I

Rationality

I.1 Introduction

The present work is largely about irrationality. Yet the discussion will hardly make sense without a prior analysis of the notion of rationality. This is embarrassingly rich. There is a bewildering multitude of entities that are said to be rational or irrational: beliefs, preferences, choices or decisions, actions, behavioural patterns, persons, even collectivities and institutions. Also, the connotations of the term 'rational' range from the formal notions of efficiency and consistency to the substantive notions of autonomy or self-determination. And in the background of the notion lurks the formidable pair of 'Verstand' vs. 'Vernunft', be it in the Kantian or in the Hegelian senses.

I begin with the focus on rationality as a formal feature of individual actions (I.2). This will provide what, following a similar terminology in Rawls,[1] I shall call the thin theory of rationality. It is thin in that it leaves unexamined the beliefs and the desires that form the reasons for the action whose rationality we are assessing, with the exception that they are stipulated not to be logically inconsistent. Consistency, in fact, is what rationality in the thin sense is all about: consistency within the belief system; consistency within the system of desires; and consistency between beliefs and desires on the one hand and the action for which they are reasons on the other hand.

The broad theory of individual rationality goes beyond these formal requirements (I.3). Rationality here involves more than acting consistently on consistent beliefs and desires: we also require that the beliefs and desires be rational in a more substantive sense. It is not too difficult to spell out what this means in the case of beliefs. Substantively rational

1 Rawls (1971), pp. 396ff., invokes 'the thin theory of the good to explain the rational preference for primary goods', while acknowledging that a fuller theory is needed to account for 'the moral worth of persons'.

beliefs are those which are grounded in the available evidence: they are closely linked to the notion of *judgment*. It is more difficult to define a corresponding notion of a substantively rational desire. One way of approaching the problem is by arguing that *autonomy* is for desires what judgment is for belief, and this is how I shall in the main proceed.

The notion of rationality can also be extended in a different direction, from the individual to the collective case. Once again I shall begin with the more formal considerations (I.4). At this level, rationality may either be attached to collective decision-making (as in social choice theory) or to the aggregate outcome of individual decisions. In both cases the individual desires and preferences are taken as *given*, and rationality defined mainly as a relation between preferences and the social outcome. A broader theory of collective rationality (I.5) will also have to look at the capacity of the social system or the collective decision mechanism to bring the individual preferences into line with the broad notion of individual rationality. A collectively rational arrangement in this sense is one which fosters autonomous wants, or is able to filter out non-autonomous ones.

In this chapter I am concerned with rationality, in later chapters with irrationality. One way of looking at the relation between these two notions is the following. Rationality tells the agent what to do; if he behaves otherwise, he is irrational. I shall argue against this view. There are many cases in which rationality – be it thin or broad – can do no more than exclude certain alternatives, while not providing any guide to the choice between the remaining. If we want to *explain* behaviour in such cases, causal considerations must be invoked in addition to the assumption of rationality. In fact, I argue below that if we require rationality in the broad sense, this will be the rule rather than the exception.

I.2 Individual rationality: the thin theory

Along the lines suggested by Donald Davidson,[2] rational action is action that stands in a certain relation to the agent's beliefs and desires (which I collectively refer to as his *reasons*). We must require, first, that the reasons are reasons for the action; secondly, that the reasons do in fact cause the action for which they are reasons; and thirdly, that the reasons cause the action 'in the right way'. Implicit in these requirements is also a consistency requirement for the desires and beliefs themselves. In what

2 See in particular the essays collected in Davidson (1980).

follows, the focus will mainly be on consistency, but first I have a few words to say about the three clauses that went into the definition of rational action.

The first clause can be taken in two ways. One might either say that the reasons are reasons for the action when, given the beliefs of the agent, the action in question is *the best* way to realize his desire. Or, more weakly, that the reasons are reasons for the action if it is *a* way of realizing the desire (given the beliefs). This distinction is related to, yet different from, the problem raised in the last paragraph of I.1 above. It is different because the question of unicity (is there *one* rational course of action?) must be distinguished from the question of optimality (is the rational course the *best*?). There might well be several alternatives that are equally and maximally good. I shall discuss these issues below. Here I only want to note how extremely thin is the theory of rationality we are dealing with here. If an agent has a compulsive desire to kill another person, and believes that the best way (or a way) of killing that person is to stick a pin through a doll representing him, then he acts rationally if he sticks a pin through the doll. Yet we might well want to question the substantive rationality of that desire and that belief.

