The Economic Journal, **101** (*July* 1991), 751–785
Printed in Great Britain

RATIONAL CHOICE: A SURVEY OF CONTRIBUTIONS FROM ECONOMICS AND PHILOSOPHY*

Robert Sugden

The theory of rational choice has a central place in modern economics. In mainstream economics, explanations are regarded as 'economic' to the extent that they explain the relevant phenomena in terms of the rational choices of individual economic agents. Theories which seem not to have this structure – John Maynard Keynes's (1936) General Theory, with its references to psychological propensities and animal spirits,[1] is a classic example – are regarded as suspect until their 'microfoundations' have been properly constructed. But what exactly do we *mean* by rational choice?

Some economists insist that economic theory is purely descriptive or 'positive': its purpose is to predict human behaviour, and nothing more. On this view, rational-choice theory describes certain regularities in human behaviour. If the theory works – that is, if it generates predictions that are generally in accord with our observations – then this provides us with all the justification we need for accepting its assumptions as working hypotheses. Conversely, if the theory's predictions fail to accord with our observations, then we must look for a better theory. We may be entitled to be more sceptical of animal spirits than of well-behaved preference orderings, but only on grounds of parsimony, or because (and if) we have more experience of successfully predicting human behaviour in terms of preference orderings.

Economists who take this line sometimes also say that the concept of rationality has no real content. Rationality, they say, equals consistency: a person is rational to the extent that his or her choices are consistent with one another. But this means only that those choices are consistent with one another *when viewed from the perspective of some theory* – that is, that they can be predicted by that theory. On this account, the expression 'rational-choice theory' is a tautology: any *theory* of choice must postulate some consistent pattern that is to

* This paper was begun at the University of California, Davis, where I was a visiting professor in spring 1990. It was completed at the University of East Anglia, as part of the Foundations of Rational Choice Theory project, supported by the Economic and Social Research Council (award R 000232269). The ideas contained in the paper developed out of discussions with many people, including Richard Arneson, Michael Bacharach, Giacomo Bonanno, David Copp, Robin Cubitt, David Gauthier, Jean Hampton, Martin Hollis, Gregory Kavka, Graham Loomes, Edward McClennen, Judith Mehta, Susan Mendus, Philip Pettit and Chris Starmer.

This survey paper is the twelfth and last in the series published in the JOURNAL. It was commissioned by me, and not by the now-retired Surveys Editor, Andrew Oswald [JDH].)

[1] 'Most, probably, of our decisions to do something positive, the full consequences of which will be drawn out over many days to come, can only be taken as a result of animal spirits – of a spontaneous urge to action rather than inaction, and not as the outcome of a weighted average of quantitative benefits multiplied by quantitative probabilities. Enterprise only pretends to itself to be mainly actuated by the statements in its own prospectus, however candid and sincere.' (Keynes, 1936, pp. 161–2.)

be found in people's choices, and then anyone who chooses as the theory predicts must be acting consistently with that pattern and therefore rationally.

The arguments of this paper will be of no interest to a convinced positivist. I shall be concerned with a different view of rational-choice theory, in which the theory is seen as having a genuinely normative content: it tells us how, as rational agents, we *ought* to choose. If the theory also has predictive power, this is because, in a non-tautological sense, human beings have some tendency to act rationally. In this paper, I shall ask what we can say about how rational agents ought to choose.

At least within economics, and probably more generally, there can be said to be a received theory of rational choice: expected utility theory. This theory is thought to apply not only to individual decision-making in 'games against nature', but also – by way of postulates about individuals' rationality being common knowledge – to games in which rational individuals interact strategically. I shall be examining the philosophical foundations of this theory.

It is often said that the theory of rational choice used within economics embodies an instrumental conception of rationality – a conception whose classic statement can be found in the work of Hume. Modern economics can certainly trace a direct line of descent from Hume and the Scottish Enlightenment, and Hume would be the most obvious candidate for the position of economists' philosopher. But it is an open question whether the theory of rational choice, as we know it today, can be given a consistent foundation in terms of a Humean conception of rationality. This is one of the issues to be explored in this paper.

A second issue to be explored is one about which many economic theorists are beginning to be concerned. In game theory, it is conventional to suppose not only that rational players choose according to expected utility theory, but also that it is common knowledge that they so choose. The main project of game theory has been to work out the implications of this formulation of rationality. Some of these implications turn out to be very puzzling, and raise questions about the project itself. The classic presentations of expected utility theory are constructed in terms of games against nature: the chooser is uncertain about which of a set of mutually exclusive states of nature is the true one. Game theory prescribes that that a rational player should treat his opponent's acts as if they were states of nature while also recognising that the opponent is rational in the same sense. I shall be asking whether this conception of rational play is coherent.

This connects with a third issue: is the theory of rational choice self-defeating when applied to a world in which rational agents interact with one another? A number of philosophers have argued that it is. Roughly, the suggestion is this: in terms of the theory, choices are rational to the extent that they lead to the satisfaction of the chooser's preferences. But, it is said, a person who accepts the theory and who behaves according to its prescriptions may end up satisfying his preferences less well than he would have done, had he accepted some other 'irrational' principle of choice.

I. INSTRUMENTAL RATIONALITY

To suppose that a theory of rational choice is possible is to suppose that in some way, choices can be influenced by reason. But how does reason influence choice? The mainstream tradition of economic theory has been strongly influenced by nineteenth-century utilitarianism and by the Scottish Enlightenment thinking of Hume and Adam Smith. One distinguishing feature of this body of ideas is its *instrumental* conception of rationality.

Hume provides the most famous statement of the instrumental view of rationality. 'Reason alone', he says, 'can never be a motive to any action of the will'; and 'reason is, and ought only to be the slave of the passions' (Hume, 1740/1978, pp. 413, 415). The ultimate motive for any act must be some kind of pure feeling or 'passion'. Since all we can say about a state of feeling is that it exists within us, there is nothing in such a state on which reason can get a grip. The qualification 'ultimate' is important here. Hume recognises that some of our desires may be formed as a result of rational reflection, but insists that any such reflection must take some desires as given. Actions can be motivated only by desires, and no desire can be brought into existence by reason alone.

This theory of motivation is controversial, but it is accepted by many modern philosophers. Michael Smith (1987), for example, argues for it in the following way. Following Elizabeth Anscombe, he distinguishes between two kinds of mental states. One kind, represented by beliefs, is intended to fit the world as it is; if a belief does not fit the world, it is untrue and has to be rejected. The other kind, represented by desires, has a different 'direction of fit' with the world: a desire is realised by making the world fit the desire. To recognise this categorical distinction is to understand why desires cannot be derived from beliefs. When we ask what motivates a person's action, we are necessarily working in the realm of desires: the motive for an action must involve some kind of goal, and 'having a goal' is a mental state with which the world must fit.[2]

What role is left for reason in determining choices? Passions in themselves, Hume says, cannot be called unreasonable or irrational: 'a passion must be accompay'd with some false judgment, in order to its being unreasonable; and even then 'tis not the passion, properly speaking, which is unreasonable, but the judgment.' And: 'a passion can never, in any sense, be call'd unreasonable, but when founded on a false supposition, or when it chuses means insufficient for the design'd end' (Hume, 1740/1978, p. 416). Thus reason is to be seen as an instrument for achieving ends that are not themselves given by reason. We may say that an act is irrational if it is not the best means of achieving the ends that the actor himself had a view when choosing the act. Even an irrational act will normally be motivated by some desire of the actor. What makes the act irrational is that this desire is founded on a false belief (that the act *is* the best

[2] A different defence of a Humean theory of motivation is given by Williams (1981, pp. 101–13). An opposing view can be found in Nagel (1970).

means of achieving the actor's ends); and ultimately, the irrationality resides in the belief rather than in the act.[3]

Hume's account of rationality undoubtedly has many echoes in modern economics. But this is not to say that it provides a justification for the standard theory of rational choice. There are at least two features of rational-choice theory that present problems for a Humean interpretation.

First, consider whether, given a complete description of a person's subjective attitudes to a choice problem – given, that is, a complete description of his desires and beliefs – reason is always able to tell him what to do. Economists usually want the theory of rational choice to have this property. A rational person is supposed to have preferences which need not be susceptible to any form of rational appraisal; but once these preferences are given, his choices are determined. Only one exception is normally allowed. This is the case of indifference, where a person regards two or more courses of action as exactly equal in desirability.[4] I shall say that a theory of choice is *determinate* if, given a complete description of a person's desires and beliefs, the theory prescribes a unique course of action for every contingency – except in the case of indifference. I shall be arguing that determinacy is not implied by Hume's theory of motivation.

The second significant feature of rational-choice theory is that it requires a person's preferences to satisfy certain conditions of consistency, such as transitivity. It is not clear that such conditions are implied by Hume's conception of rationality. For Hume, pure passions cannot be said to be consistent or inconsistent with one another; if we are to claim that (say) non-transitive preferences are irrational, we must show that they are grounded in inconsistent beliefs. Later I shall consider whether it is possible to show this.

These difficulties could be avoided if we did not treat all the assumptions of rational-choice theory as axioms of rationality. Instead, we might say that the theory rests on psychological assumptions about the nature of human desires – assumptions whose truth or falsity is an empirical matter. This was the strategy followed by utilitarian economists of the late nineteenth century such as Jevons (1871/1970). Jevons's idea is to start with a theory of psychology in which all pleasures and pains can be reduced to a single dimension of 'utility', and in which all desires can be reduced to the desire to maximise utility. It then follows straightforwardly that preferences are complete and transitive. Combining this with a Humean theory of motivation, we arrive at an internally consistent and determinate theory of instrumentally rational choice. Jevons recognises that the psychological assumptions of his theory are simplifications, but argues that they are sufficiently realistic for the subject-matter of economics. Like Alfred Marshall (1920/1949, p. 1), Jevons seems to think of economics as being about mankind in the ordinary business of life, with 'business' understood in a fairly narrow way: it is not about all possible forms of rational choice. For

[3] On Hume's conception of reason, see Stroud (1977, chapter 7).

[4] This exception causes real problems for positivists. Savage (1954, p. 17) glances at them quickly and then passes by. Samuelson (1947, pp. 90–113), in his revealed preference theory, faces up to them by making it an axiom of his system what whatever is chosen is strictly preferred to everything else. But this works only because Samuelson restricts his analysis to choices made by price-taking consumers.

this purpose, he argues, it is sufficient to study 'the lowest rank of feelings' (Jevons (1871/1970, p. 93)).

Modern rational-choice theorists have opted for the more ambitious strategy of grounding the theory on axioms of rationality. I shall argue that the axioms from which the standard theory is derived are far stronger than can be justified by an appeal to a Humean conception of rationality.

II. KANTIAN RATIONALITY

The Humean conception of rationality is often contrasted with another conception, which derives from the work of Immanuel Kant (1781/1896, 1785/1949).[5] It would be hard to argue that the standard theory of rational choice owes much to Kant. But from a philosophical point of view, Kant's conception of rationality is the most prominent alternative to the instrumental one. It provides a standpoint from which the standard theory might be criticised.

On the instrumental view, action is ultimately determined by psychological states. Thus an instrumental theory of choice offers a causal explanation of human action that has essentially the same structure as the explanations we find in the natural sciences. Kant accepts that, if we set out to *explain* the choices people make, we must adopt this kind of approach: human beings are part of the physical world, and so we must conceive of their actions as having physical explanations. But when we reason, Kant says, we cannot do other than conceive of ourselves as *autonomous*: we must think of ourselves as capable of forming beliefs and of reaching conclusions that are not determined for us by outside causes. Similarly, when we reason about what actions to take, we must conceive of ourselves as being able to determine our actions. Reason, Kant says, must regard itself as the author of its principles, independently of alien influences. In this context, Humean passions must count as alien influences: we cannot reason coherently about what to do while thinking of our choices as wholly determined by our psychological states. Kant does not attempt to prove that in any sense we really are autonomous. His claim is that we cannot engage in any form of reasoning without presupposing our own autonomy.

