Elster excerpt

I

WHEN RATIONALITY FAILS

I.1 Introduction

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 an event or state of affairs that actually materializes is among those excluded by the theory. In saying this, I am not espousing naïve falsificationism, but simply making the common-sense observation that it is worse for a theory to predict wrongly than to predict weakly but truthfully.[1] In the former case it must be replaced or modified, not supplemented.

My concern here is not with scientific theories in general, but with failures of rational-choice theory. As argued below, rational-choice theory is first and foremost a normative theory and only secondarily an explanatory approach. It tells people how to act and predicts that they will act in the way it tells them to. To the extent that it fails to give unambiguous prescriptions, it is indeterminate. To the extent that people fail to follow its prescriptions – to the extent, that is, that they behave irrationally – the theory is

[1] The Popperian view that it is better to predict strongly than weakly, because strong predictions are more likely to be falsified and therefore more surprising if not falsified, is quite consistent with this assertion. Popper was concerned with the ex ante choice of research strategy, whereas I am here discussing ex post properties of predictions.

inadequate. In this book as a whole, the emphasis is on the indeterminacy of rational-choice theory.[2] The inadequacy of the theory is also a constant theme, closely intertwined with that of indeterminacy. I argue, in fact, that *failure to recognize the indeterminacy of rational-choice theory can lead to irrational behaviour.*

Let me sketch, as a foil to the later discussion, how the problems of indeterminacy and inadequacy arise in another normative domain – the theory of distributive justice. My point of departure will be John Rawls's view that theories of distributive justice are constrained by data, namely people's intuitions about particular moral problems.[3] Any theory of justice can be judged, therefore, by the criteria of determinacy and adequacy. A theory of justice is determinate if it allows us to tell, at least in principle, how a given allocation problem should be resolved.[4] It is adequate if its prescriptions about particular cases correspond to our considered intuitions about these cases. In addition, of course, the theory must be independently plausible – that is, correspond to some general moral principle that can be defended in abstraction from particular cases.

Among the major theories of justice, utilitarianism and Robert Nozick's theory are intended to be determinate. Rawls's theory is explicitly not advanced with this intention. It is concerned only with the justice of the basic structure of society,[5] not with justice in particular contexts such as the allocation of scarce medical resources or the selection of soldiers for military service. For such problems, the theory has to be supplemented by local principles (II.5). I believe, although the argument cannot be made here,[6] that none of the major theories is adequate. Each yields consequences that are strongly counterintuitive,[7] even when we allow intuitions to be refined and modified pari passu with the construction of the theory. They stand in need of being replaced rather than supplemented.

[2] In Elster (1983a, 1984), the emphasis is mainly on cases of inadequacy.

[3] Rawls (1971), pp. 19–20.

[4] In practice, there are many obstacles to determinacy. Some theories require information about preferences and productive capacities that may be impossible to collect, if only because people may not find it in their interest to reveal them. Others require irretrievable information about events in the distant past or about hypothetical events that would have ensued if these past events had been different.

[5] Rawls (1971), p. 8.

[6] This is largely a euphemism for 'I don't yet have it'.

[7] Yaari and Bar-Hillel (1984); Frohlich, Oppenheimer and Eavey (1987).

notions

These remarks are also relevant to the methodology of rational choice. Like the theory of justice, the theory of rational choice is constrained at both ends. On the one hand, the notion of rationality has to be independently plausible as a normative account of human behaviour. On the other hand, it has to yield prescriptions about particular cases that fit our preanalytical notions about what is rational in such cases. As in the case of justice, these notions are somewhat elastic. As we construct a theory of what is rational, some intuitions about what is rational in particular contexts may change. In particular, theory may force determinacy on our intuitions in situations where initially they were indeterminate. The theoretical notion of an equilibrium, for instance, can serve as a guide to intuition and action when otherwise we would not know what to think or to do. Other, more recalcitrant intuitions can force us to modify the theory.[8]

In what follows, I first set out the bare bones of rational-choice theory (I.2), including a discussion of whether desires can be rational. I then consider failures of rationality that are due to a lack of determinacy (I.3) and go on to discuss failures due to a lack of adequacy (I.4). In the final section (I.5), I briefly discuss how rationality can be supplemented or replaced by other guides to action. In chapter IV, the same issues are examined with respect to political choices.

I.2 Rational action

As I said, rational-choice theory is first and foremost normative.[9] It tells us what we ought to do in order to achieve our aims as well as possible. It does not, in the standard version, tell us what our aims ought to be. (Some nonstandard comments on this problem are offered later.) From the normative account, we can derive an explanatory theory by assuming that people are rational in the normatively appropriate sense. The privileged, but not exclusive status of this assumption is discussed in I.5

The central explananda of rational-choice theory are *actions.* To explain an action, we must first verify that it stands in an optimizing

[8] For an example, see Elster (1984), p. 121, n.17. A recent debate of similar issues is Binmore (1987).

[9] The following draws heavily on Elster (1983a), ch. 1, and Elster (1986a).

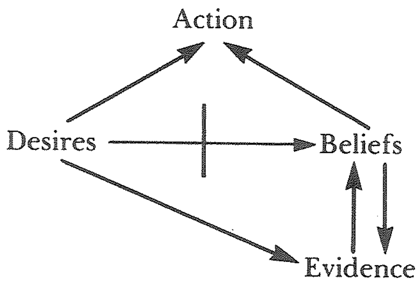

Fig. I.1

relationship to the desires and beliefs of the agent. The action should be the best way of satisfying the agent's desires, given his beliefs. Moreover, we must demand that these desires and beliefs be themselves rational. At the very least, they must be internally consistent. With respect to beliefs, we must also impose a more substantive requirement of rationality: they should be optimally related to the evidence available to the agent. (The substantive rationality of desires is discussed later.) In forming their beliefs, the agents should consider all and only the relevant evidence, with no element being unduly weighted. As a logical extension of this requirement, we also demand that the collection of evidence itself be subject to the canons of rationality. The efficacy of action may be destroyed both by the gathering of too little evidence and by the gathering of too much. The optimal amount of evidence is determined partly by our desires: more important decisions make it rational to collect more evidence. It is determined partly by our prior beliefs about the likely cost, quality and relevance of various types of evidence. Schematically, these relations can be represented as in Fig. I.1.

Rational action, then, involves three optimizing operations: finding the best action, for given beliefs and desires; forming the best-grounded belief, for given evidence; and collecting the right amount of evidence, for given desires and prior beliefs. Here, desires are the unmoved movers, reflecting Hume's dictum that 'reason is, and ought only to be the slave of the passions'.[10] Hume did not mean that reason ought to obey every whim and fancy of the passions. In particular, he would not have endorsed the direct shaping of reason by passion found in wishful thinking, illustrated

[10] Hume (1739), p. 415.

by the blocked arrow in the diagram. To serve his master well, a slave must have some independence of execution: beliefs born of passion serve passion badly.[11]

It follows from this sketch that rational-choice theory can go wrong at three levels, and that in each case the failure may be due either to indeterminacy or to irrationality. There may not exist a uniquely optimal action, belief or amount of evidence. Or people may fail to carry out the action, form the belief or collect the evidence as rationality requires them to do. Such failures of rationality are discussed in I.3 and I.4. Here I want to consider whether one can impose substantive rationality conditions on the desires of the agent.[12] The first idea that comes to mind is that it is rational to have the desires and emotions having which tends to make one happy. The proposal, however, turns out to be flawed.

To have a strong desire for something that is manifestly out of reach can make one desperately unhappy. Sometimes it seems natural to say that desires such as this are irrational. A person with moderate means who is tormented by desires for expensive luxury goods might well be called irrational. But we would not usually say that a person who lives in a totalitarian regime is irrational if he does not get rid of the desire for freedom that makes him deeply miserable.[13] Human beings are more than happiness machines:

> *Example 1. Psychiatric treatment of Soviet dissidents.* The Serbsky Institute for Forensic Psychiatry in Moscow has become notorious for its treatment of political dissidents as mentally ill. 'Some psychiatrists have buttressed their argument about the dissenter's poor adaptation by pointing to the tenacity with which he acts out his beliefs despite the odds. . . . The dissenter does indeed operate in dangerous territory; the reaction of the regime is often harsh. But he is fully aware of the risks inherent in his non-conformist behaviour; his moral integrity compels him to take them. Some dissenters have parried the psychiatrists on this point by asking whether Lenin and his colleagues were "poorly adapted" when, in their struggle against

[11] Veyne (1976), p. 667.