The second clause of the definition is needed to exclude what we may call 'coincidences of the first class', in which a person has reasons for acting in the way he does, but is caused to do so by something other than these reasons. One might do by accident what one also has reasons for doing. Also, compulsive behaviour might occasionally be quite adequate to the occasion.

The third clause is needed to exclude 'coincidences of the second class', when the reasons do in fact cause the action for which they are reasons, but do so 'in the wrong way'. That reasons can cause an action 'in the wrong way' can be seen from the cases in which reasons cause an action for which they are not reasons. Davidson, for instance, argues that weakness of will can be explained along these lines.[3] The present case, however, is more complex, since the action which is caused by the reasons in the wrong way is the very action for which they are reasons. To see how this is possible, we invoke Davidson's notion of non-standard causal chains. An example from the external world is this: 'A man may try to kill someone by shooting at him. Suppose the killer misses his victim by a mile, but the shot stampedes a herd of wild pigs that trample the intended victim to death.'[4] We do not

3 Davidson (1980), Ch. 2. 4 *Ibid*. p. 78.

then want to say that the man killed the victim intentionally, since the causal chain is of the wrong kind. Correspondingly for the case of mental causality that concerns us here:

> A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope, he could rid himself of the weight and the danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never *chose* to loosen his hold, nor did he do so intentionally.[5]

Beliefs and desires can hardly be reasons for action unless they are consistent. They must not involve logical, conceptual or pragmatic contradictions. I shall first discuss consistency criteria for beliefs, and then at somewhat greater length for desires.

To evaluate the consistency of beliefs is not difficult, at least on the more superficial level at which we can assume that the beliefs have already been identified. At a deeper level we must accept Davidson's argument that identifying the beliefs of a person and assessing their consistency cannot be separated from each other. The process of belief imputation must be guided by the assumption that they are by and large consistent.[6] But once we have established the base line or background of general consistency, one may raise the question of local inconsistency of beliefs. The following holds only with this proviso.

We may look on beliefs either as subjective probability assessments, or as somehow *sui generis*. On the first reading, consistency simply means conformity to the laws of probability, so that the point probabilities of exclusive and exhaustive events add up to 1, the probability of the combination of any two of them is 0, etc. Similarly, the probabilities of compound events must have the right kind of relation to the probabilities of the elementary events, so that, say, a conjunction of independent events has a probability equal to the product of the component events.

For beliefs taken *sui generis* the obvious consistency criterion would seem to be that a set of beliefs are consistent if there is some possible world in which they are all true, i.e. if it is not possible to derive a contradiction from them. Jaakko Hintikka has shown, however, that this is insufficient.[7] His criterion is that the beliefs are consistent if there exists a possible world in which they are all true *and believed*. The need for the last

5 *Ibid*. p. 79. 6 *Ibid*. Ch. 12 and *passim*.
7 Hintikka (1961). For some applications, see Elster (1978a), pp. 81ff.

clause arises in cases of higher-order beliefs, i.e. beliefs about beliefs. Thus Niels Bohr at one time is said to have had a horseshoe over his door. Upon being asked whether he really believed that horseshoes bring luck, he answered, 'No, but I am told that they bring luck even to those who do not believe in them.'[8] Rigging the story a bit, this comes out as follows:

(1) Niels Bohr believes 'The horseshoe will not bring me luck.'
(2) Niels Bohr believes 'Horseshoes bring luck to those who do not believe they will bring them luck.'

Here there is no contradiction between the beliefs within quotation marks in (1) and (2), but we get an inconsistency if to these two beliefs we add (1) itself. So if we admit – as I think we should – that on intuitive grounds we would want to call the belief system inconsistent, we need the complex criterion to get a result in line with intuition.