There is a second dimension to the Kantian notion of autonomy. For Kant, an autonomous person is one whose actions are governed by laws that he has imposed on himself. This is central to one of Kant's conceptions of rational choice. I say 'one of', because Kant accepts the Humean conception of instrumental rationality, which he sees as the source of *hypothetical imperatives*. (If you want to achieve Y, then do X.) But for Kant there is another form of rationality, which generates *categorical imperatives*. (Do X, regardless of your wants.) On this conception, a choice is rational if it is prescribed by some principle which the chooser can will be to be a universal law for all rational agents. To ask whether an action is rational, we must not (as in the

[5] My discussion of Kant relies heavily on O'Neill (1975, 1989).

instrumental approach) ask how it connects with the psychologically given desires and beliefs of the actor; we must instead examine the coherence of the principles which – from the viewpoint of the actor, conceiving himself as autonomous – determined the action.

It is crucial for Kant that categorical imperatives are dictated by reason alone. Thus for Kant, reason alone *can* be a motive for an action of the will. Categorical imperatives are imperatives that would be recognised by any agent possessing the faculty of reason. Thus they are not merely independent of the particular desires of any particular agent; they are independent of any facts, however general, about human psychology or human society. The autonomous agent imposes his own laws; but if each agent arrives at these laws by the use of reason, all will arrive at the same laws. Kant argues that any such law must have the formal property that each agent can, without logical inconsistency, will that the law should be binding on all agents: the law must be *universalisable*.

Many variants of this central idea have been proposed and analysed.[6] A common move has been to drop Kant's insistence that moral laws must be independent of human desires, while retaining the idea that they must be universalisable. Thus we may use information about desires among the data from which we derive prescriptions for action, but the method of derivation may be different from that of instrumental rationality. For example, consider the Prisoner's Dilemma. Two players, *A* and *B*, must each decide whether to 'cooperate' or to 'defect'; defection is the dominant strategy for each, but joint cooperation is better for both than joint defection. On the standard, instrumental account of rationality, rationality requires each to defect. But can either player will that the maxim 'Defect in Prisoner's Dilemmas' is a universal law?

From a pure Kantian perspective, there seems to be no logical inconsistency in their willing this. But, clearly, neither player would *desire* that this maxim should be a universal law. If either player had to choose a rule that was to apply universally, and if in making this choice he was guided by his desires or preferences, he would choose a rule which, if followed by both players, would lead both to cooperate. From this we might conclude that each player has a reason for cooperating: cooperation is prescribed by a universalisable principle, while defection is not. Notice that each player's desires or preferences still favour defection in the game itself, since defection is the dominant strategy. But if we work within the Kantian tradition, reasons may override desires: it may be rational to do what one does not desire to do. As the example of the Prisoner's Dilemma suggests, this line of thought threatens to undermine game theory. I shall say more about this later.

[6] For example, Hare (1952, 1963) argues that it is a property of our language that moral propositions are universalisable. Rawls's (1971) 'veil of ignorance' is a formalisation of the requirement that principles of justice should be universalisable. Rule utilitarianism – the moral principle that each person should follow those rules which, if universally followed, would maximise overall welfare – represents another form of universalisation. Regan (1980) reviews various forms of rule utilitarianism and proposes a universalisable rule of his own (see Section IX below).

III. RATIONALITY AS CONSISTENCY

I have suggested that the utilitarian economists of the late nineteenth century had a consistent theory of instrumentally rational choice. From the 1930s, however, economists began to become embarrassed by the old-fashioned utilitarian psychology they were carrying round, and a conscious attempt was made to jettison it. An alternative foundation for the theory of choice was sought in a purer concept of rationality, free of any psychological assumptions. The culmination of this programme was Savage's *Foundations of Statistics* (1954). Savage's theory draws on the work of Ramsey (1931), who showed that subjective probability could be defined in terms of preferences over gambles, and on that of von Neumann and Morgenstern (1947), who provided the first axiomatic derivation of expected utility theory. In Savage's theory, the central concepts of modern decision theory – subjective probability, Bayesian learning, and the maximisation of expected utility – are derived from a compact system of simple axioms. When economists and game theorists feel obliged to justify their use of these concepts, they still turn to Savage. I shall be treating Savage as the best spokesman for conventional rational-choice theory.

Savage sees the principal purpose of his theory as normative. He says that he is concerned with 'the implications of reasoning for the making of decisions', and draws analogies between his axioms and the principles of logic. He says that it has often been asked 'whether logic cannot be extended, by principles as acceptable as those of logic itself, to bear more fully on uncertainty'; and he presents his own work as an attempt to do exactly this (p. 6).[7] In this he is echoing Ramsey (1931, p. 166), who describes his own work as an enquiry into 'the logic of partial belief'. The main value of logic, Savage says, is that it provides criteria for detecting inconsistencies among beliefs and for deriving new beliefs from existing ones. Analogously: 'the main use I would make of [my postulates] is normative, to police my own decisions for consistency and, where possible, to make complicated decisions depend on simpler ones' (p. 20).

In Savage's theory, uncertainty is represented by a set of mutually exclusive *states of the world*, one and only one of which *obtains*. Any set of states of the world is an *event*. The objects among which choices are made are *acts*; an act is defined by a list of *consequences*, one for every state of the world. An individual has *preferences* over acts. States of the world, consequences and preferences are primitives in the formal theory. Savage presents a set of axioms which impose conditions of consistency on preferences. He then proves the following theorem. If a person's preferences satisfy these axioms, then those preferences can be represented by a utility function, unique up to positive linear transformations, which assigns a utility index to every consequence, and by a unique probability function, which assigns a probability index to every event. These functions can be used to assign an expected utility index to every act. Of any two acts, the one with the higher expected utility will be preferred.

If we are to evaluate Savage's theory, we must ask what he means by 'preference' and 'probability'. Savage defines probability in terms of

[7] In this section, unattributed page references are to Savage (1954).

preference. (Assuming that more money is preferred to less, a person who obeys all Savage's axioms and who prefers £100 conditional on event A to £100 conditional on event B is defined to believe A to be more probable than B.) Savage sees this as a way of expressing a 'personalistic' or *subjective* conception of probability, in which 'probability measures the confidence that a particular individual has in the truth of a particular proposition'. Subjective views of probability 'postulate that the individual concerned is in some ways "reasonable", but they do not deny that two reasonable individuals faced with the same evidence may have different degrees of confidence in the truth of the same proposition' (p. 3).

Savage insists that preference must be interpreted in terms of choice. As an informal definition of preference, he says that the statement 'the person prefers f to g' (where f and g are two acts) means 'if he were required to decide between f and g, no other acts being available, he would decide on f' (p. 17). Recognising (but not resolving) the problem of making a behavioural distinction between preference and indifference, he emphatically rejects the idea of using introspection as a source of information about preferences. 'I think it of great importance', he says, 'that preference, and indifference, between f and g be determined, at least in principle, by decisions between acts and not by response to introspective questions' (p. 17). Thus although Savage's axioms are formulated in terms of the concept of preference, it seems that he regards choice as the more fundamental concept: the idea is to construct a theory of rational choice, not of rational preferences.[8]

It is crucial for Savage that preferences are *complete*: for every pair of acts f, g, one and only one of the propositions 'f is strictly preferred to g', 'f and g are indifferent' or 'g is strictly preferred to f' must be true. This is one of the axioms of his theory, and is essential for the proof of his theorem. From a more philosophical point of view, completeness is essential for Savage's project of constructing a theory of rational *choice*. If preferences were not complete, we would not be entitled to assume that choices revealed preferences; and it is only by assuming this that we can translate principles about the consistency of preferences into principles about the consistency of choices. But why, we might ask, should completeness be a requirement of *rationality*?[9]

Imagine that someone, say Jane, has to choose between two acts f and g, where f gives £15,000 if a fair coin falls heads 10 times in succession, while g gives £10 with certainty.[10] Suppose Jane says: 'I really don't know which of these I prefer. I have a very strong desire to have £15,000, but f gives me only one chance in 1,024 of winning. My desire for £10 is quite weak, but g gives

[8] Savage's interpretation of preference is, I think, the standard one in modern economics. Gauthier (1986, pp. 26–9) offers an alternative interpretation, in which a person's preferences are revealed both in his choices and in the statements he makes about his attitudes.

[9] It is surprising how few theorists have been willing to separate rationality from determinacy. Among the few to have done so is Levi (1986). Levi presents a theory of 'unresolved conflict' in which a rational person may find that more than one feasible option is rationally choosable, even though he is not indifferent between them. In game theory, the concept of rationalisability (see Section VII) is a non-determinate theory of rational choice.

[10] The following discussion of determinacy and framing is based on work I have done with Chris Starmer.

me this for certain. I just can't see how to strike a balance between these conflicting thoughts.' Jane does not have any primitive desire in relation to the choice between f and g: what she has are desires which point in opposite directions, and which she cannot make commensurable. If we interpreted preference in terms of desire, we could say that f and g do not stand in *any* relation of preference or indifference to one another.

As it stands, this conclusion has no bearing on Savage's theory: Savage defines preference in terms of choice, not desire. But if Jane's desires are as I have described, her choice between f and g might be random, or be determined in a non-rational way by her psychological response to the 'framing' of the choice problem (see below). Thus Jane might choose *either* f *or* g. Perhaps she would choose f in some cases and g in others, even though – from the viewpoint of Savage's theory – there was no difference between the cases. Savage's theory provides us with a way, but only one way, of describing such behaviour: we must say that Jane is indifferent between f and g.

This description, however, is unlikely to be satisfactory. Suppose we introduce a third act, h, which gives £10.01 with certainty, and we ask Jane to choose between f and h. And suppose, as seems quite plausible, that she gives the same answer as before: she doesn't know which of the two she prefers. If we interpret this as another case of indifference, we have g indifferent to f and f indifferent to h. So by transitivity – another of Savage's axioms – we have g indifferent to h. But if we ask Jane which she would choose of g and h, she may have no difficulty at all in opting for h: she may be quite sure that she prefers more money to less. Jane's responses are incompatible with Savage's axioms. But are they irrational?

The imaginary case of Jane may be related to a growing body of experimental evidence of *framing effects* (see, e.g. Tversky and Kahneman (1986)). We may take a single choice problem and describe (or 'frame') it in two different ways, such that almost anyone who saw both descriptions would, on reflection, agree that they were logically equivalent. But if people are presented with the problems separately, with some interval of time between so that the logical equivalence of the problems is concealed, they may respond in systematically different ways to the two descriptions. (A famous example, due to Tversky and Kahneman, concerns the difference between describing public health policies in terms of 'lives saved' and in terms of 'lives lost'.) Such patterns of behaviour resist explanation, not only by expected utility theory itself, but also by the many competing generalisations of that theory that economists have put forward in recent years. The usual response of economists is that such behaviour is simply irrational. Tversky and Kahneman (1986) do not disagree: they merely insist that it is a mistake to ground a theory of how people actually choose on assumptions about rationality.

But consider Jane again. She has to choose one of the two acts f and g, even if she can find no adequate reasons for her choice. Suppose she chooses f. When we ask her why she chose f, she points to some aspect of the framing of the problem which made f salient for her. She admits that, if the problem had been framed in a different way, she would have chosen g. She does not present this

as a *reason* for choosing one act rather than the other; she is simply reporting how she came to choose in a situation in which reasons were inadequate. If we accept a Humean theory of motivation, there seem to be no grounds for calling Jane irrational. Her actions are not based on any false beliefs. Nor (to use Savage's language) do her thought processes show anything analogous with an error of logic. The truth is that her desires do not provide her with enough data from which to work out what she should do. If, as Hume maintains, the presence or absence of a desire is ultimately a matter of psychological fact, then whether or not a person's desires have the right structure for reason to work on is a matter of fact too: it is beyond rational appraisal. We must admit the possibility that reason is like one of those algorithms that operations researchers are trained to use: a useful instrument, but only for decision problems that happen to have the right structure.