[12] I discuss this question in Elster (1983a, secs. I.4 and III.4). For the closely related question of whether it makes sense to assess emotional reactions as rational or irrational, see Elster (1985a).

[13] We might want to say that, however, if his desire for freedom is caused by the fact that he does not have it. For a brief discussion of such 'counteradaptive preferences' see Elster (1983a), pp. 111–12.

→ implies that rationality is Ō-possib.

the tsarist régime, they were constantly subject to harassment and arrest'.[14]

If anything, it is the conformist – the happy slave – rather than the dissident who appears to be irrational. Unconscious adaptation to the inevitable is a heteronomous mechanism, while rational desires must be autonomous.[15] One cannot be rational if one is the plaything of psychic processes that, unbeknownst to oneself, shape one's desires and values. This preanalytical idea is at least as strong as the intuition that rational desires are desires having which one is happy. Sometimes the two ideas point in the same direction. People who always most want what they cannot get are neither autonomous nor happy. People who adapt to their environment by a process of conscious character planning are both autonomous and happy.[16] At other times, as with the unconscious conformist and the autonomous dissident, the two ideas diverge. Tocqueville captures this ambiguity of conformism when he asks, 'Should I call it a blessing of God, or a last malediction of His anger, this disposition of the soul that makes men insensible to extreme misery and often even gives them a sort of depraved taste for the cause of their afflictions?'[17]

Could one entertain a similar proposal with respect to belief rationality? Could one argue, that is, that it is rational to have the beliefs having which one tends to be happy? In general, we would expect that one's happiness is best promoted by having beliefs which are well grounded in the evidence, since these are by definition the beliefs most likely to be true. Successful action requires correct factual beliefs. Yet in special cases this connection fails. To keep away from dangerous substances it may be necessary to have an exaggerated notion about the dangers of drug abuse.[18] High levels of motivation and achievement often require an unrealistically positive self-image, whereas people with more accurate self-perceptions tend to lose the motivations to go on with the business of living. They are sadder, but wiser:[19]

[14] Bloch and Reddaway (1978), p. 255.

[15] Elster (1983a), ch. 3.

[16] Ibid., pp. 117–19.

[17] Tocqueville (1969), p. 317.

[18] Winston (1980).

[19] Lewinsohn et al. (1980); see also Alloy and Abrahamson (1979) and, for a discussion of their findings, Elster (1985a).

Example 2. Stability of marriage. 'Expectations about divorce are partly self-fulfilling because a higher expected probability of divorce reduces investments in specific capital and thereby raises the actual probability'.[20] 'It is far from clear that a bride and a groom would be well advised to believe, on their wedding day, that the probability of their divorce is as high as .40'.[21] The low-probability expectations of divorce are only partly self-fulfilling. Our misplaced confidence in ourselves motivates us to achievements that make it somewhat less misplaced, but still less than fully justified.

A belief which is unjustified and indeed false may well be instrumentally useful, but it seems odd to call it rational. Rationality, as usually understood, is a variety of intentionality. For something to be rational, it has to be within the scope of conscious, deliberate action or reflection. Useful false beliefs obtain by fluke, not by conscious reflection upon the evidence. Although one cannot in the short run choose one's desires or one's emotional patterns, one can over time shape and bend them to some extent. Beliefs, by contrast, resist manipulation for instrumental purposes. Believing at will, for the sake of the good consequences of having the belief, is a self-defeating enterprise because one cannot – conceptually cannot – at one and the same time believe something and believe that the belief has been adopted on noncognitive grounds.[22] It is easy, therefore, to understand why exhortations to self-esteem, propagated by manuals on self-help therapy, have very limited success.[23]

I.3 Indeterminacy

To explain and predict events or states of affairs, a theory must have determinate implications about what will happen under given initial conditions. Ideally, the implications should be not only determinate, but unique. Among all possible events or states, exactly one should be singled out by the theory. Outside quantum

[20] Becker (1981), p. 224.

[21] Nisbett and Ross (1981), p. 271.

[22] For this argument, see Williams (1973) and Elster (1984), sec. II.3. A recent challenge by Cook (1987) places too much weight on a (hypothetical) example in which the belief adopted at will is also the one that is better grounded in the evidence. A nonhypothetical example of a decision to adopt an unfounded belief would have been more convincing.

[23] Quattrone and Tversky (1986), p. 48.

mechanics, this is the explanatory ideal of science. A theory which does not yield unique predictions is incomplete. It may still, of course, be vastly superior to there being no theory at all. It can be very valuable to know that certain things will *not* happen. Also, for practical purposes it may not matter much which of the events consistent with the theory is actually realized. This said, the prospect of unique prediction dominates and guides scientific work.

In economics, and increasingly in the other social sciences, the neoclassical theory of choice holds out the promise of uniqueness. By its relentless insistence that all behavior is maximizing, it can draw on a basic mathematical theorem which says that every well-behaved function has exactly one maximum in a well-behaved set.[24] Moreover, in economic contexts many functions and sets are well behaved in the relevant sense. For the consumer, there is usually exactly one consumption bundle that maximizes utility within the set of purchases that satisfy his budget constraint. For the producer, there is exactly one combination of the factors of production which maximizes profit per unit of output.

Here I discuss a variety of circumstances under which rational-choice theory fails to yield unique predictions. There may be *several* options which are equally and maximally good. More important, there may be *no* option with the property that it is at least as good as any other.

The problem of multiple optima is, with one notable exception, relatively trivial. It arises when the agent is indifferent between two or more alternatives, each of which is deemed superior to all others. In such cases, rational-choice theory must be supplemented by other approaches to predict which of the equi-optimal alternatives will actually be chosen or 'picked'.[25] If they are very similar to one another, it is not important to be able to make this prediction. Nobody cares which of two apparently identical soup cans on the supermarket shelf is chosen. If the options differ from one another in offsetting ways, as when a consumer is indifferent between two cars with different strengths and weaknesses, the choice is more consequential. The car dealers will certainly care about the choice. I believe, however, that most

[24] Technically, the function must be continuous and the set be compact and convex.
[25] Ullmann-Margalit and Morgenbesser (1977).

cases of this kind are better described by saying that the consumer is unable to rank and compare the options (as discussed later). If he really were indifferent, a reduction of one dollar in the price of one car should induce a clear preference, but I do not believe it usually would.

The exception referred to is game theory, in which multiple optima abound. In noncooperative games with solutions in mixed strategies, it can be shown that an agent will always be indifferent between the strategy prescribed to him by the solution and any other linear combination of the pure strategies that enter into the solution, always assuming that the other players stick to their solution behaviour. John Harsanyi argues that the lack of a good reason for the agent to conform to the solution in such cases is a flaw in game theory as traditionally conceived. In his substitute solution concept, only 'centroid' or equiprobabilistic mixed strategies are allowed. This proposal reflects the idea that when there are several optima, one is chosen at random by 'what amounts to an unconscious chance mechanism inside [the agent's] nervous system'.[26] Here rational choice is supplemented by a purely causal mechanism. I have more to say about randomized strategies in II.3.

Nonexistence of rational choice is a more serious difficulty than nonunicity. The problem arises at all three levels distinguished earlier: when gathering evidence, when deriving beliefs from the given evidence, and when deriving an action from the given beliefs and desires. I shall consider them in the reverse order.

If the agent has an incomplete preference ordering, that is, is unable to compare and rank all the options in his feasible set, there may be no action which is optimal.[27] It would be misleading to say that the agent is irrational: having complete preferences is no part of what it means to be rational. On the contrary, to insist that preferences must be complete and all pairs of alternatives be comparable can be a form of hyperrationality – that is, of irrationality. Other forms of hyperrationality are considered in I.4.

> *Example 3. Choice of career.* 'Life is not long, and too much of it must not pass in idle deliberation how it shall be spent: deliberation,

[26] Harsanyi (1977a), p. 114.

[27] A special and important case is that of moral conflict, discussed in Levi (1986).