To define consistency criteria for desires, we must first look more closely into the nature of the action in question. Roughly speaking, an action may be seen either as *doing something* or as *bringing about something*. When I take an apple from the fruit bowl, I am not setting up a causal process in the external world: I just do it (at will). By contrast, when I break the window by throwing an ash-tray at it, I bring about a change in the world by setting up a causal process that soon becomes independent of my will. (True, under other descriptions these characterizations may be reversed, but I am now concerned with the description under which the action is performed intentionally.) The explanations of these two actions are not quite assimilable to each other, although they both fall under the general scheme of rational action. I want an apple, and I take it: nothing more needs to be said. I may add, at the risk of some pedantry, that I believe there is an apple there; also, if I want a stronger form of explanation, that an apple is at the time what I want most, compared to the other options I believe to be available. In short, I *prefer* the apple. There is no need to go beyond this and add, falsely, that I take the apple *in order to* bring about a certain sensation in my taste organs, or to maximize a certain sensation. This would be true only in non-standard cases. I should add, however, that taste sensations may yet have explanatory force, at one remove: they are involved in the emergence and reinforcement of the preferences. They may be invoked in explaining my desire, not in describing it (see also II.10 below).

8 The story is told in Segrè (1980), p. 171.

In the ash-tray case, however, we must invoke more than mere preferences to explain the action, assuming it is not a mere *acte gratuit* as in Gide's *Les Caves du Vatican*. To understand the action we must postulate a *plan*, and specify a future state of affairs for the sake of which it was undertaken. The goal – breaking the window – could have been achieved by many means. One explanation of my action is simply that I believed throwing the ash-tray was one way of achieving the goal; a more ambitious explanation that I believed it to be the best way. If someone asks, 'Why did he throw the ash-tray?', this might be because he wants to know whether it was an expressive act of anger, or had the instrumental purpose of breaking the window; or to inquire into the reasons for breaking the window; or to understand why the ash-tray rather than some other object was chosen. Focussing on the last question brings out the distinction between preferences and plans. Choosing the ash-tray over the coffee mug is a different kind of action from that of choosing the apple over the orange. And I shall now go on to argue that quite different consistency criteria come into play for actions guided by preferences and by plans respectively.

The consistency criteria for preferences involve, minimally, *transitivity*: if I prefer a to b and b to c, I should prefer a to c. More complex consistency criteria are required when preferences are defined for options with a more complex internal structure. I shall consider two such complications, stemming from *probability* and *time* respectively.

Preferences may be defined over *lotteries*, i.e. over probability combinations of options, some of which may themselves be lotteries. This can be important practically, and is also crucial for the construction of a utility function that allows comparison of intensity of preferences.[9] It is usual then to assume the *dominance principle*: if one prefers a to b and $p > q$, then one should rationally prefer the option of getting a with probability p and b with probability $(1-p)$ to that of getting a with probability q and b with probability $(1-q)$. Also, one usually assumes the *reduction principle* that if a compound lottery – a lottery having lotteries among the options – is reduced to simple lotteries in the obvious way, the preferences should remain the same. Both assumptions have been challenged.[10]

Preferences may also be defined over whole *sequences* of options, bringing time into the picture in an essential way. In particular, we may define the notion of *time preferences* as an expression of the relative

9 For details about the construction, see Luce and Raiffa (1957), Ch. 2.
10 See for example Dreyfus and Dreyfus (1978) or Kahneman and Tversky (1979).

importance that at one point of time one accords to various later times or periods. Time preferences typically involve *discounting* the future, i.e. attaching less weight to future consumption or utility than to present. Such preferences are subject to two kinds of irrationality, which we may call respectively *incontinence* (or more neutrally impatience) and *inconsistency*. Incontinence involves discounting the future over and above what can be justified by mortality statistics and similar considerations. On the thin theory of rationality, we are not entitled to say that incontinence is irrational, unless the agent, at the time of acting incontinently, also believes that all things considered it would be best to wait. We would then be dealing with a case of weakness of will, briefly mentioned earlier. We might, on the other hand, espouse a broad theory of rationality that would enable us to characterize incontinence as irrational even when no such conflict is present.[11] By contrast, inconsistent time preferences are irrational even on the thin theory.[12] Consistency of time preferences is defined by requiring that a plan made at time t_1 for the allocation of consumption between times t_2 and t_3 should still remain in vigour when t_2 arrives, assuming that there has been no personality change or changes in the feasible set. With inconsistent time preferences one is never able to stick to past plans. It can be shown that consistent time preferences must be exponential, so that the future is discounted at a constant rate. George Ainslie has argued for the pervasiveness of non-exponential time preferences in human life, and shown that it is possible for an agent to exploit strategically this feature in order to overcome his incontinence.[13] Briefly the idea is that by grouping together several future choices the chances are increased that in each of them one will take the option with a later and greater reward. On the other hand this solution to the problem of impulsiveness may be as bad as the original difficulty, since the habit of grouping choices together may lead to rigid and compulsive behaviour.

