[Elster, Jon (1993), Some unresolved problems in the theory of rational behaviour, Acta Sociologica 36 (3): 179-190]

Some Unresolved Problems in the Theory of Rational Behavior

Jon Elster

In an article written in 1977 the author offered a survey of unresolved problems in rational choice theory. The present paper is an attempt to rethink this issue. On the one hand, it emphasizes the question of indeterminacy, i.e. situations in which the rational choice is not well defined. The paradoxes of backward induction find their place here, as do the existence and importance of genuine uncertainty (as distinct from risk). On the other hand, the article discusses the question whether preferences can be said to be rational. Examples include time preferences, attitudes to risk, regret and the 'taste for fairness'. The examples are chosen with a view to showing that rational choice theory is not a predictive theory, hut essentially a hermeneutic one. As part of the enterprise of self-understanding, the construction of rationality is partly discovery and partly decision. There is no right answer to all questions.

Fifteen years ago I wrote a long survey essay with the same title as this paper, 'Some unresolved problems in the theory of rational behavior'.[1] The present paper is to some extent, but only to a small extent, an attempt to take stock of what has happened in the meantime. I shall not really be discussing whether any of the problems I identified then have been solved. Instead, I shall consider some problems that seem to be live issues today. My selection of issues is somewhat idiosyncratic, because my real purpose is largely a methodological one. I want to convey a feeling for the nature of the rational choice enterprise. Rational choice theory is far more than a technical tool for explaining behavior. It is also, and very importantly, a way of coming to grips with ourselves - not only what we should do, but even what we should be.

The very notion of an 'unresolved problem' is, therefore, somewhat naive. Consider the paradigm of an unresolved problem in mathematics, Cantor's continuum hypothesis. Originally stated by Cantor in the late 19th century, the hypothesis asserts that there is no infinite set which is strictly larger than the set of natural numbers while strictly smaller than the set of real numbers. The problem was whether this assertion can be derived from the axioms of set theory, is independent of them, or contradicts them. Kurt Gödel showed around 1940 that the hypothesis is consistent with the axioms, and Paul Cohen then showed around 1965 that it is genuinely independent of the axioms. These results, once understood and accepted, were irrevocable and definitive. Anyone who can understand the problem and follow the proof is compelled to accept the solution.

A 'problem' in rational choice theory cannot have a solution in this sense. This is because the theory is, in an important sense, hermeneutic or interpretative. To understand this, let us consider one of the problems I discussed in my earlier survey, the existence of games without a non-cooperative solution. That a game has a solution means, roughly speaking, that without communicating with each other the players can tacitly converge towards a set of actions that are fully anticipated by all. A necessary condition for a set of actions being a solution is that they are optimal against each other, so that nobody can improve the out-come for himself by unilaterally changing his strategy. A set of actions with this feature is called an equilibrium of the game. If the game has only one equilibrium, it is *ipso facto* the solution. When there are several equilibria, one of which dominates all the others, in the sense of being better for some and worse for none, it is the solution. When the game has several undominated equilibria, we are in trouble. In such cases, there is no obvious way of selecting one equilibrium as being *the* solution to the game.

Over the past fifteen years or more, John Harsanyi and Reinhart Selten have been working on this problem. Recently they published an important book on the subject, *A General Theory of Equilibrium Selection in Games.*[2] I have neither the space nor the competence to summarize their results here. But I would like to quote from Robert Aumann's preface to their book. Aumann writes, 'Although the theory selects a unique equilibrium, as a theory it need not be unique'. To me this suggests that our analysis of rationality could follow the methodology outlined by John Rawls in his work on distributive justice.

According to Rawls, an attempt to construct a theory of justice must start with intuitions about what it would be fair or just to do in particular cases. To be successful, the theory must do two things. First, all intuitive judgments and no counter-intuitive judgments should follow from the theory. Second, the theory itself should be independently plausible. Rawls then goes on to say that the first requirement is too strong. The theory itself may force us to shed or modify some of our intuitive judgments, by helping us to see similarities or differences that otherwise would not have appeared to us. In the next round, the theory will have to adjust itself optimally to these revised intuitive judgments. The process may eventually reach an end, in which there is perfect fit between the theory and the intuitions, a state that Rawls calls 'reflective equilibrium'.[3] There may, however, be several reflective equilibria. When there is lack of fit between a theory and the intuitions, fitness-increasing revisions can take several forms. The final reflective equilibrium may then depend on which direction we take. A theory of justice, considered as a reflective equilibrium, may well be able to define, for any given situation, the unique just distribution. But the theory itself need not be unique.

