The Nature and Scope of Rational-Choice Explanation

JON ELSTER

How do rational-choice explanations explain? What are their limits and limitations? I want to discuss these questions in three steps. In Section I the topic is the more general category of intentional explanation of behavior. Section II adds the specifications needed to generate rational-choice explanation. Section III considers more closely the power of rational-choice theory to yield unique deductions. In particular, this concerns the possible nonunicity and even nonexistence of optimal choice.

I. Intentionality

To explain a piece of behavior intentionally is to show that it derives from an intention of the individual exhibiting it. A successful intentional explanation establishes the behavior as an *action* and the performer as an *agent*. An explanation of this form amounts to demonstrating a three-place relation between the behavior (B), a set of cognitions (C) entertained by the individual, and a set of desires (D) that can also be imputed to him. The relation is defined by three conditions that form the topic of this section. First, we must require that the desires and beliefs are *reasons* for the behavior. By this I mean:

(1) given C, B is the best means to realize D.

E. Ullmann-Margalit (ed.), Science in Reflection, 51–65.

The presence of such reasons is not sufficient for the occurrence of the behavior for which they are reasons. An actor might be asked to shudder as part of a scene. Even with the requisite beliefs and desires, he might find himself unable to shudder at will. More importantly, even if the behavior does occur, the reasons do not suffice to explain it. The sight of a snake on the set might cause the actor to shudder involuntarily. This also holds if we assume that the actor is in fact able to shudder at will, viz. if his intention to shudder is preempted by the sight of the snake. We must add, then, a clause ensuring that his behavior was actually caused by his intention to behave in that way:

(2) C and D caused B.

The reasons, that is, must also be causes of the action which they rationalize.[1] To see why this is also insufficient, we must look into the ways in which beliefs and desires can act as causes. Consider a rifleman aiming at a target. He believes that only by hitting the target can he achieve some further goal that he values extremely highly. The belief and the desire provide reasons for a certain behavior, viz. pulling the trigger when the rifle is pointed toward the target. They may, however, cause him to behave quite differently. If he is unnerved by the high stakes, his hand might shake so badly that he pulls the trigger at the wrong moment. If he cared less about hitting the target, he might have succeeded more easily. Here the strong desire to hit the target acts as a cause, but not qua reason. To act qua reason, it would at the very least have to be a cause of the behavior for which it is a reason. Now consider Donald Davidson's well-known example:

> A climber might want to rid himself of the weight and danger of holding another man on a rope, and he might know that by loosening his hold on the rope, he could rid himself of the weight and the danger. This belief and want might so unnerve him as to cause him to loosen his hold, and yet it might be the case that he never chose to loosen his hold, nor did he do so intentionally.[2]

Here conditions (1) and (2) are fulfilled, yet the beliefs and desires do not cause the behavior qua reasons. The example differs from that of the rifleman in that the beliefs and desires of the climber cause the very same behavior for which they are reasons, but it is similar in that they do not cause it qua reasons. It is a mere accident that in the case of the climber they happen to cause the very same behavior for which they are reasons. Hence we must add:

(3) C and D caused B qua reasons.

As in other cases, we may ask by virtue of which features the cause produced its effect. When the falling of a stone leads to the breaking of the ice, we point to the weight of the stone, not to its color, to explain what happened. When the desire of the rifleman causes him to miss the target, we point to something like psychic turbulence or emotional excitement, not the strength of the desire. The latter reflects the agent's evaluation of the importance of the goal compared to other goals that he might entertain. Hence the strength of the desire is primarily relevant for its efficacy qua reason, and only to the extent that the desire causes behavior qua reason for the behavior is its strength also relevant for its causal efficacy. The emotional halo surrounding the desire is irrelevant for its efficacy qua reason but may influence its efficacy qua nonrational cause. To be sure, these are loose and metaphorical manners of speaking. We do not yet have a good language for getting emotions and their relevance for action into focus. Yet I take it that no one would deny the phenomenological reality of the facts I am describing, or the need for something like clause (3) in order to exclude a certain kind of accidental coincidence, just as clause (2) was needed to exclude another kind of coincidence.