11 On the issue of the irrationality of time preferences, see Maital and Maital (1978). They defend time preferences as rational because utility-maximizing, i.e. as rational in the thin sense of the term. See also the demonstration by Koopmans (1960) and by Koopmans, Diamond and Williamson (1964) that discounting the future is logically implied by a set of reasonable (although not compelling) assumptions about the shape of the utility function.

12 Strotz (1955–6); see also Elster (1979), Ch. II.5. I take this occasion to point to a serious mathematical error in my earlier treatment of inconsistent time preferences. In particular, the argument in Elster (1979), pp. 73ff., concerning the 'allocation of consistent planning' is largely incorrect. I am grateful to Aanund Hylland for spotting this mistake. It is corrected in the forthcoming Italian edition.

13 Ainslie (1982).

In addition to incontinence and temporal inconsistency, time also introduces the danger of *inconstancy*, or irrational preference change (including change of time preferences[14]). Not all preference change, of course, is irrational; indeed at times it may be irrational not to change one's preferences in the face of learning. I shall postpone the discussion of this issue, however, since here we clearly appeal to the broad notion of rationality. True, in I.4 below I shall give an example of endogenous preference change that could perhaps be said to be irrational on purely formal criteria, but in general we must invoke substantive considerations of autonomy.

In the theory of rational choice preferences are often required to be *complete* as well as consistent, meaning that for any pair of options one should be able to express a preference for one of them or, failing this, indifference. From the point of view adopted here, there are no strong arguments for this condition. In fact, one could argue that it is irrational to commit oneself to a preference for one of the options if one knows very little about either. At the very least it would be irrational to put much trust in such preferences.[15] For the purposes of model building, however, it is clear that a full ordering of the available options is a much more powerful notion than a partial ordering. But if one is guided by reality rather than by convenience, there seems to be a choice between postulating partial or incomplete preference orderings, and postulating complete preferences subject to endogenous change as the agent learns more about the alternatives. Postulating preferences that are both complete and stable seems too remote from the real world.

In addition to consistency and completeness, it is often assumed that preferences have the property of *continuity*. Very broadly speaking, this means that if one prefers *a* over *b*, and *a* undergoes a very small change (as small as you please), then the preferences should not be reversed.[16] This requirement is violated in the case of so-called non-Archimedean

14 See Meyer (1977) and Samuelson (1976) for this idea.

15 Cyert and de Groot (1975), pp. 230ff. A related but importantly different argument is offered by Tocqueville (1969, p. 582): in a democracy people 'are afraid of themselves, dreading that their taste having changed, they will come to regret not being able to drop what once had formed the object of their lust'. Whence the tendency of the Americans to eschew durable consumer goods. Whereas Cyert and de Groot argue that a rational person should anticipate that his tastes will change because of new experience, Tocqueville suggests that the Americans – rationally or not – act on the assumption that their taste will change irrationally in the future.

16 For a more precise statement, see Rader (1972), pp. 147ff.

preferences, an important special case of which is the lexicographic preference structure involving a hierarchy of values. If I am starving and am offered the choice between an option involving one loaf and listening to a Bach record and another involving one loaf and listening to Beethoven, then my love for Bach may make me prefer the first option. If, however, from the first option is subtracted even a very small crumb of bread, as small as you please, then I switch to the second because at starvation level calories are incomparably more important than music. There is nothing irrational in this preference switch, and so continuity cannot be part of rationality.[17] For model-building purposes, however, the condition is very important, since preferences that are transitive, complete and continuous can be represented through a real-valued *utility function*.