I submit, therefore, that to construct a theory of rationality we must follow a similar procedure. We must begin with preanalytic, intuitive notions about what, in various situations, we ought to do if we want to do as well for ourselves as we can. There is no reason to expect that all these intuitions are coherent with each other. For instance, in games with several undominated equilibria one intuition is that we should choose the action with the highest security level, that is, use a maximin strategy. Another intuition is that we should act on the assumption that other players in similar situations will act in similar ways. A third intuition is that our action should be the best response to their actions. But these intuitions are inconsistent with each other. Maximin strategies are usually not best replies to each other.

When intuitions conflict, we have to make choices about which way to go. A given succession of such choices can then lead to a coherent theory that uniquely tells us what, in any given situation, is the rational thing to do. Or, more modestly, it may lead to a precise circumscription of the situations in which rationality has unique implications for action. Initial intuitions about rational action can be modified in two ways as a result of theoretical reflection. Either we may come to think that the rational thing to do is not what we first thought, but some other action. Or we may come to believe that there is no uniquely rational action under the circumstances.

In any case, the conclusion I want to draw is that rational choice theory is not a predictive theory, but essentially a hermeneutic one. As part of the enterprise of self-understanding, the construction of rationality is partly discovery and partly decision. There is no right answer to all questions. Of course, many questions do have right answers. We can construct local theories of rationality, which apply to special problems and which are very robust in the sense of resting on strong and shared assumptions. But full consensus is not to be expected. It might obtain by accident, but would soon disintegrate.

I now pass to some more specific issues of taking stock. The list of unresolved problems I drew up in 1977 includes the following. First, the problem of games without solution, just referred to. Second, the anomalies created by lexicographic preference structures. Third, problems related to subjective probability. Fourth, the relation between maximizing, satisficing and natural selection. Fifth, the relation between rational behavior, traditional behavior and random behavior. Sixth, the problem of explaining altruistic behavior. And lastly, problems of endogenous preference change. I think I was wrong in attaching great importance to lexicographic preferences. In most real-life situations, what may look like a lexicographic preference is just a very steep trade-off. The other problems, however, do seem to retain their importance. I shall not, however, go through the tedious process of discussing them one by one. Instead, I shall discuss some current areas of disagreement and exploration, with the emphasis on conceptual rather than on technical issues. Indirectly, I shall touch on all the issues raised in my earlier survey, but I shall not stop to point out the connections.

Let me first, in a very general way, distinguish between two types of challenge to rational choice theory or, for that matter, to any theory. There are two ways in which theories can fail to explain: through indeterminacy and through inadequacy. A theory is indeterminate when and to the extent that it fails to yield unique predictions. It is inadequate when its predictions fail. Of these, the second is the more serious problem. A theory may be less than fully determinate, and yet have explanatory power if it excludes at least one abstractly possible event or state of affairs. To yield a determinate prediction, it must then be supplemented by other considerations. The theory is weak, but not useless. It is in more serious trouble if the event or state of affairs that actually materializes is among those excluded by the theory.

In rational choice theory the emphasis may be on prescription rather than on prediction. The same kinds of failures may then occur. The theory may fail to tell people what to do. In that case, the theory is indeterminate. Or people may fail to do what the theory tells them to do. In that case, people are irrational. I shall discuss problems of indeterminacy and problem of irrationality in that order.

The problem of games without solution. is a problem of indeterminacy. I shall no say more about it here. Instead, I shall consider two other sources of indeterminacy: the logic of backwards induction and decision-making under uncertainty.

The paradox of backward induction is related to what is known as the examination paradox or, alternatively, as the hangman' paradox. It is not, however, identical to that problem, and I shall leave the connection unexplored. Imagine, now, the following kind of game between two players. In the opening move, one player can choose between quitting the game and continuing to play. If he quits, the players receive their payoffs. If he continues, it is the other player's turn to move. She can either quit or continue to play. If she quits, payoffs are distributed. If she continues, the first player will have the next move. The play can go on like this up to, say, the fourth round in which the second player has the choice between two actions, each of which brings the game to an end and distributes payoffs to the players.

