concerns in view.

The first chapter develops two principal themes. The first is a conception of human beings as globally maximizing machines. The second is a contrast between functional and intentional explanation, with an argument to show that each has its appropriate but limited domain – for functional explanation, the biological sciences, for intentional explanation, the social sciences.

Let us trace some parts of Elster's account. The process of natural selection yields local maxima. In organic reproduction random genetic mutations occur. If a mutation enhances reproductive capacity it will spread through the population until it comes to be universally present; if it diminishes this capacity it will die out. Thus organisms of a given type move up the gradient of reproductive capacity until a position is reached from which, given the environment, no future mutations yield increased capacity. This position is a local maximum.

That the process of natural selection yields local maxima is not a truism. The relation between genetic and environmental changes might exclude the effective adaptation of organisms to their environment. Our existence is testimony to the fact that natural selection has not proved wholly ineffective. But that we exist is not itself a truism.

Although natural selection yields local maxima, it does not in general yield global maxima. To reach his chosen peak, a mountain climber must ascend. But not from each place. Not all upward paths from his starting point need bring him to the top. He must avoid gradients that lead only to local, minor peaks, below the main peak which is his objective. And at times he may find it necessary to descend, to cross valleys that lie between him and the peak. Or, to change the metaphor, il faut reculer pour mieux sauter.

Evolution, Elster argues, is incapable of these strategies necessary to the mountain climber. It can not wait; it can not refuse a favourable mutation so that it may later be able to accept a more favourable one. And it can not accept an unfavourable mutation so that it may later accept an even more favourable one; it can not employ indirect strategies. But in producing human beings, evolution has outdone itself. For what characterizes the perfect rationality of which we are capable is just the ability to use these strategies – waiting and indirection – in order to achieve global maxima. We are global maximizing machines.

A local maximizing machine treats its environment as parametric, a constant in which its behaviour alone is considered variable. A global maximizing machine must be capable of regarding its environment strategically, of taking account of the presence of other maximizers whose behaviour is also variable, and of adjusting its own choice of strategy to its understanding of how that choice will affect the choices of others who are themselves engaged in a similar process of adjustment.

> The transparency and symmetry of this interaction ensures that the fate of the actors is in their own hands, whereas a community of parametrically rational actors will be in the grip of causal forces that elude them and that perpetually make their plans come to naught. (19)

Only actors capable of strategic rationality are able to engage in mutual-

ly dependent cooperative activity. Even if cooperation, once attained, proves stable and so immune to the problem of free-ridership, yet the attainment of a cooperative equilibrium is impossible for parametrically rational agents, who can attain only equilibria that are individually accessible.

Elster illustrates the failure of several writers engaged in the use of biological explanations to avoid the fallacy of supposing that natural selection can yield outcomes accessible only to strategically rational actors. Perhaps the most interesting of these examples is the attempt by Gary Becker to show that altruism may increase genetic fitness simply because its indirect effects on the behaviour of others may afford benefits to the agent outweighing the cost of the direct effects. Elster argues that the framework of expectations that must exist among the interacting parties for these indirect effects to occur is one that depends on strategic rationality.

Functional explanations are frequently appropriate to account for the attainment of local maxima. But intentional explanations are needed to account for the attainment of global maxima. This view, if sound, opens up several important avenues for further exploration. It may suggest why conscious intentionality, involving the capacity to provide semantic representations of possible future states, is in general a fifth wheel in understanding animal behaviour. It may also suggest why in principle human behaviour can not be understood through such a device as operant conditioning. These are not matters that Elster develops. I mention them only as examples of the reflections suggested by his discussion to one reader (myself), in the hope that, together with the brief sketch of parts of his argument that I have offered, they will encourage others to give Elster's full account the careful reading and analysis that it deserves.

The second chapter begins with the incident from which the book takes its title. Ulysses demands to be bound hard and fast, so that he will be unable to respond to the song of the Sirens, and demands further that when, hearing their song, he seeks to be loosed from his bonds, they should instead be tightened. Ulysses, Elster notes, is imperfectly rational. Were he fully rational he would not need the precommitment he seeks; he would not be moved by the Sirens' song. But were he merely a creature of desire he would have no use for precommitment.

Precommitment, binding oneself, is the theme around which Elster organizes the discussion in the second chapter, although it is but one of many issues that actually figure in his account. Binding oneself, he proposes, is carrying out a decision that, without enlarging the set of options later available to one, increases the probability that one will carry out a second decision, and does this by setting up some causal process in the external world. (I have in fact simplified Elster's definition, but the

full account is unnecessary for our purposes.) Elster distinguishes binding oneself from rearranging one's inner space through such techniques as private side bets and consistent planning, in which no external causal process is involved. The rearrangement of inner space is, at least in my view, one of the essential devices open to the perfectly rational person. Without it Prisoner's Dilemma problems would be irresolvable. Of this dilemma more later, but Elster does not consider the rearrangement of inner space except as a strategy for precommitment. Even here his discussion barely scratches the surface of one of the central issues facing any account of rationality.