Unless (like Jevons) we are prepared to make empirical assumptions about the structure of people's desires and beliefs, we cannot say that there is an objectively correct answer to Jane's choice problem, waiting to be found out by reason. And the whole point of Savage's subjectivist approach is to avoid such psychological assumptions. Within this approach, there is simply no way of asking which act should rationally be chosen in any decision problem, considered in isolation: we can only ask whether the choice of one act in one problem is consistent with the choice of another act in another problem.

If we start from an instrumental conception of rationality, then, it seems that completeness is not a necessary feature of rational preferences. Savage, however, clearly wants to claim that his axioms are axioms of rationality. This raises the question of whether Savage's theory should be understood as a theory of instrumental rationality. Notice that for Savage, rationality is understood in terms of the consistency of choices *with one another*, and not in terms of their consistency with any given system of desires and beliefs. Savage does not start with measures of utility and probability and then work out what it would be rational for a person to choose; he starts with a consistent pattern of choices and then derives measures of utility and probability from these choices. Certain patterns of choices are deemed to be irrational, by virtue of their internal inconsistency, and quite independently of any reference to the chooser's mental states. This does not fit easily with the idea of reason being the slave of the passions.

I must confess that I find it hard to formulate an appropriate conception of rationality to fit Savage's theory; but here is one tentative suggestion. Suppose we say that a choice is rational if it is one for which the chooser can give a determining set of reasons. By 'determining', I mean that those reasons identify that particular choice as the one that should be made. (The usual qualifications about indifference should be taken as read.) We do not start with any presuppositions about what those reasons might be: we do not assume any particular theory of motivation, or make any particular assumptions about psychology. But when a rational person makes a decision, he is seen as committing himself to the reasons that determine that choice for him. Then two decisions can be said to be inconsistent if they cannot be determined by any

single set of reasons. Savage's project, then, is to identify the restrictions that would be imposed on choices by *any* consistent set of reasons.

This interpretation allows us to make sense of the idea – one that is clearly important for Savage – that the concepts of consistency and inconsistency can be applied directly to decisions. It also allows us to understand how completeness might be seen as a principle of rationality. On the current interpretation, to say that a person prefers x to y is to say that he is committed to a set of reasons which imply that if he has to choose between x and y, he should choose x. Thus if a person's preferences are incomplete, there are situations in which he will make choices without having reasons for those choices. We might want to say that such choices are, if not exactly irrational, at least non-rational. To have all one's choices supported by reasons might be seen as an ideal of rationality, or perhaps of autonomy. (An autonomous agent acts in accordance with principles that he has chosen for himself. The person who has no reasons for his choices is not acting on any principles: in this sense, he is not fully autonomous.) Whether this ideal is attainable, however, remains an open question.

IV. TRANSITIVITY AND THE CONCEPT OF A CONSEQUENCE

Savage's theory is built on axioms about the consistency of preferences. The transitivity axiom is the most familiar of these, and the one whose status as a principle of rationality is usually thought to be the most secure. In examining Savage's idea of consistency, I shall concentrate on this axiom. The argument which follows, however, could easily be recast so as to apply to Savage's other main axiom of consistency, the 'sure-thing principle' (Loomes and Sugden (1986); Broome (1991, chapter 5)).

Savage's axioms are formulated in terms of preferences over acts, where acts are made up of consequences. The concept of a consequence is primitive in Savage's theory. Savage (1954, p. 13) gives only the informal definition: 'A consequence is anything that may happen [to the person who is choosing]'. It is tempting to think that the formal theory imposes no restrictions on what may count as a consequence. That this is not in fact the case is a matter of some significance, both for the present argument, and for some that will come later in this paper.

Savage's theory requires that there is a given set X of possible consequences, and that every function from the set of all states of the world into X is an act. In other words, the theorist is free to construct acts by arbitrarily assigning consequences to states of the world. This property of Savage's theory has been highlighted by Broome (1991, ch. 5), who calls it the 'rectangular field assumption'. Further, for every pair of acts f, g, a preference relation is defined; and preference is interpreted in terms of choice. So to say that a person has any kind of preference relation between two acts f and g is to imply that it is possible to confront that person with a choice between those two acts. This feature of acts – that any pair of acts must be capable of constituting a meaningful choice problem – is clearly required by Savage's approach, in which preferences are

defined in terms of observable choice behaviour. But even if we were to allow an introspective interpretation of preference, it would seem most natural to interpret a preference as a disposition to choose one thing rather than another. One of the main ways in which we come to know our own preferences is by noting how we in fact choose, or by constructing hypothetical choice problems for ourselves and monitoring our responses. If a pair of acts cannot constitute a meaningful choice problem, then it is doubtful whether the concept of a preference between them is meaningful either.[11]

The implication of all this is that consequences must be defined so that any assignment of consequences to states of the world is a meaningful act, and so that any pair of such acts is a meaningful choice problem. Thus the description of a consequence may not include any reference to any particular choice problem. Suppose we take a choice problem in which F is the set of feasible acts. In some particular event E, some act f gives consequence x. Then the description of x may not include any reference to the event E; nor may it include any reference to any property of the act f; nor may it include any reference to any properties of any other acts in F.

This is not mere hair-splitting. Savage's proof of the existence of a utility function depends on our being free to construct pairs of acts arbitrarily, and to identify preferences between such acts. (For example, suppose we wish to assign a utility index to some consequence x. We can do this by constructing a choice between two acts, one of which gives x in all states of the world, and the other of which gives w (which is preferred to x) in some event E and y (which is less preferred than x) otherwise. If we choose E so that these two acts are indifferent, we can use the subjective probability of E to fix the utility of x relative to that of w and y.) So anyone who uses the concept of utility and who justifies this by appealing to Savage's axioms is committed to this restriction on what may count as a consequence; descriptions of 'things that may happen' that do not satisfy this restriction cannot be given utility numbers. The same conclusion holds if we appeal to von Neumann and Morgenstern's (1947) axioms, which require another version of the rectangular field assumption.

Now consider the transitivity axiom. Let f, g and h be acts, and suppose that someone, say Cathy, would choose f from the set of feasible acts $\{f, g\}$, g from $\{g, h\}$, and h from $\{h, f\}$. Let us take it as given that these choices reflect strict preferences. Could such a pattern of choices be rational? There is at least one theory of choice which implies an answer of 'Yes': *regret theory* (Bell, 1982; Loomes and Sugden, 1982, 1987). The fundamental idea behind this theory is that the psychological experience of 'having x' can be influenced by comparisons between x and the y that one might have had, had one chosen differently. If, for example, I bet on a horse which fails to win, I may experience something more than a reduction in my wealth: I may also experience a painful sense of regret arising out of the comparison between my current state of wealth

[11] The argument in the final three sentences of this paragraph derives from Broome (1990). Broome, however, argues that 'non-practical' preferences – preferences that do not bear on any conceivable choice problem – can be interpreted as judgements about overall goodness, where such judgements are seen not as subjective mental states but as issuing from rational deliberation about what really is the case.

and the state that I would have enjoyed, had I not bet. On a Humean view of rationality, regret is just another kind of passion, to which reason must be a slave: there is no sense in which the feeling of regret can be called reasonable or unreasonable.

If the possibility of regret is admitted, then the experiences associated with choosing an act may depend, not only on the nature of the act itself, but also on the nature of other acts in the feasible set. Let $(f, \{f, g\})$ stand for the state of having chosen f from the set $\{f, g\}$. Then if Cathy's choices are an instance of the kind of cyclical preference that regret theory permits, we can say that she prefers $(f, \{f, g\})$ to $(g, \{f, g\})$, that she prefers $(g, \{g, h\})$ to $(h, \{g, h\})$, and that she prefers $(h, \{h, f\})$ to $(f, \{h, f\})$. When Cathy's choices are described like this, there does not seem anything obviously inconsistent about them.

It might be objected that Cathy's choices are inconsistent with Savage's axioms only because I have interpreted 'consequences' and 'acts' too narrowly. I have assumed that it matters to Cathy, not only what she gets, but also what she fails to get. If *anything* that may happen to a person can count as a consequence, why cannot we treat the conjunction of 'what Cathy gets' and 'what she fails to get' as the description of a consequence? If we could define consequences in this way, then an entity such as $(f, \{h, f\})$ could be described as a list of consequences, one for each state of the world, and thus would constitute an act in the Savage sense. And then there would be no violation of transitivity.

The problem is that this cannot be said without rejecting Savage's concept of a consequence, and with it, Savage's expected utility theorem. 'Getting x and regretting not having chosen an option that would have given y' is a description of a state of affairs that includes a reference to a feature of the choice problem in which that state of affairs is embedded, and so is not a consequence in Savage's sense. To put the same point another way, an entity such as $(f, \{h, f\})$ is not an act in Savage's sense, because we cannot construct meaningful choice problems out of arbitrary pairs of such entities. (Consider what it could possibly mean to face a choice between, say, $(f, \{h, f\})$ and $(f, \{g, f\})$ when g and h are different acts.)

The essence of the problem is this: the appeal of the transitivity axiom depends on our being able to say that the description of a consequence includes everything that is relevant for determining a person's preferences. If some feature (like regret) that might be relevant is left out of the description, and if acts are then defined in terms of consequences so described, we have no good reason to expect the preferences of a rational person to satisfy transitivity. But Savage's definition of a consequence imposes restrictions on what we are permitted to include in the description of a consequence, and these can prevent us from including some relevant features. The implication seems to be that, within Savage's theory, the transitivity axiom cannot be defended as a necessary property of rationality.

V. RATIONALITY IN GAMES

Savage's theory of rational choice is designed for problems in which the occurrence or non-occurrence of an event is independent of the nature of the acts among which a person has to choose. (Recall that in Savage's framework, the set of conceivable events and the set of conceivable consequences are defined independently of one another. Thus the description of an event can make no reference to any particular consequence, act, or set of feasible acts.) Such problems are 'games against nature': it as if the state of the world that obtains is chosen by some natural force which takes no account of the chooser's preferences, opportunities or actions. Some deep problems are raised when we try to adapt this theory so that it applies to the choices of rational individuals who are playing games against other rational individuals.

To keep the discussion simple, I shall look only at two-person games. (What I shall say can easily be generalised to *n*-person games, but the discussion would be messier.) I shall follow the traditions of game theory in dealing only with *one-off* games, that is, games that are played once only, so that the players of a game cannot be guided by previous experience of playing that game. I shall confine myself to *non-cooperative* games, that is, games in which the players are unable to communicate with one another, except through their actions in the game. And to avoid the complications introduced by the concept of 'correlated equilibrium' (Aumann, 1987), I shall assume that the players have no way of correlating their strategy choices. I shall ask how, as rational agents, the players of such games ought to act.

Is this the question that game theory is designed to answer? This is not clear. Kohlberg (1989), for example, asks rhetorically whether we can predict what rational players will actually do in a game, and replies: 'Except when the game is unusually simple... our answer must be "No!"'. He then goes on to offer an understanding of equilibrium which detaches it from the question of what rationality requires of players. Other theorists, as I shall show, take the opposite view. In any case, the question that I have posed is surely a significant one.

Game theorists start from a *mathematical description* of a game, which lays out the options faced by the two players and the consequences of every possible combination of choices. Consequences are described in terms of the utilities of the players. For the present, I shall assume that the game is described in the *normal form* (that is, by means of a list of strategies for each player, and a pair of utility indices for every combination of strategies that might be played). In analysing games, game theorists make much use of the concept of *common knowledge*, which derives from Lewis (1969). A proposition is common knowledge if each player knows it to be true, each knows that the other knows it to be true, and so on.[12]

[12] This, I think, is the usual understanding of 'common knowledge' in game theory. Lewis (1969, p. 56) gives a less restrictive definition framed in terms of reasonable beliefs: x is common knowledge if there exists some state of affairs A such that (i) each player has reason to believe that A holds; (ii) A indicates to each player that (i) is true; and (iii) A indicates to each player that x is true. Thus A might be the state in which the relevant players are in eye-to-eye contact when some event occurs which no one in their situation could fail to notice; x might be the proposition that the event has occurred.