We now stipulate that payoffs are as shown in Figure 1 overleaf.

In the last round the second player has the choice between an action that offers 3 to herself and 6 to the other player, and another action that offers 3 to the first and 4 to herself. In the third round the first player can ensure 4 for himself and 1 for the first player by quitting. In the second round the second player can ensure 2 to herself and 1 to the first player by quitting. In the opening round the first player can ensure 2 for himself and a loss of 1 for the second player by quitting.

It seems clear what will happen, assuming that both the players are rational, know each other to be rational, and so on. In the last round, the second player will clearly

```
  I ------> II ------> I ------> II ------>(6,3)

  |          |         |          |

  |          |         |          |

(2,-1)     (1,2)     (4,1)      (3,4)
```

Figure 1.

ensure 4 for herself rather than 3. In the penultimate round, therefore, the first player will rather quit and get 4 than let the second player have a chance to make a move that will leave him with 3. But this means that in the second round the second player will quit to get 2 rather than to give the first player a chance to make a move that will leave her with 1. But this means that the first player will quit in the first round to get 2, rather than let the second player have a chance to make a move that will leave him with 1. This out-come is worse for both than what could have been achieved by going on to the last round. As in the Prisoner's Dilemma, individual rationality leads to collective suboptimality.

The reasoning leading up to the conclusion that the first player will quit in the first round seems compelling. But it harbors a paradox. The reasoning requires the first player to go through a thought experiment, including the assumption that the last round has been reached. But we know that this round will never be reached if the players are rational. The assumption that it is reached can be entertained only if one or both players are assumed to be irrational. But in that case the backwards induction argument, which requires both players to be rational, cannot be carried out. This is the paradox of backwards induction.[4] It seems to suggest that a game like the one just described has no rationally prescribed behavior.

The conclusion is hard to accept. The logic of backwards induction is intuitively compelling. It may not be the last intuition we give up, but it is more compelling than most. There is a very active search, therefore, for a way out of the dilemma. I cannot report these efforts here, except to say that none has so far succeeded in summoning general agreement. If, in the end, backwards induction has to go, large parts of economic theory will also have to be thrown overboard. In particular, many recent advances in bargaining theory will unravel.

Another source of indeterminacy is brute uncertainty - so-called to distinguish it from strategic uncertainty in game-theoretic situations. There are many situations in which we know too little about the choice situation to be able to make a well-considered decision. Career choice may be taken as an example. Suppose than I am about to choose between going to law school or to a school of forestry - a choice not simply of career but of life style. I am attracted to both professions, but I cannot rank and compare them. If 1 had tried both for a lifetime, I might have been able to make an informed choice between them. As it is, I know too little about them to make a rational decision. What often happens in such cases is that peripheral considerations move to the center. In my ignorance about the first decimal - whether my life will go better as a lawyer or as a forester - I look to the second decimal. Perhaps I opt for law school because that will make it easier for me to visit my parents at weekends. This way of deciding is as good as any - but it is not one that can be underwritten by rational choice as superior to, say, just tossing a coin. Let me cite Samuel Johnson on the topic:

Life is not long, and too much of it must not pass in idle deliberation how it shall be spent: deliberation, which those who begin it by prudence, must continue it with subtlety, must after long expense of thought conclude by chance. To prefer one future mode of life to another, upon just reasons, requires faculties which it has not pleased our Creator to give to us.[5]

In some cases where we are able to find out what to do, the cost of finding out may be prohibitive. I have looked at child custody decisions, which illustrate this problem.[6] According to custody legislation in most Western countries, custody shall follow the best interest of the child, that is, be given to the parent that will best promote the child's emotional and intellectual development. Now I think that in many cases it is virtually impossible to say which of the parents is the more fit for custody. But even assuming that one can find out, it is not clear that it is in the child's interest to try to find out what is in the child's interest. Custody litigation is protracted and ugly. All parties suffer, but the child most of all. A swift decision by a presumption rule or even the toss of a coin might be a better procedure.