Elster distinguishes several strategies of precommitment: manipulation of the feasible set of actions, especially by making public side bets (offering security to a potential lender, wagering that one will succeed in stopping smoking); manipulation of the character, especially by inducing altered preferences; manipulation of information, especially by inducing altered beliefs (Pascal's advice to would-be believers: act as if you believe). However, it is not so much the general discussion, as the particular matters pursued – the logic of decisions to believe, the nature of inconsistent time preferences, endogenous preference changes - that hold the reader's attention. And a thorough exploration of any one of these matters would exceed the space available here. Let me merely touch on one or two points that I find of special interest.

Elster's discussion of endogenous changes of preference focusses on Von Weiszäcker's seminal proof that we may specify circumstances such that an agent has a long-term preference for x over y if and only if he has a sequence of short-term preferences leading from y to x; in other words, if and only if from y, he prefers r_1 to y; from r_1, he prefers r_2 to r_1; and so on until some r_n is reached from which he prefers x to r_n. Thus if an agent finds a conflict between his present preference for y over x and a long-term preference for x over y he may be led – and in some cases may lead himself – by a series of steps, each preferred at the time it is taken, from y to x. (He may have to be led, because some of the preferred steps may not lie in his feasible set of options.)

This process Elster terms *persuasion*. It occurs when an individual who prefers y to x comes to prefer x to y after a series of changes each of which itself accords with his or her preferences. Elster distinguishes persuasion from *seduction,* which occurs when an individual who prefers y to x comes to prefer x to y after being coerced into doing or choosing x. Neither process is, in Elster's view, ethically neutral. We might consider (my example, not his) an individual who comes to prefer alcoholism to abstinence despite his initial preference for abstinence through a sequence of short-term preferences; he prefers one drink each day to none, then two to one, then three to two, and so on until he prefers being drunk to being sober. But not all instances of persuasion or seduc-

tion are of this kind; both are devices that may be chosen by persons who would prefer to prefer *x* to *y* even though initially they prefer *y* to *x*.

The political implications of Ulysses' quandary also invite brief discussion by Elster; he examines the interesting paradox present in most democracies that 'each generation wants to be free to bind its successors, while not being bound by its predecessors (94).' From this, he goes on to examine Marx's theory of the capitalist state. I shall not attempt to sketch his argument, which is directed to showing the inadequacy of the empirical grounds on which Marx invoked precommitment devices (the abdication by capitalists of power so that their interests might better be served), but I do want to note his interesting comparison of Marx with Hobbes. Whereas Hobbes takes the role of the state to be to provide a cooperative solution to a universal Prisoner's Dilemma, Marx holds that the state enables the dominant class to attain such a solution among themselves while preventing the workers from realizing a cooperative solution to their own dilemma. This is a contrast that invites further exploration.

The chapter concludes by emphasizing three characteristics of being human: rationality, in the sense of deliberately sacrificing present gratification for future gratification; weakness of will; recognition of irrationality and the capacity to bind oneself as protection against it. 'This second-best or imperfect rationality takes care both of reason and passion. What is lost, perhaps, is the sense of adventure (111).'

The third chapter offers a catalogue of problems. Elster holds (rightly, in my view) that 'the "rational-choice" approach to human behaviour is without much doubt the best available model' (112); nevertheless, there are competitors, and there are evident weaknesses in the model. Any piece of human behaviour, Elster suggests, may be seen as the end product of two successive filtering devices, one of which is determined by the structural constraints that reduce the set of abstractly possible actions to the set of feasible actions, and the other of which is the mechanism that singles out one member of the feasible set. On the rational choice model this second mechanism 'is the deliberate and intentional choice for the purpose of maximizing some objective function (113).' But suppose that the initial filter – the structural constraint – is sufficient to reduce the feasible set to one member, or to trivially differing members. Then rational choice is of no importance. This is the *structuralist* view. Or suppose that the second filter – the mechanism that selects from the feasible set – is non-intentional. Perhaps it is random; perhaps it is *traditional*. (A broader form of the traditionalist view might be the position that selection is determined by *cultural practices,* which need not all be traditional in character. Indeed, there are some who might consider rational selection to be merely a case of selection by cultural practice.)

Parametric rationality, Elster claims, is almost always well-defined. Strategic rationality is not. In particular, strategic rationality requires equilibrium points, at which each individual's action is a best response to the others' actions. In some situations there are multiple equilibrium points with no criteria for rational selection among them; in others there are no such points.