Although these clauses would have to be satisfied in a fully satisfactory intentional explanation, we usually impose less stringent requirements. An analogy would be the detective story that proceeds by inquiring into motive and opportunity. When a person engages in a certain kind of behavior, we already know that he had the opportunity. If he did it, he could do it (in one sense of "could"). If, in addition, we find that he had a motive and also knowledge of the opportunity, we usually conclude that we have found an intentional explanation of the behavior, even if the kind of coincidences excluded by clauses (2) and (3) might conceivably have been operating. In some special cases we might want to reduce the likelihood of the first kind of coincidence by also establishing that the agent had the ability to perform the behavior in question, e.g., the ability to shudder at will or the ability to hit a target. While this does not fully eliminate the possibility of coincidence, it does so for most practical purposes. The point is that satisfaction of clauses (2) and (3) requires us to scrutinize the actual mental machinery at work, which is something we are only exceptionally able to do. By contrast, establishing motive, opportunity, knowledge, and ability is a much easier task (which is not to say that it is at all an easy one).

The nonsufficiency of clause (1) in establishing an intentional explanation is related to the difference between explaining and predicting action. If (1) were sufficient for explanation, we could also use it for prediction. There is, however, no regular lawlike connection between having certain desires and

beliefs on the one hand and performing a certain action on the other.[3] However, just as for practical purposes clause (1) goes a long way toward explaining behavior, one may with some practical confidence predict that motive, opportunity, etc., will result in action. The present paper, nevertheless, is mainly concerned with first-best explanation.

II. Rationality

Rational-choice explanation goes beyond intentionality in several respects. For one thing, we must insist that behavior, to be rational, must stem from desires and beliefs that are themselves in some sense rational. For another, we must require a somewhat more stringent relation between the beliefs and desires on the one hand, and the action on the other.

Minimally, we require that

(4) the set of beliefs C is internally consistent;

(5) the set of desires D is internally consistent.

One might think that these are required not just for rational-choice explanation but for intentional explanation more generally. If, for instance, there is *no* way of realizing a given desire, because it is internally inconsistent, how could anyone choose the *best* way to realize it? The answer, of course, is that the agent must believe that the desire is feasible. This belief, in turn, is internally inconsistent. For the belief that a certain goal is feasible to be consistent, there must be some possible world in which it is feasible. And that implies that there must be some further world in which it is realized, contrary to the assumption. Yet purposive action may spring from such inconsistent mental states. Someone may believe that the best way of trisecting the angle by means of ruler and compass is by first drawing a certain auxiliary construction. That drawing can then be explained in terms of the logically inconsistent goal of trisecting the angle in this way, and the belief that the goal is feasible and best attained by first taking that step. If this is not an intentional explanation, nothing is, but we might not want to call it a rational-choice explanation.

True, this example is controversial, because the implicit notion of rationality might seem to be too stringent. In fact, it seems to confuse irrationality with lack of mental competence. To this one may answer that while there need not be anything irrational in wanting to bring about a goal that happens to be logically inconsistent, rationality requires that we should

be aware of the possibility that it might be unfeasible. To believe, unconditionally, in the feasibility of a certain mathematical construction can be irrational, regardless of its actual feasibility. This, however, pertains to the well-groundedness of the belief, not to its internal consistency; I return to this issue below. There are, however, other and more clear-cut examples of actions deriving from internally inconsistent desires or beliefs. The belief "It will rain if and only if I do not believe it will rain" is logically inconsistent,[4] yet people might decide, on the basis of this belief, to bring their umbrella along for a trip across the Sahara. Also one may cite the less exotic phenomena of intransitive preferences, inconsistent time preferences, subjective probabilities over exhaustive and exclusive events that do not add up to 1, etc.[5]

One might want to demand more rationality of the beliefs and desires than mere consistency. In particular, one might require that the beliefs be in some sense substantively well grounded, i.e., inductively justified by the available evidence. This, to be sure, is a highly problematic notion; yet here I assume throughout that it is a meaningful one. The analysis of rational belief then closely parallels that of intentional action. Again there are three conditions to be satisfied:

(1b) the belief must be the best belief, given the available evidence;

(2b) the belief must be caused by the available evidence;

(3b) the evidence must cause the belief "in the right way."