These two examples suggest the following consideration. Human beings do not simply have material and emotional needs. They also have, for whatever evolutionary reasons, intellectual needs. One such need is the need to find meaning and patterns in the events we observe. This need is satisfied by genuinely scientific theories, but also by pseudoscientific views of all sorts, ranging from astrology to functionalist sociology. Another is the need to have and be able to cite reasons for our actions and decisions. Sometimes we know that we could find the decision that would have been optimal if found costlessly and instantaneously. By investing more time, effort and money we may be able to rank the options. We may also know, or be in a position to know, that the benefits from finding out are small compared to these costs. Yet because of what one might call *an addiction to reason* we do not use a lottery, but go on looking for reasons, until eventually we find one. I believe the child custody case brings this out with special poignancy. To promote the best interest of the child, the compulsive rationalist searches for evidence of fitness and unfitness of the parents while, in the meantime, the damage done to the child by the process of searching exceed the benefits to be expected from the search. It is more than rational in such cases to resist the sirens of reason

In an article from 1913 Otto Neurath characterized the belief that we can always have good reasons for our decisions as *pseudo rationalism* Whereas Cartesian rationalism sees its chief triumph in the clear recognition of the limits of actual insight', pseudo-rationalism 'leads partly to self-deception, partly to hypocrisy'. By way of conclusion to this part of the paper can do no better than to quote his further comments on this distinction:

The attitude of Thomas Hobbes in the matter of religion ... rarely finds approval. His idea that some order is better than none enrage every pseudorationalist who hopes to reach decision by an adequate measure of thinking. Hobbes' intolerance is purely external, a means to an admitted political end. He simply feels unable to decide which of the positive religions is preferable. It appears to me that this behaviour of Hobbes is the only one possible for an honest rationalist in many affairs of life; however, whether rationalism is at all suited to regulate public life is another question. But once tradition and community feeling are weakened, there is no choice but that between rationalism, which undoubtedly leads to drawing lots, and pseudorationalism which falsifies thinking and feeling ...

Let us go back to the parable of Descartes. For the wanderers lost in the forest, who have no indication at all as to which direction to follow, it is most important to march on energetically. One of them is driven in some direction by instinct, another by an omen; a third will carefully consider all eventualities, weigh all arguments and counter-arguments and, on the basis of inadequate premises of whose deficiencies he is unaware, take one definite direction which he considers the correct on The fourth, finally, will think as well as he can, but not refrain from admitting that his insight is too weak, and quietly allow himself to decide by lot. Let us assume that the chances of getting out of the forest are the same for the four wanderers; nevertheless there will be people whose judgment of the behaviour of the four is very different. To the seeker after truth, whose esteem of insight is highest, the behaviour of the last wanderer will be congenial, and that of the pseudorationalist third wanderer most repellent. In these four kinds of behaviour we can perhaps see four stages of development of mankind without exactly claiming that each of them has come into full existence. [7]

It is clear from Neurath's account of 'presudorationalism' that it is in fact a form of irrationality. And this brings me to the second category of unresolved problems I want to discuss. Let me begin with the trivial observation that many forms of behavior appear to be irrational. And I can add the equally uncontroversial observation that many of these will be universally recognized to be truly irrational. Some kinds of mental illness induce behavior that nobody would think of calling rational. But when from the class of apparently irrational behavior we subtract the class of uncontroversially irrational behavior, we are left with a large number of controversial cases. Economists relentlessly try to persuade us that many of these are actually instances of rational behavior. I shall consider four examples where it seems to me that the jury is still out. They are myopia, regret, indignation and revenge. Many, many others could have been cited, but these are all I have the space to discuss,

Myopia is the tendency to structure inter-temporal choices so that welfare in the present is weighted more heavily than welfare in the future, over and above what might be justified on the basis of mortality tables. There is strong evidence that people tend to behave in this way. And it is clear that sometimes it gets them into trouble, Unless one has a large fortune, an iron constitution and a good lawyer, total disregard of the future is likely to be disastrous, And even whet' some account is taken of the future, myopia tends to make one's life as a whole worse than it could otherwise have been. It is tempting, there-fore, to conclude that myopic behavior is irrational.