What then does the rational individual do when confronted with a situation in which strategic considerations lead to no determinate conclusion? Some theorists of rational choice, such as John Harsanyi (whom Elster discusses briefly), endeavour to develop a general theory of rational behaviour which effectively eliminates such difficult cases. That such a theory is often not exemplified in actual behaviour need be no objection to it; the theory is normative rather than explanatory. In the real world imperfectly rational agents just get tired, or – taking the costs of decision into account – fall back on satisficing considerations, or perhaps follow some established social practice. Elster discusses, briefly and inconclusively, the relation between satisficing and maximizing approaches to rational behaviour; the key issue here is whether satisficing is simply a species of maximizing.

In any discussion of strategic rationality the problems associated with the Prisoner's Dilemma must play a central role. The core idea, as is now well known, is that in a Dilemma, if each person acts on a clearly maximizing strategy (indeed, on a dominant strategy assuring that one's behaviour is best whatever others choose to do) then the outcome fails to be Pareto-optimal; each could be in a more preferred position without cost to the others. (The Prisoner's Dilemma is, in an obvious way, the converse of the perfectly competitive market. In the market there is an accessible equilibrium which is optimal. In the generalized Dilemma, no equilibrium is optimal.)

Prisoner's Dilemma situations cry out for some form of reciprocal altruism, in which an optimum is achieved through strategies that ensure that each gains more from the benefits others confer on him or her than he or she loses by the costs undergone in conferring benefits on them. But how to show that such reciprocal altruism is rational? Hobbes is the first to face this problem, in his discussion of the Foole, and his efforts, like all that have come after him, have not been judged very successful. Elster canvasses some recent discussions (although he seems to be unaware of my own attempt in 'Reason and maximization,' published in this *Journal* in 1975). A form of evolutionary reductionism proposed by Robert Trivers is described that appears to be in some important respects close to my own 'logically reductionist' solution. Both positions argue that the tendency to behave altruistically accords with self-interest although particular altrustic acts do not accord with interest; both insist that concern for others is not merely simulated, and that 'the altruism is

the more efficient because it is *not* derived from calculated self-interest (145).' Elster acknowledges that he lacks the competence to evaluate Trivers' analysis.

Elster's own last words, that '*rational man* can ... be moved by concern for others' (146) seem clearly inadequate to the problem. Those who are moved by concern for others may find themselves in Prisoner's Dilemma situations just as those who are not so moved (as Howard Sobel has conclusively shown). What is required is a selective concern that expresses itself as a willingness to cooperate (with others who are also so willing) in situations with Prisoner's Dilemma structure. The problem to which reciprocal altruism is the solution is one arising out of logically possible structures of interaction, not one arising from self-directed rather than other-directed concern. What must be shown is not that rational men and women need other-directed preferences, but that they have an interest in developing a disposition to cooperate.

The final chapter turns to such forms of irrationality as hate, love, and self-deception. Hate and love may give rise to unrealizable intentions. I can love and seek the love of someone only if I can respect her, but anyone who would love me can not be worthy of respect. Although Elster's discussion here is not without interest, it does not seem to me to raise questions of deep import for our grasp of rationality (or to afford deep insight into the themes it treats). Perhaps this reflects my biases. But I shall leave this last chapter without further ado.

If there is a single theme that brings together the complex strands of Elster's thought, it is that the explanation of human actions must be intentional, founded on a presumption of rationality, where rationality is understood on the rational-choice model. Elster's catalogue of problems is not meant to give comfort to those who espouse some other view of social-scientific explanation. Perhaps I am more optimistic than he about the full reach that the rational-choice model can have. But we are clearly in the same camp here, and differences of detail do not promise profitable brief discussion.

Although the middle chapters of the book raise issue after issue of real significance and interest, it is only in the first chapter that the reader will find a sustained argument. Here it seems to me that Elster offers an impressive case for the view that humans are global maximizing machines and that functional and intentional explanations are to be kept respectively to the biological and social sciences. Of course, his treatment is no 'last word,' but I think his arguments shift any burden of proof to those who hold otherwise.

There are a few rather serious misprints and errors in the book. On page 24, line 4, 'used' should surely be 'mixed'. On page 25, Dawkins' example of hawks and doves is incorrectly set out. Given Elster's account of the payoffs, the proportion of doves to hawks would converge

towards 1:2, not 5:7. In Dawkins' account convergence to 5:7 will occur, but Dawkins' payoffs are not correctly represented by Elster. In the diagram on page 81, the lines AA (representing the farmer's budget) and BB (representing the industrial worker's budget) are interchanged. These matters are apt to cause the reader unfamiliar with game theory or economics some unnecessary trouble.

April 1982

DAVID GAUTHIER
University of Pittsburgh