At this point two observations suggest themselves. First, to maximize utility is not to engage in the carrying out of a plan, choosing the best means to realize an independently defined end. In the modern theory of utility, it is essentially a short-hand for preferences, and implies nothing about more or less pleasurable mental states that could be seen as the goal of the behaviour. Now there are good reasons for thinking that this ordinal conception of welfare carries things too far, for surely we know from introspection that pleasure, happiness and satisfaction are meaningful notions, if only we could get a conceptual handle on them, which may prove difficult. My point here is that even if one should succeed in defining a cardinal measure of utility, it would be a mistake to believe that action could then always be explained in terms of utility maximization in the same sense as, say, investment may be explained in terms of profit maximization. The latter operation is (in the standard models[18]) conceived of as a plan undertaken consciously and *ex ante*, whereas the conscious and deliberate attempt to maximize utility tends to be self-defeating. It is a truism, and an important one, that happiness tends to elude those who actively strive for it. Much of Ch. II is devoted to a further analysis of this idea. Here I only want to stress that even if actions may sometimes be explained as *attempts* to maximize utility in this *ex ante*

17 For a strong argument to this effect, see Georgescu-Roegen (1954). The rhetoric of Marcuse (1964) can be understood within this framework: if the preferences can be mapped into the real line, we are indeed dealing with 'one-dimensional man'. Similarly Borch (1968, p. 22) observes that the postulate of continuous preferences amounts to saying that 'everything has a price'.

18 For a discussion of non-standard models, see Elster (1982a), Ch. 6.

sense, we would not be justified in thinking that the attempt would succeed; rather the contrary.[19] On the other hand, as I observed earlier, when the utility-maximizing consequences of behaviour *can* be invoked to explain it, they do so by providing a causal explanation of the preferences. Pleasurable inner states enter importantly into the explanation of behaviour, but not as the conscious goal of behaviour.

Secondly, we may usefully contrast *rational man* with *economic man*. The first involves – in the thin sense which we are discussing now – nothing but consistent preferences and (to anticipate) consistent plans. The second is a much better-endowed creature, with preferences that are not only consistent, but also complete, continuous and *selfish*. To be sure, economists have constructed a large variety of models involving non-selfish preferences,[20] but their reflex is nevertheless to attempt to derive all apparently non-selfish behaviour from selfish preferences.[21] This may perhaps be a good research strategy: when setting out to explain a given piece of behaviour, assume first that it is selfish; if not, then at least rational; if not, then at least intentional. But there can be no way of justifying the substantive assumption that all forms of altruism, solidarity and sacrifice really are ultra-subtle forms of self-interest, except by the trivializing gambit of arguing that people have concern for others because they want to avoid being distressed by their distress. And even this gambit, as Allan Gibbard has pointed out, is open to the objection that rational distress-minimizers could often use more efficient means than helping others.[22]

19 As emphasized in van Parijs (1981) and Elster (1982a), one should distinguish between explanation in terms of intended and in terms of actual consequences of behaviour, although there is of course no general presumption that the intended consequences will fail to materialize – except for the class of cases that form the subject of Ch. II below.

20 See the useful survey and discussion in Kolm (1981a).

21 See in particular the important synthesis of biological and game-theoretic considerations in Axelrod and Hamilton (1981). They use a model of sequential Prisoner's Dilemmas to show (i) that genuinely altruistic motivation can arise out of natural selection by purely selfish criteria and (ii) that some cases of apparently altruistic motivation can be explained by assuming no more than selfish rationality. In other words, if people behave altruistically, it is either because they have been programmed to feel concern for others or because they have calculated that it pays to fake concern for others. The first explanation, while in a sense reductionist, allows rational resistance to the economic reductionism embodied in the second. Yet there probably are cases that are resistant also to biological reductionism, unless one postulates that fitness-reducing altruism can be explained by the fact that 'it is not worth burdening the germ plasm with the information necessary to realize such an adjustment' (Williams 1966, p. 206).

22 Gibbard 1986.