But there is another side to the problem. Myopia is a form of preference a preference for the present over the future. It is a temporal analogue to selfishness tile tendency to prefer one's own welfare to that of other people. Like selfishness, it may be regrettable or even stupid - but that is not reason to call it irrational. It is just a preference, a taste, and *de gustipus non est disputadum..*

Table 1:

	t_1	t_2	t_3
A:	1	5	5
B:	4	2	2
C:	2	2	4
D:	5	5	1
E:	6	1	1
F:	1	1	6

But there is still another way of looking at the matter. Myopia may be seen as a cognitive problem - a defective telescopic faculty, as Pigou said. Just as the distant past is less vivid than the recent past, it is harder to relate to the distant future than to the present and the near future. It appears that securities and futures markets are excessively sensitive to current information, and insufficiently sensitive to past Information. If true, this observation would be a clear case of irrational behavior, Why, then, should we not say the same about the inability to take proper account of future gains and losses?

There is a further issue to be explored. Although strong time preferences are certainly welfare-reducing in a life-time (ex post) perspective, we should not assume that the only rational behavior is to choose the option that produces the largest sum of undiscounted values. Consider the streams of utility over three periods (Table 1).

If a person is observed to choose B over A, a natural explanation might be in terms of myopia. Although he gains in the present, his life overall is made worse off. however, if he also prefers C over A, this interpretation will have to be revised. It Is then conceivable that the person is moved by a desire that his welfare shall never fall below the level of 2, or that he is fulfilling an intrapersonal maximin principle. This interpretation is strengthened if he can also be shown to prefer C over D.

The intrapersonal maximin principle satisfies a basic requirement of rationality: no year shall be preferred simply because it is close in time or, for that matter, distant in time. The principle of maximizing total utility over one's lifetime also satisfies that requirement. Compared to the latter, the maximin principle is welfare-reducing. However, there is no reason to assume that it has been created by a heteronomous psychic mechanism. The desire for security, comfort and stability, even at the expense of total utility, could be entirely autonomous. By contrast, a person who chose E over all the other alternatives might plausibly be assumed to be in the grip of the pleasure principle. In a still further contrast, the person who prefers F over A and D might be moved by the desire to have, at least once in his life, a year of high living. This desire, too, satisfies the criterion of intertemporal impartiality. Although the preference is welfare-reducing, I see no reason for assuming that it must be heteronomous. Here, *rationality does not imply maximization.* It does imply, however, that all consequences are considered, and that none are given lesser weight simply because they come early in the time sequence.

The upshot of this discussion is neither that myopia is rational nor that it is irrational. I simply wanted to make you sensitive to the complexities of the issue. Should rationality be considered from the point of view of one's life as a whole, or from the point of view of the moment of choice? Should the objects of intertemporal choice be conceptualized as future states or as present and possibly defective representations of future states? Should the self that makes the choice be considered as the authoritative spokesman for my successive states, or should it be disqualified because of its very proximity to the choice? It should be clear that these are not issues that can be 'resolved' in anything like the way in which mathematical problems are resolved.

I next want to turn to regret, but I shall do so in a somewhat indirect manner. One of the most interesting developments of rational choice theory over the last decade has been the development of so-called non-expected utility theory. This is a theory, or rather a set of competing theories, that try to account for various phenomena that do not fit into the expected utility model that is the standard tool of economic theory. The best-known of the anomalies is the so-called 'Allais paradox'. Like the other anomalies, it has to do with decision-making under risk; that is, with choices that can have various outcomes with various probabilities. For my purposes here there is no need to explain exactly what the paradoxes consist in. Nor do I need to explain in detail what the competing non-expected utility theories amount to. It is sufficient to say that some of the theories clearly imply that people who behave in the anomalous way are irrational, whereas others do not, or do not obviously, have this implication.

One theory explains the anomalies in terms of anticipated regret. Suppose I have the choice between two actions, taking an umbrella or leaving it at home. If I leave it home and it rains, I shall get wet. In addition, I'll feel regret that I didn't take the umbrella. This contingency is, in other words, doubly bad. It can be shown, although I cannot do it here, that by taking account of feelings of regret, over and about preferences about the physical states themselves, some of the anomalous forms of behavior can in fact be explained.

Does this mean that the anomalous behavior is rational? Once again, intuition is ambiguous. On the one hand, regret seems just like another preference. To include regret in the decision calculus we have to expand the space of outcomes, but there is nothing irrational about that. To include altruism in a rational choice model, we also have to expand the space of outcomes, to include other people's pleasure as well as my own, but it is hard to see why anyone would object to this practice.