> which those who begin it by prudence, and continue it with subtlety, must, after long expence 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'.[28] Suppose that I am about to choose between going to law school or to a school of forestry – a choice not simply of career but of lifestyle. I am attracted to both professions, but I cannot rank and compare them.[29] If I had tried both for a lifetime, I might be 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 on 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.

The nonexistence of an optimal action can also arise because of peculiar features of the feasible set. In planning models with infinite horizons and no time discounting, one can run into the paradox that 'there is always an incentive to postpone consumption: on the other hand postponing consumption for ever is clearly not optimal'.[30] While a theoretical possibility, this problem is not central to actual decision making. By contrast, the difficulties stemming from incomplete preferences are real and important. In addition to the problem of intrapersonal comparisons of welfare referred to in Example 3, the difficulty of making interpersonal comparisons can prevent us from ranking the options, if the ranking takes account of the welfare others derive from them.[31]

At the next level, nonexistence of an optimal belief can arise in two ways: because of uncertainty and because of strategic interaction. 'Uncertainty' here means radical ignorance, the lack of ability to assign numerical probabilities to the possible outcomes associ-

[28] J. Boswell, *The Life of Samuel Johnson,* A.D. 1766 (Aetat 57), a letter from Johnson to Boswell dated 21 August 1766. I owe this reference to John Broome.

[29] If I know myself well, I may be able to predict that whatever I do I shall end up preferring the occupation I choose, or perhaps the one that I do not choose, but this is not to know which choice will make me more happy.

[30] Heal (1973), p. 300.

[31] Sen and Williams (1982), p. 17.

assigning numerical probabilities

ated with the various options. If such assignments are possible, we face a problem of decision making under risk, in which the rational decision rule – to maximize expected utility – can be counted on to yield an optimal choice. Farmers deciding on a crop mix or doctors deciding whether to operate act under risk. They can rely on well-defined probabilities ~~derived from past frequencies~~. Stock market speculators, soldiers and others who have to act in novel situations cannot rely on frequencies. If they have sufficient information and good judgement, they may be able to make good probability estimates to feed into the expected utility calculus. If they have little information or poor judgement, rationality requires them to abstain from forming and acting upon such estimates. To attempt to do so would, for them, be a form of hyperrationality.

> *Example 4. Nuclear waste.* 'Different geological mechanisms may be capable of generating the release of radioactive waste in the environment. Among these are groundwater flow, faulting, diapirism, erosion, fall of meteorites, magma intrusion, and modification of the drainage level of water. An approach to geological confinement is often sought by trying to quantify the probability of occurrence of any of these events and their nuisance value to man. Then, by combining these probabilities and nuisance values, one tries to assess the safety coefficient of the repository and to compare it to the accepted safety coefficients for present risks. This approach does not seem realistic to us *because basically the earth's development has not been a random phenomenon* (possibly apart from the fall of meteorites) and no geologist can seriously give reasonable figures for these probabilities'.[32] Here is a case in which objective probabilities and judgemental, subjective probabilities are equally out of reach.

When the situation is recognized as one of uncertainty, rational-choice theory is limited, but not powerless. Sometimes we are able to dismiss an option in the presence of another that, regardless of which state of the world obtains, has better consequences. Having done this, however, we are often left with several options for each of which there is some state in which it has better consequences than one of the others. Decision theory tells us that in choosing among these we are allowed to take account only of the best and

[32] De Marsily et al. (1977), p. 521. Italics added.

the worst consequences of each action.[33] This may also narrow the field a bit, but often more than one option will be left. In choosing among these, one may adopt the rule of thumb to choose the option with the best worst-consequences (maximin), but there are no grounds for saying that this is more rational than to choose the option with best best-consequences (maximax). To illustrate, consider the following matrix of outcomes as dependent on actions and states:

	S_1	S_2	S_3
A_1	3	4	5
A_2	1	2	8
A_3	2	0	7
A_4	0	0	6

Outcome A_4 can be excluded from consideration since under any state of affairs it yields worse consequences than A_2. Among the remaining, A_3 can also be excluded since both its best consequence and its worst consequence are worse than those of A_2. Of the remaining, maximin reasoning would make us prefer A_1 over A_2, while maximax would lead to the opposite choice. Psychological theories may be able to explain which choice will be made, but rational-choice theory, by itself, is indeterminate.

A special case arises when we have to choose among several scientific theories. Let us assume that each theory assigns numerical probabilities to the events that can occur, but that the choice of one theory rather than another is a matter of uncertainty. Controversies about the effect of CO_2 release in the atmosphere are of this kind. As shown by Jørgen Aasness and Aanund Hylland in unpublished work, this kind of theoretical uncertainty is less devastating than total ignorance, since we can use the content of the theories even if we do not know which of them is correct. If we can assume that one of them is correct, many abstractly possible states of affairs can be excluded and the optimal decision may differ from what it would have been had there been no restrictions on what could happen.

[33] Luce and Raiffa (1957), p. 296; Arrow and Hurwicz (1971). Other proposals are discussed in I.5.

To illustrate the point, assume that we have to choose between acts A_1 and A_2. There are two theories T_1 and T_2. According to T_1, state S_1 occurs with probability 1/3 and state S_2 with probability 2/3. According to T_3, state S_3 is certain to occur. The act–state matrix is the following:

	S_1	S_2	S_3
A_1	12	0	6
A_2	3	15	5

The best and worst outcomes of A_1 are, respectively, 12 and 0; those of A_2, 15 and 3. Define now X as the set of all triples $(\frac{2}{3} \cdot b, \frac{1}{3} \cdot b, 1 - b)$, where b ranges from 0 to 1. X is a subset of the set Y of all triples $(p_1, p_2, 1 - p_1 - p_2)$, with p_1 and p_2 ranging from 0 to 1. Y can be understood as the set of all abstractly possible probability vectors for the states S_i, whereas X is the restricted set that incorporates the information provided by the two theories. The standard theory of choice under uncertainty, defined over the full set Y, says that A_2 should be chosen, since it has a better best-consequence and a better worst-consequence than A_1. On the restricted set, however, A_1 is the best choice. It has a utility of $2b + 6$, whereas A_2 yields only $2b + 5$.

Later chapters provide several illustrations of choice under uncertainty. Chapter II, in particular, surveys a wide range of cases in which one might as well toss a coin to make up one's mind. Chapters III and IV consider in more detail the choice of custodial parent and political reform as examples of decision making under uncertainty.

Consider next strategic interaction as an obstacle to rational-belief formation. Often, rational choice requires beliefs about choices to be made by other people. These beliefs, to be rational, must take account of the fact that these others are similarly forming beliefs about oneself and about each other. Sometimes, these beliefs are indeterminate, when the situation has multiple equilibria with different winners and losers. The games of Chicken and Battle of the Sexes are well-known examples. Each of these games has two equilibria, each of which is better for both players than the worst outcome and preferred by one party to the other equilibrium. In the absence of enforcement or commitment devices, there is no way in which a player can form a rational belief about what the other will do.[34]

> *Example 5. Rational expectations.* To make decisions about consumption and investment, economic agents must form expectations about the future state of the economy. According to an earlier view, these are 'adaptive expectations', or extrapolations from current and past states. This view is unsatisfactory, because it assumes that people react mechanically without using all the information available to them. For instance, following the quadrupling of oil prices in 1973, we would expect expectations to change more radically and rapidly than what would be predicted by the theory of adaptive expectations. The theory of rational expectations, which emerged as a dominant paradigm in the 1970s, assumes that people are forward looking, not backward looking, when forming their expectations and that, moreover, they make the best use of the information available to them. Essentially, people predict the future development of the economy using a correct economic model. Since expectations are part of the model, rational expectations must be self-fulfilling. The problem[35] is that often there are several sets of expectations about the economy that, if held by everybody, would be self-fulfilling. In the absence of government intervention to eliminate some of the equilibria, rational agents will not be able to form mutually supporting, self-fulfilling expectations.

Uncertainty and strategic interaction, taken separately, create problems for rational belief formation. When both are present, they wreak havoc. In planning for war, generals are hindered both by uncertainty about whether their sophisticated systems will work and by strategic complexities. The old dictum – Don't base your plans on the enemy's intentions but on his capabilities – no longer applies, if it ever did, since generals are equally uncertain about the effectiveness of the weapons of the enemy (and about the degree of uncertainty among the generals on the other side).