The following assumptions are standard in game theory:

1. The mathematical description of the game is common knowledge.
2. Each player is rational in the sense of expected utility theory, treating the strategies that his opponent might choose as events in the Savage sense; and this is common knowledge.
3. Every logical or mathematical theorem that can be proved about the game is common knowledge.

I shall call these three assumptions *common knowledge of rationality* (CKR). Notice that Assumption 2 is not strictly compatible with Savage's axiom system. For Savage, the description of an event can make no reference to any act, but the 'event' that one player's opponent plays a particular strategy in a particular game cannot be described without reference to the game itself, and hence to the first player's set of feasible acts. Nevertheless, it is standard practice in game theory to use expected utility theory, and even to appeal to Savage for intellectual support in doing so. This practice leads to serious problems.

VI. NASH EQUILIBRIUM

Common knowledge of rationality implies a very important feature of game-theoretic reasoning. Suppose that one player arrives at some conclusion by logical deduction from the mathematical description of the game and from the truth of CKR. Then his opponent must know that he has reached this conclusion (and, indeed, it must be common knowledge that he has reached it.) This property is what Bacharach (1987) calls *transparency of reason*.

From this property, we can derive another important implication. (The germ of this idea can be found in von Neumann and Morgenstern (1947, pp. 146–60); Bacharach (1987) gives a formal proof.) Let us say that a game has a *unique solution* if we can show, using some set of premises which include CKR and which are common knowledge between the players, that for each player there is a particular strategy (pure or mixed) which must be chosen. Then these strategies must be best replies to one another: they must constitute a *Nash equilibrium*. The idea is simple. Suppose we can show that A must play the strategy S_A and that B must play S_B. Since all theorems are common knowledge, and since the players have access to the same premises as we have used, A must be able to predict that B will choose S_B. As a rational agent, A must choose a strategy that is a best reply to S_B. So if S_A is uniquely rational, S_A must be a best reply to S_B. Similarly, S_B must be a best reply to S_A.

So if a game has a unique solution, this solution must be a Nash equilibrium. Notice, however, that this does not imply that if a game has a unique Nash equilibrium, that equilibrium constitutes the unique solution of the game. In order to reach that conclusion, we need an additional premise: that the game has a unique solution. Such a premise might be seen as a game-theoretic analogue of completeness, expressing the idea that a theory of rational choice must be determinate.

One reason for scepticism about any such premise stems from the existence of games in which there are no pure-strategy equilibria. Consider the game

Table 1

Matching Pennies

		B's strategy	
		Heads	Tails
A's strategy	Heads	2, 0	0, 2
	Tails	0, 1	1, 0

shown in Table 1, which is a variant of Matching Pennies. This game has a unique Nash equilibrium, in which each player plays 'heads' with probability 1/3. So if the game has a unique solution, this is what the solution must be. The standard interpretation of a mixed strategy is that the player in question consciously makes use of some random device. In this case, we might imagine that each player decides to roll a six-sided die, and then plays 'heads' if and only if the die shows one or two. But the supposition that this pattern of behaviour is uniquely rational implies a contradiction. It is a property of expected utility theory that a probability-mix of two consequences can never be strictly preferred to both of those consequences. So there are simply no beliefs that *A* could hold that would make it uniquely rational for him to play 'heads' with probability 1/3. The only belief that would make this strategy rational at all is the belief that *B* will play 'heads' with a probability of exactly 1/3; and if *A* believes this, then any strategy, pure or mixed, is as good as any other.

Aumann (1987) suggests a re-interpretation of mixed strategies as a way of getting round this problem. We should, he says, interpret '*A* plays "heads" with probability 1/3' as a proposition, not about how *A* determines what he will do, but about *B*'s subjective beliefs about what *A* will do. On this interpretation, a Nash equilibrium is a relation that holds between beliefs and not between strategies. Aumann then makes the bold claim that if it is common knowledge that each player is rational in the Savage sense, then their beliefs must be in equilibrium. Since Aumann allows for the possibility that the players might be able to correlate their strategies, his claim is made in relation to the concept of correlated equilibrium. But if we assume that correlation is not possible, Aumann's argument implies that the players' beliefs must be in Nash equilibrium.[13] In the Matching Pennies game, then, each player must attach a probability of 1/3 to the event that the other player chooses 'heads'.

To see how Aumann's argument works, let the strategies open to *A* be $S_{A1}, \ldots, S_{Am}$ and let $\mathbf{p}_A = (p_{A1}, \ldots, p_{Am})$ be any probability distribution over *A*'s strategies. Let $S_{B1}, \ldots, S_{Bn}$ and $\mathbf{p}_B = (p_{B1}, \ldots, p_{Bn})$ be defined similarly. For Nash equilibrium, it is necessary and sufficient that if any strategy of either player has a non-zero probability, then that strategy must be optimal for that player in relation to the probability distribution over the other player's

[13] Binmore and Dasgupta (1985, pp. 2–10) adapt Aumann's argument in this way to provide a defence of Nash equilibrium.

strategies. Thus $(\mathbf{p}_A, \mathbf{p}_B)$ is a Nash equilibrium if and only if the following two conditions are true:

1. For all $i = 1, \ldots, m$: if S_{Ai} is not optimal for A, given $\mathbf{p}_B$, then $p_{Ai} = 0$.
2. For all $j = 1, \ldots, n$: if S_{Bj} is not optimal for B, given $\mathbf{p}_A$, then $p_{Bj} = 0$.

On Aumann's interpretation, this amounts to saying that if a strategy is not optimal for one player in relation to *his own* beliefs, then *the other player* must believe that it will not be played. For this to hold in general, each player's beliefs must be transparent to the other. (If a strategy is not optimal for A, given his beliefs, then it will not be played; but we need to explain how B, who does not have direct access to A's beliefs, comes to believe that it will not be played.) Aumann's claim is that the beliefs of rational players will be transparent in this way.

Aumann models each player as receiving some 'substantive information that leads him to make [whatever choice he makes]' (1987, p. 9). This information is private to that player, say B; A merely has a subjective probability distribution over all the possible descriptions of it. Prior to receiving this information, B also has a subjective probability distribution over all possible descriptions of it. And then comes the crucial assumption: these two probability distributions are the same. Aumann calls this the *common prior* assumption.

Given this assumption, we may interpret a probability such as p_{Ai} in two ways. First, it is the probability that B attaches to the event that A receives the information that leads him to play S_{Ai}. But it is also the prior probability that A attaches to the event that he receives the information that leads him to play S_{Ai}. This information could not lead A to play S_{Ai} unless that strategy were optimal for him, given the beliefs he holds about B, namely $\mathbf{p}_B$;[14] and A, being rational, must know this at the outset. Thus if S_{Ai} were not optimal for A, given $\mathbf{p}_B$, then A could not conceive of receiving any information that could lead him to play that strategy. That is, he would hold the belief $p_{Ai} = 0$. And this is Condition 1 for Nash equilibrium. Condition 2 can be derived in the same way.

What, we may ask, is this 'substantive information' which is playing such a crucial role in the argument? Aumann does not tell us. But it must somehow be capable of telling a player which of two strategies he should choose when he is indifferent between them.[15] Consider the Matching Pennies game. If Aumann is right, A must believe that, with probability $1/3$, B will play 'heads'. So A must be indifferent between playing 'heads' and playing 'tails'. But both A and B must believe that there is a probability of exactly $1/3$ that A will receive some information that will lead him to play 'heads' – even though he will still be completely indifferent between the two strategies. It seems we must

[14] In Aumann's theory, the information received by the two players may be correlated; thus A may have to revise his prior probabilities in the light of the information he receives. In adapting Aumann's argument to apply to Nash equilibrium, I am assuming that the information received by the two players is uncorrelated.

[15] An alternative interpretation is possible. Following Harsanyi (1973), we might assume that the payoffs as specified in the formal description of the game are not the true ones. The true payoffs for any player are given by the stated ones plus or minus some vanishingly small random disturbance terms, whose values are known only to that player. Aumann's 'substantive information' might then be identified with the values of those disturbance terms. This approach avoids the problem of mixed-strategy equilibria, but only by refusing to accept the reality of any payoff matrix which could imply such an equilibrium.

interpret this 'information' as some kind of psychological impulse which prompts A to make a particular choice. The common prior assumption requires that both players must attach the same prior probabilities to the different impulses that might act on A. From this we can deduce that both players must believe that the impulse that leads A to play 'heads' has a probability of $1/3$. So by pure deductive analysis, using no psychological premises whatever, we have come up with a conclusion about what rational players must believe about the properties of a psychological mechanism. This is an astounding result. Can it possibly be true?

In a logical sense, the result clearly is true: *if* there is common knowledge of rationality, and *if* there are common priors, then the players of the Matching Pennies game must believe that the impulse that leads A to play 'heads' has a probability of $1/3$. But the result can be read the other way round: if it is not the case that the players must hold this belief, then the conjunction of CKR and the common prior assumption is false.

We need to look more closely at the common prior assumption. Aumann's argument for this assumption is distinctly sketchy. The core of the argument consists of the claim that the common prior assumption 'expresses the view that probabilities should be based on information; that people with different information may legitimately entertain different probabilities, but there is no rational basis for people who have always been fed precisely the same information to do so' (1987, pp. 13–4). Despite Aumann's claims to the contrary, it is hard to see how this position can be compatible with Savage's subjective conception of probability. (Recall Savage's remark that two reasonable individuals faced with the same evidence may have different degrees of belief in the same proposition.) More seriously, the argument fails to justify the common prior assumption. It is true, of course, that reason cannot *require* two people who have received exactly the same information to draw different conclusions. But this leaves open the possibility that rational beliefs might be under-determined by evidence. Aumann's argument depends on the unstated assumption that, for any given information, there is a process of valid reasoning that generates a unique set of rational beliefs. To assume that such a process exists, in the absence of any demonstration that it does, is just as much an act of faith as the assumption that all games have uniquely rational solutions.[16]

VII. RATIONALISABILITY

Bernheim (1984, 1986) and Pearce (1984) have explored the implications of

[16] There have been some attempts to model the process by which rational game-players arrive at beliefs about one another. The best-known of these is probably Harsanyi's (1975) *tracing procedure*. A similar idea has been presented by Skyrms (1989). In these models, the players start with probability distributions over their strategies, these distributions being common knowledge, and then progressively revise them according to rules that are also common knowledge. If this process converges, it converges to a Nash equilibrium. However, neither Harsanyi nor Skyrms offers any real demonstration that a rational player will necessarily follow this particular type of algorithm. And we are still left with the problem of explaining how rational players arrive at the common priors from which their deliberations begin. If we were entitled to assume that *some* set of common priors existed, we might be entitled to appeal to symmetry and argue that each of a player's strategies should have an equal prior probability (Harsanyi and Selten, 1988, ch. 5); but we are not.

assuming common knowledge of rationality without common priors. Independently, they reach the same conclusion – that the only strategies that may be played are those that are *rationalisable*.

Consider a game between A and B. Suppose that A chooses some strategy S_A. If A is rational in the sense of maximising expected utility, he must attach some subjective probabilities to B's strategies, such that S_A is optimal for him.[17] But A knows that B is rational in the same sense. So if A attributes a non-zero probability to any strategy of B's, he must believe it to be possible that B would rationally choose to play this strategy. This could be true only if B were to assign some subjective probabilities to A's strategies such that the strategy in question would be optimal for her. But A knows that B knows that A is rational; and so each of A's strategies to which B has assigned a non-zero probability must be one that A could rationally choose to play. And so on indefinitely. A strategy is rationalisable if it can be supported by an infinite chain of beliefs of this kind.