On the other hand, there is something about regret that does seem irrational. Rational choice theory tends to lead to recommendations of the following kind: don't cry over spilt milk, let bygones be bygones, cut your losses, and ignore sunk costs when deciding for the future. To worry about what might have happened seems peculiarly pointless - a needless source of frustration and unhappiness. Surely the best pieces of advice we can give to our children are, first, think about your future and, second, don't worry about the past. But how can we justify this advice if myopia and regret are in fact rational?

```
          I
        /   \
      /       \
  (2,2)    II
             / \
           /     \
       (3,1) (0,0)
```

Figure 2.

I want to consider two further cases of backward-looking behavior: indignation and revenge. Consider first Figure 2.

Here, player *I* can either go left and ensure a reward of 2 for both players, or go right and leave it up to player *II* to make the next move. In the latter case the second player will, if she is rational, go left and ensure a reward of 3 for the first player. The first player should therefore go right, if he is rational.

In experimental situations this is not what happens. A typical situation is one in which the experimenter tells one of two subjects that he has the right to propose a decision of ten dollars and the second that she has the right to accept it or reject it. If the proposal is accepted, both get what was proposed. If it is rejected, neither gets anything. Clearly, if both are rational the first player would propose a division of 9.99 to himself and one penny to the other. But this is not what is observed. Instead, three findings stand out. First, most people placed in the position of the first player propose a much more equal division. Secondly, when a very unequal division is proposed by the first player, the second player often decides to take nothing. Thirdly, most proposals that are made and accepted are somewhat biased in favor of the proposer.

The first two findings are hard to reconcile with forward-looking rationality. To explain them we must invoke something like norms of fairness, indignation, or a dislike to be taken advantage of. Is this rational? One might argue that it can be rational to cut off one's nose to spite one's face if one thereby sends a signal to the other person that one is unwilling to be taken advantage of. But in some of these experiments the two subjects interacted only once, and then only by the intermediary of a computer. The results were the same. Or one might say that feelings of indignation have their place in the calculus of decision. But are such feelings rational? One might say that, like regret, they simply express the person's taste. And the third finding might then be expressed by saying that the person trades off material wealth against the taste for fairness. Fairness is not an absolute value, but just another good on a par with dollars or kroner. Alternatively, one might argue that a person exhibits rationality by the ability to overcome feelings of this kind. Once again, I do not conclude - but simply offer the matter to your reflection.

Let me finally discuss the case of revenge at somewhat greater length. I define revenge as the attempt, at some cost or risk to oneself, to impose suffering upon those who have made one suffer, because they have made one suffer. Such behavior can arise in two different ways. On the one hand there is what we might call pre-social revenge - the spontaneous, unreflective urge to impose suffering on another person or, for that matter, a material object that has caused us harm. This kind of behavior is universal. On the other hand there is revenge that is regulated by social norms. Vendettas or blood feuds are found in many societies, but in far from all. The problem of revenge from the point of view of rational choice theory is that it seems pointless or worse. If the harm has already been done, why should I make it worse by exposing myself to the risks and costs of vengeance?

The following arguments have been made or could be made. First, in societies with norms of revenge an individual may be worse off if he fails to avenge an affront, because of the external sanctions to which he thereby exposes himself. Secondly, a person who demonstrates that he cares about getting revenge will often have an edge when dealing with people who don't. For that reason it may pay to cultivate a reputation for caring about revenge. Lastly, revenge can be seen as a tit-for-tat strategy in iterated collective action problems.

The first argument is very general, and not restricted to norms of revenge. All norm-guided behavior, or so the argument goes, is kept in line by the fear of external sanctions that make it individually rational for the agent to abide by the norm. There is no need to invoke internal variables like emotions to explain why people act against what appears to be their material self-interest: it is sufficient to observe that the alternative is even worse. The tangible costs of violating norms exceed the tangible costs of adhering to them.