[34] At least this holds for the symmetric version of these games. With asymmetries, tacit bargaining may lead the parties to converge to the equilibrium that favours the party who is least worried by the prospect of the worst outcome. The weak may accept a legal regime that favours the strong because, unlike the strong, they cannot survive in the state of nature.

[35] Actually, one of the many problems that beset rational-expectations theory. For a survey, see Begg (1982), pp. 61–70.

Example 6. Explaining investment. 'The outstanding fact is the extreme precariousness of the basis of knowledge on which our estimates of prospective yield will have to be made. Our knowledge of the factors which will govern the yield of an investment some years hence is usually very slight and often negligible. If we speak frankly, we have to admit that our basis of knowledge for estimating the yield ten years hence of a railway, a copper mine, a textile factor, the goodwill of a patent medicine, an Atlantic liner, a building in the city of London amounts to little and sometimes to nothing; or even five years hence'.[36] For the special case of investment in research and innovation, this lack of foreknowledge decomposes into the elements of uncertainty and strategic interaction. On the one hand, the outcome of innovative activities is inherently uncertain. One may strike gold, or find nothing. As Humphrey Lyttelton is reported to have said, 'If I knew where jazz was going I'd be there already'. But suppose one could know how the chance of finding gold is related to the amount one has invested. Under the 'winner-take-all' system of modern industry, it also matters whether one finds it before others do. If other firms invest massively, the chances that a given firm will be first past the post may be too small to make the investment worth while. If other firms do not invest, the chances are much higher. But if it is true of each firm that it should invest if and only if others do not, it has no basis for anticipating what others will do.[37] Entrepreneurs might as well follow Keynes's advice and be guided by their 'animal spirits'.

Finally, determinacy problems arise with respect to the optimal amount of information one should collect before forming an opinion. Information is useful, but costly to acquire. Ideally, the rational agent would strike a balance between these two considerations: he would acquire information up to the point at which the marginal cost of acquiring information equaled its expected marginal value. In some areas of decision making these calculations can be carried out with great accuracy. Thus 'To detect intestinal cancer, it has become common to perform a series of six inexpensive tests ('guaiacs') on a person's stool. The benefits of the first two tests are significant. However, when calculations are done for each of the last four tests to determine the costs of detecting a case of cancer (not even curing it), the costs are discovered to be

[36] Keynes (1936), pp. 149–50.
[37] Dasgupta and Stiglitz (1980).

precise probabilities

$49 150, $469 534, $4 724 695 and $47 107 214, respectively. To some these calculations suggest that the routine should be reduced, say to a three-guaiac test'.[38]

Sometimes it is impossible to estimate the marginal ~~cost and~~ benefit of information. Consider a general in the midst of battle who does not know the exact disposition of the enemy troops. The value of more information, while potentially great, cannot be ascertained. Determining the expected value would require a highly implausible ability to form numerical probability estimates concerning the possible enemy positions. (Indeterminacy of rational belief due to strategic interaction is important here.) The costs of acquiring information are equally elusive. The opportunity costs might be enormous, if the time spent gathering information offered the enemy a chance to attack or to prepare his defence, or they might be quite trivial. Under such circumstances, one might as well follow Napoleon's maxim 'On s'engage et puis on voit'.

In between these extremes – medical diagnosis and the conduct of battle – fall most everyday situations. The observations that a rational person should make 'greater investment in information when undertaking major than minor decisions',[39] while true, does not help him to decide *how* much to invest. That decision requires estimates about the probable costs and benefits of the search for information. Search theories of unemployment, for instance, assume that the unemployed worker knows the distribution of job offers or at least the general shape of the distribution. Using this knowledge, he can calculate the optimal time spent searching for well-paid jobs. This argument is of dubious value. The doctor ?! carrying out a medical diagnosis finds himself *many* times in the *same situation.* Most persons are unemployed only once or, if more than once, under widely diverging circumstances. They have no way of learning by experience what the job distribution looks like. To be sure, they know something about the job market, but there is no reason to think that they can piece together their bits of information to a reliable subjective distribution.[40] Similar argu-

[38] Menzel (1983), p. 6. The marginal value of the information is controversial, since it depends on an assessment of the value of life.

[39] Becker (1976), p. 7.

[40] On the general point, see Tversky and Kahneman (1974); Lichtenstein, Fischhoff and Phillips (1982). For a devastating criticism of optimal search theories, see Hey (1981).

ments apply to many consumer decisions, like the purchase of a car or an apartment. People know that it makes sense to spend some time searching and that it would be pointless to search forever, but between these lower and upper limits there is usually an interval of indeterminacy.

I.4 Irrationality

In this section I survey some main varieties of irrationality, including, as a special case, hyperrationality. The latter notion is defined as *the failure to recognize the failure* of rational-choice theory to yield unique prescriptions or predictions. As in Kant's critique, the first task of reason is to recognize its own limitations and draw the boundaries within which it can operate. The irrational belief in the omnipotence of reason is what I call hyperrationality. Later chapters provide numerous illustrations.

Failures to conform to well-defined prescriptions of rational-choice theory arise at all three levels distinguished in Fig. I.1. Consider first how actions can fail to relate optimally to given desires and beliefs. The paradigm case is weakness of will, characterized by the following features. (a) There is a prima facie judgement that *X* is good. (b) There is a prima facie judgement that *Y* is good. (c) There is an all-things-considered judgement that *X* is better than *Y*. (d) There is the fact that *Y* is chosen. Often, *X* is an act that is in the long-term interest of a person or corresponds to his moral will, whereas *Y* is a short-term impulse or a self-interested desire. There is no conceptual link, however, between weakness of will, myopia and selfishness.[41]

> *Example 7. Neo-Freudianism.* Freud depicted two forms of human irrationality: being under the sway of the pleasure-seeking id and being dominated by the rigid, compulsive superego. As bearer of the rational will, the ego is engaged in a two-front war against these two enemies. The nature of the id, ego and superego in Freud's theory is somewhat unclear. Are they separate homunculi, each with a will of its own and capable of engaging in strategic interaction with the others?[42] Or, more soberly, are they conflicting tendencies

[41] Elster (1985b).

[42] Kolm (1980), pp. 302–11.

of one and the same subject? In his recent reinterpretation of Freud's trichotomy, George Ainslie has clarified the matter.[43] The ego's struggle with the id is interpreted in terms of time preference functions with the peculiar feature that a larger delayed reward which is preferred to a small early reward when they are both in the distant future becomes less preferred when the time for choice approaches. (Think of a person who makes an appointment with his dentist and then cancels it the day before.) To avoid such weak-willed behaviour, the ego can ally itself with the future, for if the situation can be expected to recur, *bunching* of all the small rewards and of all the large rewards makes it easier to choose the latter. By its rigid, uncompromising character, however, bunching may be as crippling to rationality as the problem it was supposed to resolve. If the ego abdicates its will to get rid of the id, it substitutes one form of weakness of will for another. Even when the person sees that it makes sense to give himself a break, he cannot bring himself to do so.

There is another set of cases in which desires and beliefs can fail to bring about the end for which they provide reasons. They have been referred to as 'excess of will',[44] although they are not in any sense the contrary of weakness of will. Assume that if I do *X*, I shall bring about *Y*, which is what I most desire. Moreover, I am able to do *X*, in the straightforward sense in which I am able to raise my arm. The snag, however, is that *X* will bring about *Y* only if I do *X* without the intention to bring about *Y*. Doing *X* for the purpose of bringing about *Y* will not succeed. Examples of *X* and *Y* could be: drinking hot tea at bedtime and falling asleep; working hard and forgetting a humiliating experience; looking at erotic pictures and becoming sexually aroused; joining a political movement and achieving self-respect.[45] Further examples are discussed at more length below and in chapter IV.

It might appear that someone who does *X* to achieve *Y* acts rationally. He is doing what he believes is (let us assume) the best way of getting what he most desires. This would be true if the situation conformed to the standard scheme of action, depicted in Fig. I.2.

The scheme goes from beliefs and desires through the intention

[43] Ainslie (1982, 1984, 1986).

[44] Farber (1976).

[45] These examples and many others are extensively discussed in Elster (1983a), ch. 2.