This definition can be illustrated by using the Matching Pennies game. Every strategy in this game is rationalisable. For example, consider the strategy 'heads' for A. This strategy is optimal for A if he assigns a probability of at least $1/3$ to B's playing 'heads'. It would be optimal for B to play 'heads' if she assigned a probability of at least $2/3$ to A's playing 'tails'. It would be optimal for A to play 'tails' if he assigned a probability of at least $2/3$ to B's playing 'tails'. And so on.

So we can prove that only rationalisable strategies will be played. But if, as in Matching Pennies, a player has more than one rationalisable strategy, what is their status? Bernheim and Pearce seem to be saying that which strategy A chooses will depend on the subjective probability judgements he makes about B. Any such judgements are permissible, provided they do not attach non-zero probabilities to non-rationalisable strategies. In this respect, Bernheim and Pearce are allowing their theory to be non-determinate. But – and this is crucial – A must somehow commit himself to one particular probability distribution over B's strategies. For if we were to say that A cannot assign subjective probabilities, and thus that he cannot make a rational choice between rationalisable strategies, we would be rejecting the assumption that the players are rational in the expected-utility-maximising sense required by CKR. By assuming CKR, Bernheim and Pearce are building Savage's form of determinacy into their theory.

How, we may ask, does A come to commit himself to a probability distribution over B's strategies? We seem to have no way of explaining how A could arrive at any probabilistic belief about B. Knowing that B is rational, A knows that B's action will be determined by the probabilities she attaches to A's actions. Thus if A is to predict B's behaviour, he must first predict her probability judgements. But given CKR, this leads to an infinite regress. It seems that A cannot find any ground for holding one probability judgement rather than another. The infinite chain of reasons which 'rationalises' a

[17] Throughout the paper, when discussing games between two players A and B, I shall make A male and B female.

particular act on *A*'s part is internally consistent; but it cannot *explain* why *A* chose that act, or was convinced by those reasons. For that, we need to find the end of the chain; and there is none. This is essentially the argument that Ken Binmore (1987, 1988) uses to question whether CKR is a meaningful assumption.

It might be objected that such indeterminacy is an inevitable consequence of adopting a subjective view of probability. Clearly, there is a sense in which any subjective probability is ungrounded. But there is a fundamental difference between subjective beliefs about states of nature and subjective beliefs about the actions of perfectly rational agents. Consider an example involving states of nature: what is the probability that there will be rain in Norwich, ten days from now? To help us in making forecasts, we have climate records going back over many years. We also have information about current air pressures, temperatures, rainfall and so on, and a theory of meteorology which tells us how these variables change over space and time. Because this theory is incomplete and in some respects open to dispute, expert meterologists may disagree about the probability of rain in ten days' time. In this sense, any probability judgement will be subjective. Nevertheless, we can recognise certain principles of scientific method which we would expect a meterologist to use in arriving at judgements. The meteorologist could give an account of how she arrived at her judgements, even if those judgements were not the same as those of other experts.

Savage's theory, of course, tells us nothing about how we should form probability judgements about states of nature: that is not its function. Its function is to provide a definition of such judgements in terms of preferences over acts, and to impose conditions of consistency on them. But the assumption that a rational person holds subjective beliefs about states of nature might be defended by an appeal to well-established principles of inductive and scientific reasoning. But would this defence carry over to subjective beliefs about the behaviour of an opponent in a game in which there was CKR?

The problem is that CKR is a theoretical concept, not an empirical one. We are not entitled to assume that any observable game between real people is played under conditions of CKR. (Of course, were we to construct a theory of game-playing on the assumption of CKR and were we then to find that it predicted behaviour that was sufficiently similar to that of real game-players, we would be entitled to use that theory in an 'as if' way to make further predictions about real games. But that is jumping ahead of our present concern, which is with the construction of the theory itself.) So we cannot appeal to the methods of the empirical sciences, or to their informal analogues in everyday reasoning about empirical questions, in order to explain how beliefs are formed in games in which CKR is true. Any such explanation must work by deduction from the premises of the theory. And those premises are inadequate for the purpose.

What all this suggests is that in the context of games like Matching Pennies, CKR is an incoherent assumption. CKR requires that each player form subjective probabilities about his opponent's decisions, but if CKR is true,

there may be no way in which such probabilities can be formed. At first sight, the assumption of CKR seems to present a natural extension of Savage's theory of rationality: the idea of treating the actions of a rational opponent as if they were states of nature seems harmless enough, even though it is not strictly compatible with Savage's axiom system. But, I think, it is a fundamental error.

VIII. BACKWARD INDUCTION

The assumption of CKR leads to more problems in the case of sequential games (that is, games in which the players move in some sequence.) Consider the following game. First A chooses whether to stop the game (move S_1). If he does this, no payments are made. If he chooses to let the game continue (move C_1), he must pay £1 to the bank; the bank then pays £10 to B, and it is B's turn to move. Now she can choose to stop the game at this point (move S_2). If she lets the game continue (move C_2), she must pay £1 to the bank; the bank pays £10 to A and it is A's turn to move again. Again, A can choose whether to stop the game (move S_3) or to pay £1 to the bank, in which case the bank pays £10 to B (move C_3). But whatever A does at this point, this is the end of the game. I shall assume that each player's utility is linear in his or her own wealth. This gives the game shown in Fig. 1. This is a simple version of the Centipede game invented by Rosenthal (1981) and discussed by Binmore (1987).

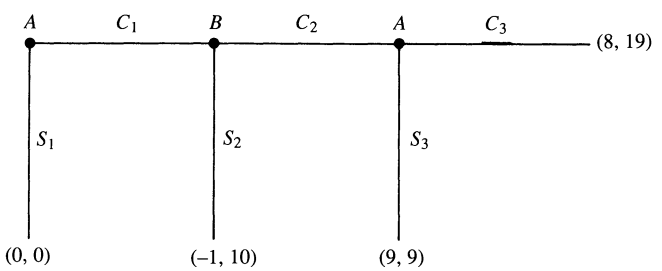

Fig. 1. The Centipede game.

There is a well-known argument that, given CKR, A must stop the game at his first move. The argument works by *backward induction*. Suppose the third node of the game is reached, so that A is called on to choose between C_3 and S_3. The A, being rational, must choose S_3 (which gives him a utility level of 9) rather than C_3 (which gives him 8). This gives us the proposition P3: 'If the third node is reached, A will play S_3'. By CKR, the truth of P3 must be common knowledge between A and B. Now suppose that the second node is reached. B knows that P3 is true, so if she were to play C_2, she would do so knowing that A would then play S_3. But then it cannot be the case that she would choose C_2 (which would give her 9), since S_2 (which gives her 10) is clearly preferable. This gives us the proposition P2: 'If the second node is reached, B will play S_2'. By CKR, P2 must be common knowledge. Finally, consider the first node of the game. A knows that P2 is true, so if he were to play C_1, he would do so knowing that B would then play S_2. But then it cannot be

the case that A would play C_1 (and get -1) rather than S_1 (and get 0). This gives us the proposition P1: 'A will play S_1'.

It is important to recognise that P1 does not falsify P2 or P3. Formally, this is because a conditional proposition with a false antecedent is true. More intuitively, if there really is common knowledge of rationality, then it cannot be the case that an irrational move is played. In deriving P3, we make the supposition that the third node is reached. Given CKR, this must be understood as the supposition that the third node is reached by rational play. P3 says that *if* the third node is reached by rational play, A will play S_3. We then use this proposition as part of a proof that the third node cannot be reached by rational play. And similarly for P2 and the second node.

But, as many commentators have noticed, this leaves us with a puzzle (Reny, 1986; Binmore, 1987; Bicchieri, 1989; Bonanno, 1989; Pettit and Sugden, 1989; Basu, 1990). We seem to have proved that A, as a rational agent, must play S_1. But, given our interpretation of rationality in terms of expected utility theory, to say that rationality requires A to play S_1 is to say that A would get a greater expected utility by playing S_1 than by playing C_1. And given CKR, this can never be shown to be true. If A asks 'What would happen if I were to play C_1?' there is no answer. If we were to try to answer it, we should have to make some prediction about what B would do, having observed that A had played C_1. But we have no basis for such a prediction, since we have already proved that A's playing C_1 is an event that will not occur, and we know that this proof is known to B. In order to make a rational decision about what to do in this event, B would first need to explain the event itself; and given CKR, no explanation is possible. The correct conclusion to draw, I suggest, is that it cannot be the case that the game is as described and that there is common knowledge of rationality. And since it seems clear that two players *could* face the game as described, the implication is that there are some games for which CKR cannot be assumed.

One way of trying to escape this problem is to use the concept of belief rather than knowledge. To say that B *believes* X is to say that B attaches a subjective probability of zero to the event that X is false. This allows us (as outsiders) to ask how B will respond if, despite this initial belief, she comes to know that X is false: if she is rational, she will presumably revise her beliefs in a way that makes them consistent with what she then knows. If we assume that B starts out, not with the knowledge but with the belief that A is rational, we can meaningfully pose the question of what it would be rational for her to do in the event that A played an irrational move.

It is possible to keep the spirit of CKR while substituting belief for knowledge. Consider the concept of *entrenched common belief in rationality* (ECBR), which is defined by the following four assumptions:

1. Each player knows the mathematical structure of the game.
2. Each player is rational in the sense of expected utility theory, treating the strategies that his opponent might choose as events in the Savage sense.
3. Each player knows the truth of every logical or mathematical theorem that can be proved about the game.

4. Properties 1–3 are matters of common belief at the start of the game; they remain matters of common belief as long as the hypothesis that they are matters of common belief can be maintained without inconsistency; that this is so is itself a matter of common belief.

A proposition is a matter of common belief if each player believes it to be true, each player believes the other believes it to be true, and so on. ECBR embodies the same concept of rationality as does CKR, but this rationality is made a matter of common belief rather than common knowledge. This common belief is 'entrenched' in the sense that it will be maintained as long as it is not contradicted by anything else the players *know* to be true.[18] This ensures that the common belief in rationality will remain intact throughout any game. (Since both players *are* rational, neither will ever do anything that could require the common belief in rationality to be revised.) I present ECBR as a suggested reconstruction of what game theorists have intended (or should have intended) when speaking of 'common knowledge'.

Returning to the Centipede game, it is easy to show that ECBR implies the propositions P_3, P_2 and P_1, just as CKR does. (The proof is almost exactly as before.) So ECBR implies that A will play S_1. This in turn implies that A's beliefs must be such that his expected utility from playing S_1 is at least as great as that from playing C_1. But what is the basis for this restriction on A's beliefs?

Suppose A asks himself what would happen, were he to play C_1. C_1's being played is inconsistent with ECBR. So after observing C_1, B can no longer rationally believe in the truth of ECBR. She must now conclude *either* that A is irrational, *or* that A believes that she is irrational, *or* that A believes that she believes that he is irrational. I can see no grounds for claiming that, in these circumstances, it would be irrational for B to form the belief that A is irrational. And if this would be a rational belief on B's part, I can see no reason for claiming that it would be irrational for A to believe that this is the belief that B would form.