Now the case of revenge is one in which this account is particularly implausible. If it were true, revenge behavior would be singularly overdetermined, because the internal, emotional motives also seem to provide a sufficient explanation. The argument also fails more generally, however, for reasons which apply to all norm-guided behavior. It is true that violators of a norm which they share with other members of their community are often exposed to sanctions by these others, ranging from raised eyebrows to crippling forms of social ostracism. But then we have to ask what reasons these others could have for sanctioning the violators. The obvious answer is that for any ordinary norm there is a meta-norm that enjoins people to punish people who fail to punish violators of the first-order norm. A system of sanctions might keep people in line even if nobody believes in the norm. But this argument soon runs out of steam. *Expressing* disapproval is always costly, whatever the target behavior. At the very least it requires energy and attention that might have been used for other purposes. One may alienate or provoke the target individual, at some cost or risk to oneself. On the other hand, when one moves upwards in the chain of actions, beginning with the original violation, the cost of *receiving* disapproval falls rapidly to zero. It is a brute empirical fact that people do not frown upon others when they fail to sanction people who fail to sanction people who fail to sanction a norm violation. Consequently, some sanctions must be performed for other motives than the fear of being sanctioned.

The second argument for the rationality of revenge behavior derives from Thomas Schelling[8] and amounts, in effect, to an argument for the rationality of appearing to be irrational. Consider for instance the game shown in Figure 2.

Assume first that both players are fully rational, and not moved by backward-looking considerations. Player *I* knows that if he goes right, Ifs self-interest will induce her to go left, thus ensuring the best outcome for *I*,

Assume next that *II* is believed to be truly irrational, because in the past she has consistently refused to let bygones be bygones. *I* knows that if he goes right, she will resent it sufficiently to go right herself, effectively cutting off her nose to spite her face. Being rational, *I* will go left. It would be a pointless play on words to say that *II*, when behaving irrationally, is in fact being rational. Her irrationality is useful to her, but it is none the less irrational.

The argument does not establish a case for the unconditional usefulness of obeying a norm of revenge. It establishes at most that one could benefit from obeying the norm when dealing with other people who don't obey it. If everyone abides by the norm an individual might well be better off not abiding by it. This argument has to be stated carefully. I do not mean that in actual feuding societies, unilateral cowardice is a rational strategy. Social ostracism by third parties might well make that option unacceptable. What I claim is that *if* the only cost of cowardice were the loss incurred in conflictual encounters, an isolated coward might do better for himself than the average norm-follower. Followers of the norm of revenge will tend to meet other followers. If they have a substantial chance of being killed in each encounter, because neither side will back down, their life expectancy is pretty poor. The coward who yields up a contested resource without protesting might do better for himself. If the tendency to engage in spontaneous revenge behavior is genetically determined, we would expect that in an evolutionary stable equilibrium some individuals would have the revenge genes and others not, with equal expected fitness for both groups.

Assume now, however, that *II* is fully rational, but deliberately engages in acts of vengeance to create an impression that she is irrational. If *I* is rational, he will not take these acts at face value. He will know that there is some probability that *II* is in fact irrational, and also some probability that she is just faking irrational behavior to build up a reputation for toughness. Depending on the actual probabilities and on what is at stake, he might well decide to abstain from provoking her. By this mechanism, it could indeed be rational to engage in acts of revenge.

Note, however, that the mechanism is parasitic on the existence of some genuinely irrational persons in the population. What drives the argument is the common knowledge that society has some rational and some irrational members, but that they do not bear their rationality or lack of it on their face. In a population of individuals known by each other to be fully rational nobody would ever exact revenge. This comment parallels a comment I made on the first argument for the rationality of revenge. There, I observed that sanctions cannot sustain a norm of revenge unless some of the sanctioners are genuinely moved by the norms. Here, I am saying that faking adherence to the norm cannot be a rational strategy unless some people genuinely adhere to it. Neither mechanism, therefore, can *fully* explain revenge practices as rational behavior.

The third argument for the rationality of revenge does not rest on the presence of some irrational believers in the norm. Rather, it asserts that threats of revenge can be part of a cooperative equilibrium, because the knowledge that defectors will be punished keeps everybody in line. Here, the purpose of a threat of punishment is simply to deter. The threat would fail if it had to be carried out. Actually revenge could occur only if someone acted irrationally, but then it is not clear that revenge would be a rational response to that act. A person who by defecting has shown himself to be irrational might not be moved by revenge. In that case, the rationality of the revenge threat might itself be called into question. This, in fact, brings us back to the question discussed earlier, about the rationality of backward induction arguments.