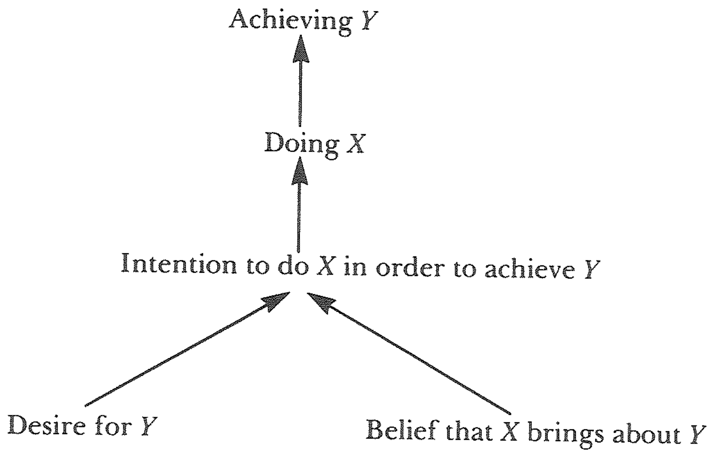
Achieving *Y*
Doing *X*
Intention to do *X* in order to achieve *Y*
Desire for *Y*
Belief that *X* brings about *Y*

Fig. I.2

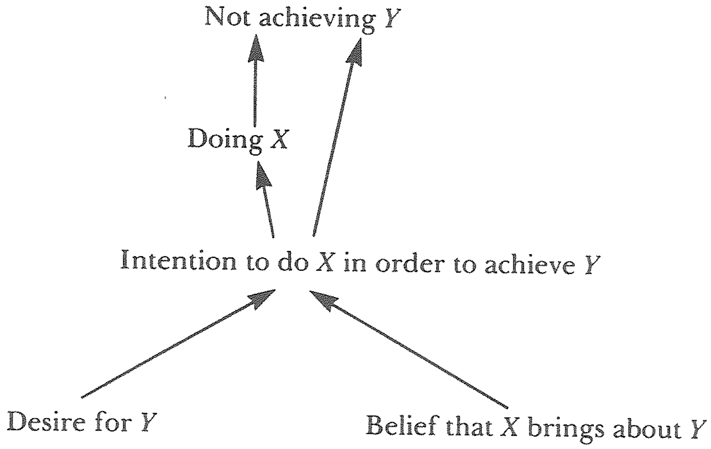
Not achieving *Y*
Doing *X*
Intention to do *X* in order to achieve *Y*
Desire for *Y*
Belief that *X* brings about *Y*

Fig. I.3

to the action and finally to the outcome of the action. There is no guarantee, of course, that the intended outcome will occur. The belief that *X* brings about *Y* could be mistaken. Extraneous factors might intervene. Actions that fail to bring about their intended outcomes for such reasons are not irrational. They fail, as it were, honourably. Matters stand differently when the failure is intrinsic to the action, as when the very intention to do *X* for the sake of *Y* interferes with the efficacy of *X* to bring about *Y*. This nonstandard scheme is shown in Fig. I.3.

Example 8. Don't wait for return of husband. This was the heading of a column in the *Miami Herald* (April 1987), in which Howard Halpern, a psychologist in private practice in New York City, answered

> the following question: 'I am a 57-year-old woman whose husband of 36 years has decided to live alone. We've sold our house and are living in separate dwellings. He speaks of "hope" and "working things out", while happily living the single life. I am unable to get on with my life in such an independent manner. We've had a great deal of joint and individual therapy, but it has not restored our relationship. We've lived together for a few months in the past two years. Each time I thought we would get back together, but then he would leave again. Is there something I should be doing besides waiting?' Mr Halpern answered, 'When you use the word "waiting", I get the impression that you have put your life on hold until your husband's hoped-for return. It is time to stop waiting. By that I don't mean you should make your separation legal – I'm not suggesting any action in particular. I think you must accept your situation as real, understand that your husband may not return and refuse to let your life be dependent on his decision. You have already made efforts to get him to return. Now you must pay attention to your own life and outline your own goals. Focusing on yourself may make you more appealing to him, but that is not the reason to do it. You must do it for yourself'. It is hard to think of advice that would be more misguided. The remark that focusing on herself might make her more appealing to him, while obviously intended to motivate her efforts, is sure to ruin their effect.

Consider next the varieties of irrationality that arise at the level of beliefs and desires. These can be subverted and distorted by causal forces in two main ways: by drives and motivations or by inadequate cognitive processing. Since the end result can be a motivational state (a desire) or a cognitive one (a belief), we have four categories, which I now proceed to illustrate.

The motivational basis of motivations

By this phrase, I do not have in mind conscious character planning, the shaping of preferences by metapreferences. Rather it refers to nonconscious motivational mechanisms that shape our desires 'behind our back'. The best known is what Festinger called 'cognitive dissonance reduction', the natural tendency of the mind to rearrange its desires and beliefs so as to reduce the tension created by high valuations of objects believed to be unattainable or low valuations of objects believed to be inescapable.[46] Also, being faced, like Buridan's ass, with two objects that appear equally attractive creates a form of cognitive dissonance extensively discussed in chapter II.

Some applications of dissonance theory focus on the adjustments of beliefs, while others emphasize the motivated change in evaluations.

> *Example 9. The Hungarian black market.*[47] One mechanism of dissonance reduction is 'I paid a lot for it, so it must be good'. A Hungarian coffee shop begins to offer high-quality coffee to customers who are willing to pay a bit extra. Since the shop has a limited quota of coffee beans, each customer who pays the high price creates an externality for the customers who pay the official price. The official cups of coffee being increasingly diluted, more and more customers are willing to pay the premium. Yet, as more and more do so, the quality of black-market coffee approaches the initial quality of the ordinary coffee. In the end, everybody pays the higher price for coffee of ordinary quality. It would appear, therefore, that everybody has lost, in a standard *n*-person Prisoner's Dilemma. The twist to the story is that because of cognitive dissonance nobody experiences any subjective loss. Since everyone is paying more for the coffee, it must be better than it used to be. The Prisoner's Dilemma yields a Pareto improvement: the shop keeper gains more and the customers are happy.

It is not obvious that desires shaped by dissonance reduction are, ipso facto, irrational. They do, after all, make people happier. Desires shaped by dissonance-increasing mechanisms are more obviously irrational. Many people, for instance, have a preference for novelty that gets them into trouble:

> *Example 10. What father does is always right.* In H. C. Andersen's story of this name, a farmer goes to the market in the morning to sell or exchange his horse. First, he meets a man with a cow, which he likes so much that he exchanges it for the horse. In successive transactions, the cow is then exchanged for a sheep, the sheep for a goose,

[46] Festinger (1957, 1964); Wicklund and Brehm (1976). Economic applications include Akerlof and Dickens (1982) and Schlicht (1984).

[47] The example draws upon Galasi and Kertesi (1987).

the goose for a hen and, finally, the hen for a sack of rotten apples. The farmer's road to ruin is paved with stepwise improvements.[48] (Actually he is not ruined, because a pair of English tourists make and lose a bet that his wife will be angry with him when he comes back with the apples.) Although the story does not say so, it is likely that the farmer would have refused to exchange his horse for a sack full of rotten apples. Curiosity and the thirst for novelty are triggered by options which are neither too similar nor too dissimilar from the current state.[49] In Johannes V. Jensen's story of the same name – a take-off on Andersen's classic tale – the farmer goes to the market with a set of rotten apples.[50] By a series of lucky accidents, he comes back with a horse. When he tells his wife about the deals, she manages to see each of them in an unfavourable light. Although the story is not fully clear on this point, it appears that she even thought a horse for a sack of apples a bad deal. Thus her perverse attitude can probably be explained by her belief that her husband cannot do anything right, not by an inherent conservatism that would be the converse of a preference for novelty. If the latter was the case, she would probably prefer the end state over the initial state, while being opposed to each of the intermediate steps.

The motivational basis of cognitions

Dissonance reduction can also take the form of belief adjustment. Workers who take jobs in unsafe industries alter their estimated probabilities of accidents.[51] As a result, when safety equipment becomes available, they may choose not to purchase it. Here, as in other cases, misformation of private beliefs (or preferences) creates a case for government intervention.[52] In addition to direct motivational interference with the cognition, there can be indirect interference with the evidence on which cognition is based. People who dread having a dangerous disease put off seeing the doctor.