So suppose that A does believe this. Now let p stand for A's estimate of the subjective probability that B attaches to A's playing C_3 if the third node is reached. Of course, C_3 is an irrational move; but we are considering a case in which A is believed to be irrational. If we are to ask what value p should take, we must ask what rational players should believe about the behaviour of irrational ones; and this requires a theory of irrational behaviour. None of the ingredients for such a theory has been introduced into our assumptions: our premises have all been about the behaviour and beliefs of rational players. We seem therefore to have no grounds for imposing any restrictions on the value of p. Nevertheless, we can prove that $p \leqslant 0{\cdot}1$. (Suppose that $p > 0{\cdot}1$. Then if the second node is reached, B maximises expected utility by playing C_2, and so it is optimal for A to play C_1. But we know that A plays S_1.) We seem to have

[18] The idea of entrenchment has some similarities with Kohlberg and Mertens's (1986) concept of 'forward induction'. It implies that the common belief in rationality will be retained even if an event occurs to which a player had assigned a zero subjective probability, provided only that that event is not inconsistent with the common belief.

succeeded in discovering a property of the theory of irrational behaviour, using only premises about rationality. How can this be possible?

The answer is that we have proved only a conditional proposition: *if* ECBR is true, *then* the theory of irrational behaviour must have certain properties. In other words: to assume ECBR is to make some implicit assumptions about irrational behaviour.[19] This merely echoes the conclusion of Section VII: we are not entitled to regard common knowledge of rationality, or common belief in rationality, simply as conceptions of rationality, and to treat them as fundamental assumptions of the theory of games. Their status is more like that of an equilibrium condition in a model for which there is no general proof that equilibrium exists.

IX. IS RATIONALITY SELF-DEFEATING? GAMES OF COORDINATION

A recurring theme in philosophical discussions of rational-choice theory is that the theory can be self-defeating. Rationality, it is said, is viewed as a matter of choosing efficient means for the attainment of given ends; but there are situations in which a person who acts on the prescriptions of the theory will do less well than one who does not.

Suppose A and B are playing a *coordination game* of the kind discussed by Schelling (1960). Each player must choose either 'left' or 'right'. If both choose 'left', or both choose 'right', each gets a payoff of 10 units of utility; otherwise, each gets a payoff of zero. (Imagine them to be drivers approaching one another on a narrow road; each must decide in which direction to steer in order to avoid a collision. I shall call this game Narrow Road.) To make things easier for the players, suppose they are able to discuss this problem with one another in advance, even though they are unable to make binding agreements. It is tempting to say that this makes the problem trivial: it seems obvious that they should agree that both will choose 'left' (or that both will choose 'right' – it makes no difference) and that each should then fulfil his or her part of the agreement. But can we show that this is *rational*?

Hodgson (1967) considers this type of problem from the viewpoint of act utilitarianism, which prescribes that each person should perform the act which maximises the sum of all persons' expected utilities. (Since we are analysing a pure coordination game, this is equivalent to the act which maximises each person's expected utility.) Hodgson argues that if it is common knowledge that A and B are act utilitarians, neither will have any reason to act on any agreement, and so they will be unable to make agreements in any meaningful sense. Consider A. The mere fact that he has agreed to play 'left' is no reason for him to play 'left': the only reason for action that an act utilitarian recognises is the maximisation of expected utility. (Act utilitarians are not bound by promises when greater utility could be achieved if those promises

[19] Selten (1975) presents a theory of irrational behaviour in which breakdowns of rationality (or 'trembles') occur in a particular random way. This theory is consistent with ECBR. However, as many commentators (e.g. Binmore, 1987) have noted, this theory is not particularly convincing at the psychological level.

were broken.) In deciding what to do, *A* must ask which action will maximise expected utility. This depends on what *B* can be expected to do: *A* should play 'left' if *B* is more likely to play 'left' than 'right'. But *A* knows that *B* is an act utilitarian too, and so the fact that she has agreed to play 'left' is no reason for her to play 'left'. This leads to an infinite regress: we cannot find an adequate reason for *A*'s playing 'left'.

The implication of this is that act utilitarianism is some degree self-defeating. If *A* and *B* are act utilitarians, they are both pursuing the same objective – the maximisation of the sum of their expected utilities. This objective will be achieved if they coordinate on playing 'left'. But because it is common knowledge that each is pursuing this objective, neither has a reason to play 'left'. In contrast, suppose that *A* believes, contrary to the prescriptions of act utilitarianism, that 'having promised to do *X*' is a reason for doing *X*. Suppose also that *B* knows that *A* believes this. Now, irrespective of whether *B* is an act utilitarian or whether she shares *A*'s belief in the significance of promises, *A* and *B* will find no difficulty in agreeing to coordinate, and then in coordinating. By *not* trying to reach the act-utilitarian objective (and by being known not to be so trying), *A* is more successful in reaching it.

This argument can be recast so that it applies to standard rational-choice theory, in which each individual seeks to maximise his own expected utility and CKR holds. Since Narrow Road is a pure coordination game, the two players have the same objective. As before, neither has any reason to play 'left', even if both have previously agreed to do so. And so they may fail to achieve their common objective: even in such a simple problem of coordination, CKR causes paralysis. If, in contrast, either player is known to have an irrational belief in the binding force of promises, then coordination can be achieved without any difficulty.

Now consider another coordination game, which I shall call Red and White. *A* and *B* are in separate rooms and are not allowed to communicate. Each has a keyboard with a red and a white key, and is told to press one of the keys. If both press red keys, each player wins £10. If both press white keys, each wins £100. If one presses a red key and the other a white one, neither wins anything. It seems obvious that each should press the white key. But again, we cannot prove this as an implication of CKR. Which key *A* should press depends on his beliefs about which key *B* will press, and we get another infinite regress.

Finally, consider a variant of Red and White, which I shall call Heads and Tails. The two keys are the same colour, but are labelled 'heads' and 'tails'. If both players press 'heads', or if both press 'tails', each wins £100; otherwise, neither wins anything. Yet again, CKR implies nothing about what the players should do; there is another infinite regress.

Schelling (1960, p. 64) conjectures that most players of Heads and Tails would choose 'heads'. (A recent experiment involving almost exactly this game, but with smaller payoffs, has confirmed Schelling's conjecture: see Mehta *et al.* (1991).) Schelling argues that 'heads' has 'some kind of conventional priority' over 'tails', and that each player typically will have grounds for believing this convention to be a matter of common belief. This

item of common belief acts as a signal, pointing out 'heads' as the obvious solution to the game. In this sort of case, 'heads' is said to be *prominent* or *salient*; the outcome in which both players choose 'heads' is a *focal point*. Schelling would argue that a similar analysis applies to the previous two games. In Narrow Road, the fact that there has been an agreement to play 'left' is the signal that makes 'left' salient. In Red and White, the fact that 'white' is associated with a higher prize than 'red' is the signal that makes 'white' salient.

The difficulty with this line of argument is to explain how a rational player gets from the proposition '"Heads" is salient' to the proposition 'Therefore I should play "heads"'. If we assume only CKR, there seems to be no way in which this can be done. This is the conclusion reached by Gilbert (1989). As Gilbert puts it, 'rational agents as such cannot be expected to do their parts in a salient solution to a coordination problem' (p. 73); if human beings *are* guided by salience, 'this is not a consequence of their rationality' (p. 61). Lewis (1969, pp. 35–7) seems to agree. He argues that a salient course of action is one which has some non-rational appeal, so that 'people tend to pick the salient as a last resort, when they have no stronger ground for choice'. Similarly, he suggests, we have a non-rational tendency 'to repeat the action that succeeded before if we have no strong reason to do otherwise'. Although Lewis does not use this example, it might also be said that we have some non-rational tendency to tell the truth and to keep our promises, when we have no strong reason to do otherwise. For Lewis, such pieces of common knowledge about non-rational impulses work rather like the common priors in Harsanyi's tracing procedure and in Skyrms's model of Bayesian deliberation. They provide the ultimate grounding for the common belief that players will choose their salient strategies and thus cut off the infinite regress of reasoning that would otherwise occur.

The implication of this is that rational players can succeed in coordinating only because it is not common knowledge that they are perfectly rational: their success depends on a residual element of non-rational action. But as Heal (1978) argues, this conclusion does not correspond with our intuitive feeling that, in a quite uncomplicated sense, choosing the salient strategy is the rational thing to do. Even so, Gilbert's argument is logically correct. The implication is that if we are to show that it is rational to choose the salient strategy, we must appeal to some conception of rationality other than CKR. What could this be?

Game theorists sometimes imagine there to be a book of recommendations for playing games which is entirely authoritative – let us call it the *authoritative volume*. This volume provides a set of prescriptions for rational play; it is certified that these prescriptions are indeed requirements of rationality. The ultimate objective of game theory is to find out what must be written in the authoritative volume.

But we may ask: to whom is the volume addressed? One answer is that it is addressed to game-players *as individuals*. This would certainly be the natural interpretation of a volume of recommendations about how to play zero-sum games. (Think of the books of advice that are written for chess-players.) We

might think of game theory as the project of writing recommendations of this kind for games in general, including non-zero-sum ones. It is this project that seems to keep leading us into infinite regresses of reasoning.

An alternative approach, however, is to address the authoritative volume to *both players*. As an analogy, think of two people playing as a team in a game against another team. Suppose that (as in Bridge) the rules of the game require the members of a team to make decisions independently of one another, but each is concerned only about the outcome for the team. In this case, it is natural to think of a set of recommendations as being addressed to a team, rather than to a single player. (Compare books on Bridge with books on Chess.) The author of such recommendations will assume that both team-members will have access to them, and that each can be sure that the other will follow them.

The point of this analogy is that the players in a coordination game can be regarded as members of the same team in a game against nature. If we address our recommendations to both players as a team, there is no problem at all in producing a set of authoritative recommendations for simple coordination games. For the game of Red and White, for example, the recommendation is simply: 'Press the white keys'. This guarantees the best possible outcome for the team, and so it is the best recommendation that can be made. More generally, suppose that two people know they will be playing some coordination game with a structure like that of Heads and Tails, but do not know how the strategies will be labelled. Then the best recommendation we can give may be: 'If some combination of strategies is salient, choose those strategies'.[20]

It may be objected that this does not address the problem pointed out by Gilbert. We can recommend a certain combination of strategies as the best possible for the players, considered as a team. But why is it rational for either player, as an individual agent, to act on those recommendations? This question leads back to the usual infinite regress, since it is rational for one player to act on them only if he expects the other to do so. But perhaps the mistake is to think that the question needs an answer at all. If the players constitute a team, then a recommendation that is addressed to them as a team is just the kind of recommendation that they need. Conventional rational-choice theory starts from the presupposition that players are independent agents pursuing separate objectives, and that recommendations must be addressed to them as individuals. But this is not the only possible starting-point for a theory of rational choice. (This thought is pursued by Hollis (1990).)

Very roughly, we need a theory of rational choice that will allow us to move from propositions of the kind 'R is the rule which, if followed by both players, would give the best results' to propositions of the kind 'Therefore both players should follow R'. Theories of universalisation (see Section II) are obvious candidates here. Regan's (1980) theory of *cooperative utilitarianism* is an example.

[20] Crawford and Haller (1990) analyse an interesting class of repeated coordination games. In these games, the players are unable to make use of ideas of prominence because they lack a 'common language' with which to describe the games. Crawford and Haller's approach is somewhat similar to that suggested in this paper: their solutions can be thought of as optimal recommendations for the players, considered as a team.

This theory starts by making a distinction between 'cooperators' and 'non-cooperators'. In the language I have been using, cooperators conceive of themselves as members of a team, engaged in a joint enterprise; the recommendations of cooperative utilitarianism are addressed to such agents. The recommendation is that cooperators should follow that rule which, if followed by all cooperators, would produce the best consequences from a utilitarian point of view.

As I have said before, this line of thought threatens to subvert game theory. In the Prisoner's Dilemma, for example, a theory such as Regan's will recommend both players to cooperate. But if we are to explain how rational agents ever cooperate, even in the simplest of coordination problems, it seems that we need something beyond the standard conception of rationality.