Let me try to connect this example with some of the earlier remarks. In my experience, believers in rational choice theory tend to say that revenge is rational because the person known to be vengeful usually gets his way. What they really mean, I believe, is that a propensity for revenge can be *useful.* Now one might want to equate the rational with the useful. One might argue, for instance, that rationality should not be defined in terms of subjective attitudes, but in terms of objective adaptation to the environment. Individuals with pointless or self-destructive attitudes will soon be eliminated by natural or economic selection, so that in equilibrium we will in fact observe only adaptive behavior. This, according to some, is what we should mean by rationality.

Now for many reasons I do not think this is a fruitful line of argument. Selection is often inefficient. When it is efficient, it may lead to polymorphism rather than to one form of behavior driving out all others. Also, the emphasis on objective adaptation makes it hard to answer the normative questions that are part and parcel of the concern with rationality. At the same time, I have been arguing myself that some kind of objective component of rationality may be necessary. Myopia and regret are irrational if they make people pointlessly miserable. But I'm not so sure about indignation. I cannot really bring myself to think that it is irrational to be willing to take a loss rather than be unfairly exploited. To have a coherent theory of rationality, I might have to swallow that conclusion, but I would rather not.

This paper has been deliberately non-conclusive and tentative. I do not think the nature of the subject matter allows for more. If one is interested in rationality exclusively for the sake of predicting behavior, some of the conundrums would disappear, but others would remain. But I do not think this is our only reason. We care about rationality because we want to be rational and want to know what rationality requires us to do.

Received March 1993 Final version accepted April 1993

Notes

[1] Ch. 3 in Jon Elster, *Ulysses and the Sirens* (Cambridge: Cambridge University Press, 1979).

[2] Cambridge, MA: MIT Press, 1988.

[3] This concept of equilibrium is not related to the game-theoretic one.

[4] For a more elaborate analysis, see Ken Binmore, Modeling Rational Players', *Economics and Philosophy* 3 (1987), 179-214.

[5] *J.* Boswell, *The Life of Samuel Johnson*, AD 1766 (Aetat 57) - a letter from Johnson to Boswell dated August21, 1766. Note that Johnson here offer two distinct arguments against the possibility of making a rational choice between different careers or, more generally, ways of life. The first is that our reason is too limited to allow us to assess and weight the long-term consequences of the options. The other is that even if we could in fact predict our future happiness under the various alternatives, the calculations would take so long that they would absorb much of the time available for living. Johnson often used the latter argument. 'We talked about the education of children; and I asked him what he thought was best to teach them first. JOHNSON. "Sir, it is no matter what you teach them first, any more than what leg you shall put into your breeches first. Sir, you may stand disputing which is best to put in first, but in the mean time your breech is bare. Sir, while you are considering which of two things you should teach your child first, another boy has learnt them both" (ibid., Ætat 54). 'He did not approve of late marriages, observing that more was lost in point of time, than compensated for by any possible advantages. Even ill assorted marriages were preferable to cheerless celibacy' (ibid., Ætat 61).

[6] Jon Elster, *Solomonic Judgements* (Cambridge: Cambridge University Press, 1989), Ch. III.

[7] Otto Neurath, 'Die verriten des Cartesius und das Auxiliarmotiv: zur Psychologie des Entschlusses' (1913). Cited after the translation in Otto Neurath, *Philosophical Papers 1913-1946,* pp.1-12. Dordrecht: Reidel, 1983.

[8] *The Strategy of Conflict* (Cambridge, MA: Harvard University Press, 1960), ch. 5.

References

Binmore, K. 1987. Modeling Rational Players. *Economics and Philosophy*. 3, 17~214.

Boswell, J. 1766. *The Life of Samuel Johnson.*

Elster, J. 1979. *Ulysses and the Sirens.* Cambridge: Cambridge University Press.

Elster, J. 1989. *Solomonic Judgements.* Cambridge: Cambridge University Press.

Neurath, O. 1983. *Philosophical Papers 1913-1946.* Dordrecht: Reidel.

Schelling, T. 1960. *The Strategy of Conflict.* Cambridge, Mass.: Harvard University Press.

Selten, R. 1988. *A General Theory of Equilibrium Selection in Games.* Cambridge, Mass.: MIT Press.