[48] von Weizäcker (1971) offers a formal model of this process.

[49] Middleton (1986).

[50] I am indebted to Hilde Sejersted for bringing this story to my attention.

[51] This example is taken from Akerlof and Dickens (1982). This otherwise excellent article is marred by the idea that people can choose their own beliefs so that, for instance, they can weigh the psychic benefits of believing that their job is safe against the cost of increased chances of accidents. Although I am sure that both the costs and the benefits of dissonance reduction influence the extent to which it occurs, I do not think they do so by virtue of conscious comparison since, as argued above, beliefs cannot be deliberately chosen.

[52] Sunstein (1986) has a general discussion of such cases.

People who fear they might be gaining weight avoid stepping on the scales.

Belief-oriented dissonance reduction is a form of wishful thinking. To the extent that it makes one feel happy, it might be thought to be a good thing. Usually, however, the pleasure of wishful thinking is of brief duration, like the warmth provided by pissing in one's pants. *Acting* on beliefs formed in this way can be disastrous and is likely to force a change in the beliefs. When action is not called for, the wishful beliefs can be more stable. The 'just-world' theory, for instance, suggests that people adjust their beliefs about guilt and responsibility so as to preserve their belief that the world is fundamentally just.[53] The best-known example is the 'blame the victim' syndrome, further discussed in II.8. While it would be perverse to say that blaming the victim is rational, it can certainly contribute to one's peace of mind. Some forms of motivated belief formation do not even have that effect. The congenital pessimist, who systematically believes that the world is as he would *not* like it to be, creates dissonance instead of reducing it. Dissonance reduction, while a threat to autonomy and rationality, is at least intelligible in terms of the 'wirings of the pleasure machine', as Amos Tversky has put it. Dissonance production indicates that the wires have been crossed and that something is radically wrong.

The cognitive basis of motivations

Under this heading fall the violations of expected utility theory that have been extensively studied over the past decade or so.[54] An important example is 'framing', that is, preference reversal induced by redescription of the choice situation.[55] People who would abstain from buying credit cards if firms impose a 3 per cent surcharge on card users may be less deterred if firms offer a 3 per cent discount on cash purchases.[56] Time preferences can be manipulated by presenting the difference between present and future consumption as a delay premium or as a speed-up cost.[57] These are

[53] Lerner and Miller (1978).

[54] A recent summary is Machina (1987).

[55] Tversky and Kahneman (1981).

[56] Thaler (1980).

[57] Loewenstein (1987).

examples in which the reference points or frames are imposed from the outside. A more intriguing problem arises if we ask about the principles that regulate spontaneous choice of frames.[58] It has been suggested that people choose the frame that induces the choice that makes them happy,[59] but it is far from obvious that nonconscious motivational mechanisms are capable of operating in this indirect manner.

Another set of deviations from expected utility theory arises because people do not treat known probabilities as the theory tells them to. (The problem comes on the top of their difficulties in estimating unknown probabilities.) Thus 'low probabilities are overweighted, moderate and high probabilities are underweighted, and the latter effect is more pronounced than the former'.[60] In other words, people exaggerate the difference between impossible events and low-probability events and, especially, between near-certain and certain events. Attitudes towards nuclear accidents and other great disasters may, for this reason, include elements of irrationality. The point is *not* that it is irrational to feel anxiety at the prospect of a low-probability nuclear accident. What is irrational is that this attitude, when combined with other attitudes that may also appear unobjectionable in isolation, can be made to yield inconsistent choices. 'It is not easy to determine whether people value the elimination of risk too much or the reduction of risk too little. The contrasting attitudes to the two [logically equivalent] forms of protective action, however, are difficult to justify on normative grounds'.[61]

The cognitive basis of cognitions

There is by now a massive body of evidence showing how belief formation can fail because people rely on misleading heuristic principles or, more simply, ignore basic facts about statistical inference.[62] Securities and futures markets seem excessively sensitive

[58] Fischhoff (1983).
[59] Machina (1987), p. 146.
[60] Tversky and Kahneman (1981), p. 454.
[61] Ibid., p. 456.
[62] Good summaries are Nisbett and Ross (1981) and Kahneman, Slovic and Tversky, eds. (1982).

to current information.[63] Baseball trainers who notice that last season's star is not living up to his past performance are rapid to conclude that he has been spoilt by success, ignoring the statistical principle that, on the average, an outstanding performance is likely to be followed by one closer to average ('regression to the mean').[64] 'Labeling' theorists of mental illness cite as evidence for their theory the fact that the longer people have been in mental hospitals, the less likely they are to get well, ignoring the alternative explanation that the probabilities of getting well may differ across people but be constant over time.[65]

Example 11. Calvinism. The previous two examples turn upon a confusion between causal and noncausal interpretation of the facts. Max Weber's interpretation of the affinity between Calvinism and economic activity invokes a similar tendency to infuse diagnostic facts with causal value. 'Thus, however useless good works might be as a means of attaining salvation, for even the elect remain beings of the flesh, and everything they do falls infinitely short of divine standards, nevertheless, they are indispensable as a sign of election. They are the technical means, not of purchasing salvation, but of getting rid of the fear of damnation'.[66] It has been argued that the mechanism invoked here is motivational, a form of dissonance reduction.[67] It could, however, be a purely cognitive tendency to confuse diagnostic and causal efficacy. When people ask themselves, 'If not now, when?' and 'If not me, who?' they commit similar fallacies, albeit very useful ones.[68] People who open only one box in Newcomb's Problem do the same.[69]

I conclude this section with a few remarks about hyperrationality. Since the concept is discussed extensively in later chapters, especially in II.3 and II.8, I content myself here with a brief

[63] Arrow (1982).

[64] Nisbett and Ross (1981), p. 164, referring to 'the sophomore slump'.

[65] Gullestad and Tschudi (1982).

[66] Weber (1958), p. 115.

[67] Barry (1978), p. 41.

[68] See Quattrone and Tversky (1986) for the latter fallacy and Elster (1985b) for a discussion of the former.

[69] For exposition and discussion of this problem, see the articles collected in Campbell and Sowden, eds. (1985). A perfect illustration is a circular letter issued by English Baptists around 1770: 'Every soul that comes to Christ to be saved . . . is to be encouraged. . . . The coming soul need not fear that he is not elected, for none but such would be willing to come' (Thompson 1968, p. 38).

enumeration of some main varieties. (a) Sometimes people attempt to eliminate uncertainty of beliefs or incompleteness of preferences, although the choice situation is essentially indeterminate. It is always possible to devise questions that will force a person to reveal his preferences or subjective probabilities, but often there is no reason to believe in the robustness of the results. If the outcome depends on the procedures of elicitation, there is nothing 'out there' which is captured by the questions. (b) Sometimes people look to the second decimal when they are ignorant about the first. In some contexts, this method of problem solving is as good as any other. In others, it can be very wasteful, if people differ in their assessment of the second decimal and spend resources arguing about it. (c) Sometimes people will reframe an indeterminate decision problem so as to make it appear determinate. If one option stands out along one dimension, that dimension may take on increased importance so as to make the choice an easier one. (d) Sometimes people seek out what is rational to do in any given situation instead of looking for more general rules that cover many similar cases. Focusing on rules rather than acts can economize on costs of decision (see chapter III) and also have superior incentive effects.[70] (e) Sometimes people ignore the costs of decision making. They search for the solution that would have been best if found instantaneously and costlessly, ignoring the fact that the search itself has costs that may detract from optimality.

I.5 Alternatives to rationality

In light of earlier sections, several questions arise. How serious are these failures of rational-choice theory? Is there any reason to think that the theory has a privileged status in the study of human action? What are the alternative accounts that could supplement or replace the theory?

The failures of indeterminacy appear to me to be quite serious. One way of assessing the power of the theory is to distinguish choice situations by two criteria: the importance of the problem and the number of agents involved. 'Small' problems, that is, problems in which the options do not differ much in value from one

[70] In addition, focusing on rules can protect one against weakness of will.

another, do not lend themselves to the rational approach. Either the options are equally good or it is not clear that it would pay to find out which is the better, or pay to find out whether it will pay to find out. 'Large' problems,[71] in which the choice can be expected to have wide-ranging consequences, also tend to fall outside the scope of the theory. Preference rankings over big chunks of life tend to be incomplete, and subjective probabilities over events in the distant future tend to be unreliable. The theory is more powerful when applied to medium-sized problems like the purchase of a car or a house, but even here the question of optimal search is largely indeterminate.