X. IS RATIONALITY SELF-DEFEATING? GAMES OF COMMITMENT

Kavka (1983) presents a thought experiment which suggests another way in which conventional rational-choice theory may be self-defeating.[21] Imagine that an eccentric billionaire has set up the following problem for you. Today is Monday. On Tuesday afternoon you will be offered the option of drinking a toxin which will induce extremely unpleasant nausea for twenty-four hours, but will have no lasting consequences. You will be paid $1 million if and only if, at midnight on Monday, it is judged that you *intend* to drink the toxin on Tuesday afternoon. You will be interviewed by a panel of psychologists, equipped with a battery of lie-detection tests; this panel will judge whether you have the required intention. They will announce their decision at the end of the interview. If they have judged you to have the intention, the $1 million will immediately be paid into your bank account. The money will then be yours, even if you do not drink the toxin: it is enough that you were judged to have intended to do so. For the purposes of the problem, you must assume that you would much prefer to have both the $1 million and the toxin than to have neither. You must also assume that the panel's procedures are extremely reliable, so that you believe with virtual certainty that you will be paid the $1 million if and only if you actually intend to drink the toxin.

This Toxin Puzzle may seem fantastic, but it illuminates some important issues involved in real choices. Take the case first discussed by Hobbes (1651/1962, p. 108), where A and B would both benefit from an arrangement whereby A (at some cost to himself) performs a service for B at one date, and then B (at some cost to herself) performs a service for A at a later date. This can serve as a paradigm of economic interaction. It will be rational for A to perform if and only if he can be assured that, conditional on this performance, B will perform to. So if it is possible for B to give this assurance, it will be rational for her to give it. But *is* this possible? Hobbes argues that 'in the condition of mere

[21] This thought experiment has some similarities with Newcomb's Problem, which is discussed by Nozick (1969) and by the contributors to Campbell and Sowden's (1985) volume. In Newcomb's Problem, you face a super-intelligent being which can predict your behaviour. This scenario is more fantastic than the Toxin Puzzle, and has less obvious relationship to economics.

nature', *B* will not be able to provide adequate assurance to *A*. According to Hobbes, *A* needs to be assured that, when the time comes for *B* to perform, it will be in her interest to do so. *B* cannot provide this assurance, since it will *not* be in her interest to perform.[22] As modern game theorists would put it, *B* can offer only 'cheap talk'.

Or take another case. Suppose it would be in *A*'s interest to do something that would harm *B*. Were *A* to do this, *B* would then have the opportunity to retaliate in a way that would harm *A*; but this would impose further costs on *B*. (A familiar example in economic theory concerns two firms. *A* is considering entry into a market which is currently monopolised by *B*. If *A* enters, *B* may either cooperate with *A* in fixing prices, or launch a price war.) If *B* could convince *A* that she *would* retaliate, it would not be in *A*'s interest to perform the original harmful action. But can *B* make a convincing threat? Again, we may say that the words 'If you harm me, I shall retaliate' are merely cheap talk.

The analysis of problems like these has been a major issue in game theory, particularly as a result of the work of Selten (1975, 1978) and Kreps and Wilson (1982*a*, *b*). It is usual to think of these problems as being about the credibility or non-credibility of threats and assurances. The Toxin Puzzle is interesting because it sidesteps the issue of credibility. The specification of the problem ensures that if you form an intention to drink the toxin, it will be believed. (Because of the skill of the panel of psychologists, all true statements of intent are credible.) By making this assumption, Kavka poses a further problem: granted that your intention, if formed, will be credible, can you form it?

It is tempting to say that the assumption that intentions are transparent is completely unrealistic, irrelevant for the analysis of real-world threats and assurances. But I am not sure that this is so. We are animals with physiological processes that we cannot entirely control. As Frank (1988) argues, we send out many signals that we have not consciously chosen. Because of this, our intentions are not entirely opaque to others. So there is some point in exploring the implications of Kavka's assumption.

Kavka's conclusion is that if you are rational, you will not be able to get the $1 million. When Tuesday afternoon comes, you will have no reason to drink the toxin. This will be true irrespective of whether you have won the $1 million or not. Since you are rational, you will not drink the toxin.[23] But you already know all this on Monday. So on Monday, you cannot have a genuine intention to drink it.

Now suppose we accept this conclusion. Suppose the billionaire makes his eccentric offer to two people – rational Rachael and irrational Irene. Being

[22] The details of Hobbes's argument are analysed by Hampton (1986, pp. 80–96) and Kavka (1986, pp. 137–56).

[23] Notice that this is not a case of weakness of will, as in the case of Ulysses and the Sirens discussed by Elster (1979). 'Weakness of will' describes cases in which a person has good reason to do *X* rather than *Y*, but psychological or physiological forces act on him in such a way that he does *Y*. In the Toxin Puzzle, the obstacle to your drinking the toxin is not psychological but rational: you have no reason to do so.

rational, Rachael does not intend to drink the toxin. She tries to feign this intention, but (as she expected from the outset) she does not succeed. She gets nothing. Irene, in contrast, believes that having formed an intention to do something gives her a sufficient reason for doing it. She simply says to the panel of judges: 'I intend to drink the toxin'. She means what she says, because she believes that when Tuesday comes round, her previous intention will give her a sufficient reason to drink the toxin; and of course, the offer of $1 million gives her more than sufficient reason to form the intention itself. She is rightly judged to have the appropriate intention. Next day, Rachael tells Irene that she would be irrational to drink the toxin: surely she can see that, having already won the $1 million, her choice is now simply between nausea and no nausea? As she drinks the toxin, Irene has an obvious reply: 'If you're so smart, why ain't you rich?'

This reply deserves to be taken seriously. We can imagine Irene's form of irrationality paying dividends in much less bizarre circumstances. She will be able to make sincere threats and to give sincere assurances in cases in which Rachael can make only insincere ones. If intentions are not entirely opaque, this gives Irene a clear advantage in many economic relationships. (This, I take it, is the idea behind the old maxim that honesty is the best policy.)[24]

It seems, then, that in some cases the conventional theory of rational choice is what Parfit (1984) calls *self-effacing*. A theory is self-effacing if it recommends us to believe some other theory if we can. In the face of the Toxin Puzzle, Parfit would say, a rational agent should try to convince himself that on Tuesday it will be rational for him to act on the intentions he forms on Monday, even though really, it will not be. If he manages to convince himself of this, he will be able to win the $1 million. If not, then he has still done as much as it is possible for him to do to achieve his ends, and so he cannot be called irrational. Kavka (1978) presents a rather similar analysis of the 'paradoxes of deterrence'.

Parfit distinguishes between a self-effacing theory and a self-defeating one, arguing that the standard theory is *not* self-defeating or inconsistent. In contrast, Gauthier (1986) and McClennen (1985, 1990) have used examples like the Toxin Puzzle to argue that the standard conception of rationality is unsatisfactory. McClennen begins by endorsing what he calls a *pragmatic* criterion of rationality: a procedure for making choices is rational to the extent that it tends to promote the ends of the person who uses it. McClennen argues in this way for the rationality of what he calls *resolute* choice. A resolute person makes his plans on the assumption that he will follow them through. As long as no event occurs which he had not foreseen, he then acts on those plans. He does not re-evaluate them as time goes on; it is sufficient for him that the plans *were* optimal when they were made. Gauthier discusses a similar mode of reasoning, which he calls 'constrained maximisation', and argues that it is rational. Machina's (1989) analysis of dynamic consistency is rather similar to McClennen's and Gauthier's conception of rationality.

In the Toxin Puzzle, Rachael's objective is to achieve the best possible

[24] Hume (1740/1978, p. 501) makes a similar point when he argues that 'a man is more useful, both to himself and others, the greater degree of probity and honour he is endow'd with'.

outcome. If she adopts the procedure of resolute choice, she will form a plan to drink the toxin, carry it through, and win $1 million. If she acts on the conventionally-rational procedure (which we may call *forward-looking* rationality) she will not win the $1 million. So resolute choice is pragmatically more rational.

The difficulty with this approach is to explain how a self-consciously rational agent can endorse a pragmatic criterion of rationality. If Rachael is capable of elementary reasoning, she will of course be able to recognise the pragmatic advantages of the procedure of resolute choice. But can she adopt the procedure *for this reason*? Suppose Rachael were able to convince the panel of judges of her intention to drink the toxin, and so get the $1 million. Tuesday comes. Why should she drink the toxin? Her objective, we have supposed, is simply to get to the best possible outcome. *Now*, drinking the toxin leads to a worse outcome ($1 million plus a day of nausea) than not drinking it ($1 million). Having decided to be resolute for purely instrumental reasons, she now finds that the same instrumental reasons tell her to break her resolution. So if these really are her reasons, she has no reason to drink the toxin. Unfortunately, she can foresee all this on Monday.

Irene's advantage is that she does not act on a forward-looking conception of rationality. We might explain Irene's perspective in Kantian terms by saying that when she chooses to carry out her previous intentions rather than to pursue her current interests, she is acting on a self-imposed law. Or, if we think in Humean terms, we may say that she has a primitive desire to act on her past intentions, and that this desire provides her with a reason for acting on them. Whichever of these accounts we give, Irene is not being resolute because, from a forward-looking point of view, it pays to be resolute (even though it does so pay). She just *is* resolute.

All this leaves us with a curious stalemate. The standard, forward-looking conception of rationality gives us a pragmatic criterion for appraising actions and decision-making procedures. On this criterion, Irene's procedure is more successful than Rachael's. But as long as Rachael accepts this criterion, she cannot use the more successful procedure. Irene's ability to use the procedure arises from her rejection of the criterion that recommends it. There seems to be a parallel here with the family of problems that Elster (1983) discusses under the heading of 'states that are essentially by-products'. Such states 'can never... be brought about intelligently or intentionally, because the very attempt to do so precludes the state one is trying to bring about' (p. 43).

We might do better to leave the pragmatic argument on one side, and instead ask whether there is anything irrational in just *being* resolute. Is it irrational to have a desire to act on past intentions, just because they are past intentions? On the Humean view, desires are ultimately beyond rational appraisal, and so we have no grounds for asserting that Irene's desires are less rational than Rachael's. It is, I think, a feature of human nature that we do sometimes have a desire to act on our past intentions, to seek to fulfil plans that we made in the past and that were important to us then, even if they are no longer so now. (Michael Slote (1989) provides an insightful analysis of such

desires.) And if there are pragmatic advantages to being resolute, we may be able to provide an evolutionary explanation of how we come to have these desires.

Defenders of orthodox rational-choice theory may make the following reply: If Irene really does have a desire to act on her past intentions, then this ought to enter into the utility indices we assign to consequences. We have been assuming that '$1 million and no nausea' is preferred to '$1 million and nausea'; it is only on this assumption that a rational agent cannot win the $1 million. Isn't the resolution of the paradox that Irene has the opposite preference?

The problem with this reply is that it is incompatible with the standard interpretation of 'utility'. Irene does not have a preference for nausea; she has a preference for being resolute. So we need to include in the description of a consequence, some reference to past intentions. The relevant preference is between '$1 million and no nausea, as a result of failing to keep a resolution' and '$1 million and nausea, as a result of keeping a resolution'. But these descriptions of consequences include references to the choice problem in which they are embedded, and so are not consequences in the Savage sense (see Section IV above). We therefore cannot appeal to Savage's axioms in order to justify our assignment of utility numbers to consequences. Of course, no axioms are sacrosanct. But we should not think that resolute choice can easily be assimilated into standard rational-choice theory.

XI. CONCLUSION

I have focused on three features of the received theory of rational choice.

First, I have examined the philosophical foundations of expected utility theory, concentrating on the version that commands most support among modern theorists – that of Savage. I have argued that Savage's axioms are much stronger than can be justified merely by an appeal to an instrumental conception of rationality. There is no adequate justification for the requirement that preferences are complete. Nor is there for the transitivity and sure-thing axioms, once it is recognised how Savage's axioms impose restrictions on what can count as a consequence.

Second, I have looked at the assumption of 'common knowledge of rationality', which underlies game theory. I have argued that this assumption is not sufficient to lead rational players to Nash equilibria. More controversially, I have argued that there can be games for which this assumption is incoherent. The heart of the problem is that Savage's theory of rational choice effectively requires each individual to be able to assign subjective probabilities to all events; what it is rational to do then depends on those probabilities. In game theory, the possible acts of opponents are treated as if they were Savage events, so that what it is rational for A to do depends on the probabilities he assigns to B's actions. But at the same time, A is assumed to know that B is rational in the same sense, so that the probabilities A assigns to B's strategies depend on what it is rational for A to do. A related kind of circularity occurs in some sequential

games, such as Centipede. Here the problem is that A cannot work out what it is rational for him to do unless he can predict the consequences of playing each of his possible first moves. He cannot do this without first working out whether, on observing a particular move by A, B will infer that A is rational or irrational; but this depends on what it is rational for A to do. These circularities make 'common knowledge of rationality' an equilibrium concept. We can start with the supposition that a particular strategy is rational, and then deduce whether that supposition is internally consistent; but this procedure does tell us what it really is rational to do. Nor, in general, is there any guarantee that a player has *any* strategy about which this supposition can consistently be made. It is of the nature of equilibrium analysis that equilibrium may not exist.

Finally, I have looked at the suggestion that rational-choice theory is self-defeating. I have argued that individuals who are known to be fully rational in the conventional sense may be less successful in reaching their objectives than they would have been, had they been (and been known to be) less rational. A conventionally rational person may be less successful at solving problems of coordination than one who acts on universalisable maxims – even when the problem is to coordinate with someone who is conventionally rational. A person who is forward-lookingly rational as the conventional theory prescribes may do less well in interactions with others than one who believes that his past intentions provide him with reasons for present action. Such findings do not quite imply that the conventional theory is logically inconsistent. If we really are sure that the theory is correct, we may accept it as a twist of fate that our knowledge of its correctness sometimes works to our disadvantage. But if we are not so sure about the theory's claims in the first place, the paradoxical nature of such a conclusion may give us further cause for scepticism.

There was a time, not long ago, when the foundations of rational-choice theory appeared firm, and when the job of the economic theorist seemed to be one of drawing out the often complex implications of a fairly simple and uncontroversial system of axioms. But it is increasingly becoming clear that these foundations are less secure than we thought, and that they need to be examined and perhaps rebuilt. Economic theorists may have to become as much philosophers as mathematicians.

University of East Anglia

Date of receipt of final typescript: January 1991

References

Aumann, R. J. (1987). 'Correlated equilibrium as an expression of Bayesian ignorance.' *Econometrica*, vol. 55, pp. 1–18.

Bacharach, M. (1987). 'A theory of rational decision in games.' *Erkenntnis*, vol. 27, pp. 17–55.

Basu, K. (1990). 'On the non-existence of a rationality definition for extensive games.' *International Journal of Game Theory*, vol. 19, pp. 33–44.

Bell, D. (1982). 'Regret in decision making under uncertainty.' *Operations Research*, vol. 30, pp. 961–81.

Bernheim, B. (1984). 'Rationalizable strategic behavior.' *Econometrica*, vol. 52, pp. 1007–28.

—— (1986). 'Axiomatic characterizations of rational choice in strategic environments.' *Scandinavian Journal of Economics*, vol. 88, pp. 473–88.

Bicchieri, C. (1989). 'Self-refuting theories of strategic interaction: a paradox of common knowledge.' *Erkenntnis*, vol. 30, pp. 69–85.
Binmore, K. (1987). 'Modeling rational players: Part I.' *Economics and Philosophy*, vol. 3, pp. 179–214.
—— (1988). 'Modeling rational players: Part II.' *Economics and Philosophy*, vol. 4, pp. 9–55.
—— and Dasgupta, P., eds. (1985). *Economic Organizations as Games*, Oxford: Basil Blackwell.
Bonanno, G. (1989). 'The logic of rational play in games of perfect information.' Mimeo, University of California, Davis.
Broome, J. (1990). 'Can a Humean be moderate?.' University of Bristol Department of Economics Discussion Paper no. 90/269.
—— (1991). *Weighing Goods*, Oxford: Blackwell.
Campbell, R. and Sowden, L., eds. (1985). *Paradoxes of Rationality and Cooperation*. Vancouver: University of British Columbia Press.
Crawford, V. P. and Haller, H. (1990). 'Learning how to cooperate: optimal play in repeated coordination games.' *Econometrica*, vol. 58, pp. 571–95.
Elster, J. (1979). *Ulysses and the Sirens*. Cambridge: Cambridge University Press.
—— (1983). *Sour Grapes*. Cambridge: Cambridge University Press.
Frank, R. H. (1988). *Passions within Reason*. New York: Norton.
Gauthier, D. (1986). *Morals by Agreement*. Oxford: Oxford University Press.
Gilbert, M. (1989). 'Rationality and salience.' *Philosophical Studies*, vol. 57, pp. 61–77.
Hampton, J. (1986). *Hobbes and the Social Contract Tradition*. Cambridge: Cambridge University Press.
Hare, R. M. (1952). *The Language of Morals*. Oxford: Clarendon Press.
—— (1963). *Freedom and Reason*. Oxford: Clarendon Press.
Harsanyi, J. C. (1973). 'Games with randomly disturbed payoffs: a new rationale for mixed-strategy equilibrium points.' *International Journal of Game Theory*, vol. 2, pp. 1–23.
—— (1975). 'The tracing procedure.' *International Journal of Game Theory*, vol. 4, pp. 61–94.
—— and Selten, R. (1988). *A General Theory of Equilibrium Selection in Games*. Cambridge, Mass.: MIT Press.
Heal, J. (1978). 'Common knowledge.' *Philosophical Quarterly*, vol. 28, pp. 116–31.
Hobbes, T. (1651/1962). *Leviathan*. London: Macmillan.
Hodgson, D. H. (1967). *Consequences of Utilitarianism*. Oxford: Clarendon Press.
Hollis, M. (1990). 'Moves and motives in the games we play.' *Analysis*, vol. 50, pp. 49–62.
Hume, D. (1740/1978). *A Treatise of Human Nature*. Oxford: Clarendon Press.
Jevons, W. S. (1871/1970). *The Theory of Political Economy*. Harmondsworth: Penguin.
Kant, I. (1781/1896). *Critique of Pure Reason*, translated by F. Max Müller. London: Macmillan.
—— (1785/1949). *Foundations of the Metaphysics of Morals*, translated by Lewis White Beck. Chicago: University of Chicago Press.
Kavka, G. S. (1978). 'Some paradoxes of deterrence.' *Journal of Philosophy*, vol. 75, no. 6, pp. 285–302.
—— (1983), 'The toxin puzzle.' *Analysis*, vol. 43, pp. 33–6.
—— (1986). *Hobbesian Moral and Political Theory*. Princeton, N.J.: Princeton University Press.
Keynes, John Maynard (1936). *The General Theory of Employment Interest and Money*. London: Macmillan.
Kohlberg, E. (1989). 'Refinement of Nash equilibrium: the main ideas.' Mimeo, Harvard Business School.
Kohlberg, E. and Mertens, J.-F. (1986). 'On the strategic stability of equilibria.' *Econometrica*, vol. 54, pp. 1003–37.
Kreps, D. M. and Wilson, R. (1982*a*). 'Sequential equilibrium.' *Econometrica*, vol. 50, pp. 863–94.
—— and —— (1982*b*). 'Reputation and imperfect information.' *Journal of Economic Theory*, vol. 27, pp. 253–79.
Levi, I. (1986). *Hard Choices*. Cambridge: Cambridge University Press.
Lewis, D. K. (1969). *Convention: A Philosophical Study*. Cambridge, Mass.: Harvard University Press.
Loomes, G. and Sugden, R. (1982). 'Regret theory: an alternative theory of rational choice under uncertainty.' Economic Journal, vol. 92, pp. 805–24.
—— and —— (1986). 'Disappointment and dynamic consistency in choice under uncertainty.' *Review of Economic Studies*, vol. 53, 272–82.
—— and —— (1987). 'Some implications of a more general form of regret theory.' *Journal of Economic Theory*, vol. 41, pp. 270–87.
McClennen, E. F. (1985). 'Prisoner's dilemma and resolute choice.' In Campbell and Sowden (1985).
—— (1990). *Rationality and Dynamic Choice*. Cambridge: Cambridge University Press.
Machina, M. J. (1989). 'Dynamic consistency and non-expected utility models of choice under uncertainty.' *Journal of Economic Literature*, vol. 27, pp. 1622–68.
Marshall, A. (1920/1949). *Principles of Economics*, 8th ed. London: Macmillan 1st ed. 1890.
Mehta, J., Starmer, C. and Sugden, R. (1991). 'An experimental investigation of focal points in coordination and bargaining,' forthcoming in proceedings of Fifth International Conference on the Foundations of Utility and Risk Theory.
Nagel, T. (1970). *The Possibility of Altruism*. Oxford: Clarendon Press.

Neumann, J. von and Morgenstern, O. (1947). *Theory of Games and Economic Behavior*, 2nd ed. Princeton: Princeton University Press.
Nozick, R. (1969). 'Newcomb's Problem and two principles of choice.' In (N. Rescher *et al.* eds), *Essays in Honor of Carl G. Hempel* (Dordrecht: Reidel); abridged version reprinted in Campbell and Sowden (1985).
O'Neill [Nell], O. (1975). *Acting on Principle*. New York: Columbia University Press.
—— (1989). *Constructions of Reason: Explorations of Kant's Practical Philosophy*. Cambridge: Cambridge University Press.
Parfit, D. (1984). *Reasons and Persons*. Oxford: Clarendon Press.
Pearce, D. G. (1984). 'Rationalizable strategic behavior and the problem of perfection.' *Econometrica*, vol. 52, pp. 1029–50.
Pettit, P. and Sugden, R. (1989). 'The backward induction paradox.' *Journal of Philosophy*, vol. 86, pp. 169–82.
Ramsey, F. P. (1931). 'Truth and probability' in *The Foundations of Mathematics and Other Logical Essays*, London: Routledge and Kegan Paul.
Rawls, J. (1971). *A Theory of Justice*. Cambridge, Mass.: Harvard University Press.
Regan, D. (1980). *Utilitarianism and Co-operation*. Oxford: Clarendon Press.
Reny, P. (1986). 'Rationality, common knowledge and the theory of games.' Ph.D. dissertation, Princeton University.
Rosenthal, R. W. (1981). 'Games of perfect information, predatory pricing and the chain-store paradox.' *Journal of Economic Theory*, vol. 25, pp. 92–100.
Samuelson, P. A. (1947). *Foundations of Economic Analysis*. Cambridge, Mass.: Harvard University Press.
Savage, L. (1954). *The Foundations of Statistics*. New York: Wiley.
Schelling, T. C. (1960). *The Strategy of Conflict*. Cambridge, Mass.: Harvard University Press.
Selten, R. (1975). 'Re-examination of the perfectness concept for equilibrium in extensive games.' *International Journal of Game Theory*, vol. 4, pp. 22–55.
—— (1978). 'The chain-store paradox.' *Theory and Decision*, vol. 9, pp. 127–59.
Skyrms, B. (1989). 'Correlated equilibria and the dynamics of rational deliberation.' *Erkenntnis*, vol. 31, pp. 347–64.
Slote, M. (1989). *Beyond Optimizing*. Cambridge, Mass.: Harvard University Press.
Smith, M. (1987). 'The Humean theory of motivation.' *Mind*, vol. 96 (1987), pp. 36–61.
Stroud, B. (1977). *Hume*. London: Routledge and Kegan Paul.
Tversky, A. and Kahneman, D. (1986). 'Rational choice and the framing of decisions.' *Journal of Business*, vol. 59, pp. S251–78.
Williams, B. (1981). *Moral Luck*. Cambridge: Cambridge University Press.