Other things being equal, decision problems with one agent or with many agents are more likely to yield determinate solutions than problems with a small number of agents. By definition, one-agent problems have no strategic indeterminacy. With many sellers and many buyers, competition forces a unique set of equilibrium prices. With one seller and one buyer, there is often a large range of mutually acceptable outcomes and much indeterminacy concerning which outcome will be realized.[72] A rough conclusion is that rational-choice theory is applicable mainly to one-agent and many-agent problems of intermediate size. Although precise quantification is impossible, indeterminacy is not a marginal problem that can be assimilated to 'friction' or 'noise'.

The factual importance of irrationality does not lend itself to a similarly systematic analysis. The central issue is whether people deal irrationally with important problems. The issue cannot be studied experimentally, since limitations on funding rarely allow stakes to be high enough and subjects to be numerous enough to get reliable results.[73] Introspection, casual observation, historical

[71] See Ullmann-Margalit (1985) for an analysis of 'big decisions' which nicely complements the analysis of 'small decisions' in Ullmann-Margalit and Morgenbesser (1977).

[72] Although noncooperative bargaining theory has done much to force determinacy in such problems (Rubinstein 1982), it is mainly of use in two-person contexts. Three-person bargaining problems remain largely indeterminate even in the noncooperative approach (Sutton 1986).

[73] To get around this problem, it has been suggested that in Third World countries experiments be conducted in which five- or ten-dollar rewards represent high stakes. To get around any ethical problems, all subjects could receive the maximal reward when the experiment was completed, even though told beforehand that they would get it only if they performed well. For reasons explained in Barry (1986), severe ethical problems would still remain.

case studies and novels suggest that irrationality is quite widespread. Drug abuse is perhaps the most striking evidence. More generally, the widespread inability to be properly swayed by future consequences of present action points to a serious deficit in rationality.[74] Studies of 'group think'[75] suggest that political and military decisions are often made in disregard of the evidence. The motivated ignorance of the Holocaust is a massive example of irrational belief formation.[76] The vast sales of self-therapy manuals suggest that many people believe that they can talk themselves into self-confidence and self-respect. I could go on enumerating cases, but they would not add much to the general idea. Irrationality is neither marginal nor omnipresent.

Although indeterminacy and irrationality are widespread, they do not affect the normative privilege of rationality. First and foremost, rationality is privileged because we want to be rational.[77] We take little pride in our occasional or frequent irrationality, although sometimes it has to be accepted as the price we pay for other things we value. In our dealings with people, we are compelled to treat them as, by and large, rational. Communication and discussion rest on the tacit premise that each interlocutor believes in the rationality of the others, since otherwise there would be no point to the exchange.[78] To understand others, we must assume that, by and large, they have rational desires and beliefs and act rationally upon them. If a person says that he wants *X* and yet deliberately refrains from using the means that he knows to be the most conducive to *X*, we usually conclude not that he is irrational but that he does not really want *X*. Sometimes, of course, we may conclude that irrationality offers the best explanation of a given kind of behaviour, but even then most of the evidence about the agent that goes into that conclusion is formed on the assumption that he is, by and large, rational.[79]

[74] Against those who say that discounting the future only shows a 'taste for the present' and that *de gustibus non est disputandum,* I would reply, first, that much time discounting is inconsistent (Elster 1985b) and, second, that even consistent time discounting beyond what is justified by mortality tables is a failure of rationality.

[75] Janis (1972).

[76] Laqueur (1980).

[77] Føllesdal (1982).

[78] Midgaard (1980); Habermas (1982).

[79] Davidson (1980).

"satisficing"

The explanatory privilege of rationality rests on two grounds. As just observed, rationality is presupposed by any competing theory of motivation, whereas rationality itself does not presuppose anything else. On grounds of parsimony, therefore, we should begin by assuming nothing but rationality.[80] Also, while rationality may have its problems, the opposition is in even worse shape. The dictum that you cannot beat something with nothing applies here, with some modifications. As will be clear from what I shall say about the alternatives to rational-choice theory, they are more than nothing, but they do not quite amount to something either.

Herbert Simon's theory of satisficing is intended to supplement rational-choice theory when it is indeterminate.[81] It has been applied to technical change,[82] consumer choice,[83] and numerous other problems. The strength and main weakness of the theory are its realism. On the one hand, it is true and important that many people are happy once their aspiration level has been reached. They stop searching once they have found something that is good enough. On the other hand, there is to my knowledge no robust explanation of why people have the aspiration levels they do, nor of why they use the particular search rules they do. The theory describes behaviour, but does not really explain it. Now, one might say that a similar criticism applies to rational-choice theory, which does not, after all, explain why people have the preferences they do. The hypothesis that people behave rationally is nevertheless simpler, more general and more powerful than the assumption that they are guided by their aspiration levels. In the theory of the firm, for instance, rational-choice theory needs only one assumption, namely that the firm maximizes profits. Satificing theory needs many assumptions, stipulating one aspiration level for each

[80] The situation is somewhat similar to the privileged status of the assumption of selfishness. We can consistently imagine a world in which everybody behaves selfishly all the time, but not a world in which everybody behaves altruistically all the time, because altruism presupposes some nonaltruistic pleasures that the altruist can promote.

[81] I disregard the interpretation of satisficing as maximizing under constraints on information-processing capacities. Limited calculating ability is only one obstacle to first-best rationality. A more important obstacle, in my view, is our inherently limited knowledge about the value of information. Also, people with severely limited cognitive capacities may not be able to understand their limits and hence are not, subjectively, constrained by them.

[82] Nelson and Winter (1982).

[83] Hey (1981, 1982).

of the many subroutines of the firm and, when that level is not attained, one search mechanism for each routine.

Simon's theory, and other theories in the same vein,[84] are intended to *supplement* rational-choice theory, both as a guide to and as an explanation of action. They are rarely intended to *replace* the rationality assumption. Proponents of these alternatives usually grant that rational-choice theory has substantial explanatory power in the absence of uncertainty, but add that most real-life decision making is characterized by a high degree of uncertainty that is costly or impossible to resolve. This is also the point of departure of the theory offered by Isaac Levi to guide and explain decision making under value conflicts and uncertainty.[85] Under conditions of unresolved value conflict, he recommends that we use lexicographically secondary values to decide among the options that are 'admissible' according to the primary, conflicting values, an admissible option being one that is optimal according to one of these values (or to some weighted average of them). Under conditions of uncertainty, he similarly recommends the use of *security* and *deferability* to supplement the expected-value criterion. Levi also argues that many apparent violations of rationality can be understood by assuming that the agents are acting in accordance with his prescriptions. Their choices reflect reasonable ways of coping with unresolved value conflicts and uncertainty rather than cognitive illusions of the kind discussed above.[86] Levi does not try, however, to account for all the apparent violations of expected utility theory.

Other theories, offered squarely as alternatives to rational-choice theory, aim to explain what they admit to be violations of rationality. They can be classified, very roughly, into psychological, biological and sociological alternatives to the economic approach to behaviour.

Psychological theories attempt to explain the observed viola-

[84] Notably Heiner (1983, 1988).

[85] Levi (1974, 1986).

[86] Levi (1986), p. 33, shows that a perfectly sensible way of handling unresolved value conflicts can lead to violation of Sen's 'property alpha', which says that if *a* is chosen in the set (*a*, *b*), *b* should never be chosen in a larger set (*a*, *b*, *c*). Similarly he argues (Levi 1986, ch. 7) that the Ellsberg and Allais paradoxes of choice under risk can be handled without imputing irrationality to the agents who make these apparently inconsistent decisions.

tions of expected utility theory referred to earlier by providing an account that (a) is simple and intuitively plausible, (b) explains all observed deviations from expected utility theory and (c) predicts no unobserved deviations. Attempts to achieve this goal include prospect theory,[87] generalized expected utility theory[88] and regret theory.[89] This is a field where nonexperts should tread warily, and I abstain from evaluating the various proposals, beyond making the presumably uncontroversial remark that only prospect theory appears to be capable of explaining framing phenomena. I note later, however, an apparent example of irrationality through framing that is more plausibly explained by a sociological alternative to rational-choice theory.

Biological alternatives take off from findings about animal behaviour. Animals can be constrained to choose between two responses, each of which has a particular reward schedule. In variable-ratio (VR) schedules we set up a constant probability of reward for each response. The one-armed bandit of the Las Vegas variety illustrates this reward schedule. It is a mechanism with no memory: if we hit the jackpot on one occasion, we are just as likely to hit it again the next time. In variable-interval (VI) schedules we set up a mechanism with memory, so that each unrewarded response increases the probability that the next response will be rewarded. In each period the experimenter uses a chance device, with constant probabilities, to decide whether food is to be made available. Once it has been made available, it stays available. The animal does not know, however, whether it is available. To find out, and to get the food, it must make the appropriate response.

The central question is whether animals allocate their attention optimally between the two responses, that is, whether they act to maximize their rewards. Faced with the choice between two VR schedules, animals often do the rational thing and allocate all their attention to the response with the highest probability of reward. Sometimes, however, they commit 'gambler's fallacy' of distributing the stakes in proportion to the odds. With two VI schedules, the findings are also ambiguous. In a VI–VR schedule, animals usually do not optimize. Instead of equalizing the marginal return

[87] Kahneman and Tversky (1979); Tversky and Kahneman (1987).

[88] Machina (1983).

[89] Loomes and Sugden (1982).

of the two responses, as rationality would require them to do, they equalize the average return. They forget, as it were, that most of the VI rewards come from a few responses, and that it is not really profitable to pay attention to this schedule beyond visiting it from time to time to collect any reward that might have come due after its last visit.

Richard Herrnstein argues that the principle of equalizing average returns ('the matching law') is a more fundamental principle than utility maximization.[90] In addition to explaining allocation of behaviour across schedules, it can explain the allocation over time. Specifically, the matching law predicts that time discounting will be steeper than the exponential discount functions usually stipulated by economists. Although the empirical verdict is not yet in, there is evidence that much animal and human discounting is nonexponential.[91] On the other hand, the matching law explains only the most naïve forms of human behaviour. People can use conscious thought processes to analyse the structure of the choice situation. Unlike animals, they are not restricted to myopic learning. The matching law may describe 'prerational' behaviour, but it is powerless to explain more sophisticated choice processes.

A sociological alternative to the economic approach is the theory of social norms.[92] I define social norms mainly by their non-outcome-oriented character. Whereas rationality tells people, 'If you want *Y*, do *X*', many social norms simply say, 'Do *X*'. Some social norms are hypothetical, but they make the action contingent on past behaviour of oneself or others, not on future goals. These norms say, 'If others do *Y*, do *X*' or 'If you have done *Y*, do *X*'. The norms are *social* if they satisfy two further conditions: they are shared with other members of the community and they are in part enforced by sanctions provided by others.

Here are some examples of social norms, chosen with a view to the contrast with rational action: (a) the norm of voting is very strong in Western democracies. It accounts for most voting in

[90] Herrnstein and Vaughan (1980); Vaughan and Herrnstein (1987); Herrnstein (1988). A cautiously optimizing approach to animal behaviour is that of Staddon (1983, 1987).

[91] Ainslie (1975, 1982, 1984, 1986).

[92] I discuss this theory at some length in a companion volume (Elster 1989) to the present book, and the following account must be read only as a sketch of that more extended argument.

national elections.[93] Selfish voters have virtually nothing to gain from voting, while the costs are non-negligible. Altruistic voters might find voting rational, were it not for problems of strategic interaction. Altruistic voting is a game with multiple equilibria, in each of which most but not all voters go to the polls.[94] (b) The norm of vengeance practised in many traditional societies is triggered by an earlier offence, not motivated by future rewards. Indeed, from the future-oriented point of view vengeance is pointless at best, suicidal at worst. (c) In most Western societies there is a norm against walking up to someone in a cinema queue and asking to buy his place. The norm is puzzling, as nobody would lose and some could gain from the transaction. (d) Norms of dress and etiquette do not seem to serve any ulterior purpose, unlike for instance traffic rules that serve to prevent accidents. wrong

Consider finally an example that could be explained both in terms of framing and in terms of social norms. Consider a suburban community where all houses have small lawns of the same size.[95] Suppose a houseowner is willing to pay his neighbour's son ten dollars to mow his lawn, but not more. He would rather spend half an hour mowing the lawn himself than pay eleven dollars to have someone else do it. Imagine now that the same person is offered twenty dollars to mow the lawn of another neighbour. It is easy to imagine that he would refuse, probably with some indignation. But this has an appearance of irrationality. By turning down the offer of having his neighbour's son mow his lawn for eleven dollars, he implies that half an hour of his time is worth at most eleven dollars. By turning down the offer to mow the other neighbour's lawn for twenty dollars, he implies that it is worth at least twenty dollars. But it cannot both be worth less than eleven and more than twenty dollars.

The explanation in terms of framing suggests[96] that people evaluate losses and gains forgone differently. Credit card companies exploit this difference when they insist that stores advertise cash discounts rather than credit card surcharges. The credit card

[93] Barry (1979a), pp. 17–18; Wolfinger and Rosenstone (1980), p. 8 and passim.

[94] For the reasoning behind this statement, see Oliver, Marwell and Teixeira (1985) or cases B, D and E in Schelling (1978), p. 220.

[95] I am indebted to Amos Tversky for suggesting this to me as an example of social norms.

[96] Thaler (1980), p. 43.

holder is affected less by the lost chance of getting the cash discount than by the extra cost of paying with the card. Similarly, the houseowner is affected more by the out-of-pocket expenses that he would incur by paying someone to mow his lawn than by the loss of a windfall income. But this cannot be the full story, because it does not explain why the houseowner should be indignant at the proposal. Part of the explanation must be that he does not think of himself as the kind of person who mows other people's lawns for money. It *isn't done,* to use a revealing phrase that often accompanies social norms.

Economists often argue that norms can be reduced to individual rationality. One version of the reductionist claim is that norms are 'nothing but' raw material for strategic manipulation; that people invoke norms to rationalize their self-interest while not believing in them. But this is absurd: if nobody believed in the norms, there would be nothing to manipulate.[97] A more serious reductionist argument proceeds from the fact that norms are maintained by sanctions. Suppose I face the choice between taking revenge for the murder of my cousin and not doing anything. The cost of revenge is that I might in turn be the target of countervengeance. The cost of not doing anything is that my family and friends are certain to desert me, leaving me out on my own, defencelessly exposed to predators. A cost–benefit analysis is likely to tell me that revenge is the rational choice. More generally, norm-guided behaviour is supported by the threat of sanctions that make it rational to obey the norm.

Against this argument, each of the following objections is a sufficient refutation. First, sometimes norms are followed in the absence of any observers who could sanction violations. Many people vote even when nobody would notice if they did not. Second, we have to ask why anyone would want to impose the sanctions. Perhaps they follow a metanorm to sanction people who violate first-order norms, but then we have to ask whether it is rational to follow that norm. In the regress that now arises, there must come a point at which the cost of expressing disapproval is less than the cost of receiving disapproval for not expressing it, since the former cost is approximately constant while the second

[97] Edgerton (1985), p. 3.

goes rapidly to zero. The chain of norms must have an unmoved mover, to which the rationalist reduction does not apply.

Among the alternatives to rational-choice theory, the (as yet undeveloped) theory of social norms holds out most promise. It is radically different from rational-choice theory, whereas the other alternatives are largely variations on the same consequentialist theme. They are different species of the same genus, whereas the theory of norms is of a different genus altogether. Other species of that genus might include the theory of neurotic behaviour, which is similarly rigid, mechanical and nonconsequentialist. Eventually, the goal of the social sciences must be to construct the family comprising both genera – to understand outcome-oriented motivations and nonconsequentialist ones as elements in a general theory of action. As long as this task is not accomplished, rational-choice theory will probably remain privileged, by virtue of the simplicity and power of the maximizing assumption. And in the event that it should one day be accomplished, rationality would still retain its privilege as a normative account of action.