Of these, the first condition presupposes some rather strong rule of inductive inference. The second is needed to eliminate the possibility that one has hit on the best belief merely by accident. It may be possible, for example, to arrive by wishful thinking at the belief which also happens to be the best.[6] The third condition is needed to exclude the possibility that by considering the evidence one might arrive at the belief which is in fact warranted by it — but by an incorrect process of reasoning. There could, for instance, be several compensating errors in the method of inference.[7] Once again, we may make the distinction between this first-best analysis of rational belief-formation and the less demanding condition that only (1b) be satisfied.

Given the satisfaction of (1b), (2b), and (3b), the belief is explained by its well-groundedness with respect to the available evidence E. One might want to make this part of the definition of rational-choice explanations:

(W) the relation between C and E must satisfy (1b), (2b), and (3b).

For reasons set out in Section III, this proposal is incomplete. It needs to be supplemented by a condition about how much evidence it is rational to collect.

Could one, similarly, demand substantive rationality of the desires? If so, what requirements would one want to impose on the rational formation of desires and preferences? Although I believe it possible to suggest the beginning of an answer to these questions, the results are not sufficiently robust to be reported here.[8] We do need, however, an additional condition on the relation between desires and behavior. This is designed to exclude akratic behavior, or weakness of the will.

Consider the man who wants to stop smoking and yet yields to temptation when offered a cigarette. In accepting it, he behaves in conformity with conditions (1) through (5). He desires to smoke: a perfectly consistent goal. He believes that he is offered a cigarette, not just a plastic imitation. Hence the best way to realize his desire is to accept it, which he does. This, however, gives only part of the picture. The account mentions that there are reasons for smoking but omits the reasons against smoking. When discussing intentional explanation, I implicitly used an existential quantifier: there exist a set of beliefs and a set of desires that constitute reasons for the action and that actually, qua reasons, cause it. But these need not be all the reasons there are. The agent may have a desire to stay in good health that would provide a reason for not accepting the offer. Moreover, he might think that this desire outweighs the immediate wish to smoke: all things considered, he had better reject the offer. And yet he might take it. To exclude such akratic behavior from being considered rational, we must add the following condition:

(6) given C, B is the best action with respect to the full set of weighed desires.

There are various accounts of how akratic behavior comes about. To my mind, the most plausible is offered by Davidson, who argues that it occurs because of faulty causal wiring between the desires and the action.[9] The weaker reason may win out because it blocks the stronger ones from operating; or the stronger reasons might lose because they cause a behavior other than that for which they are reasons. In either case, condition (1) fails to hold for the full set of desires. The action is intentional but irrational.

Is there a cognitive analogy to condition (6)? This would have to be part of condition (1b). By considering only part of the evidence, one might form a belief that is the best relative to that part but not the best relative to the whole evidence. A related, although different process is at work when one decides to

stop collecting evidence at the point where it favors the belief that, on other grounds, one wants to be true. I will return to this shortly.

III. Optimality

The explanatory force in condition (1) derives from the requirement that the explanandum be "the best" means to accomplish the agent's goal. The enormous success of rational-choice models in economics and other sciences is due to their apparent ability to yield unique, determinate predictions in terms of maximizing behavior. Although, generally speaking, explanation may take the form of elimination as well as determination,[10] the explanatory ideal in science is always to form hypotheses from which a unique observational consequence can be deduced. In this section I want to consider some difficulties with this view when applied to the social sciences. For one thing, there may be several options that are equally and maximally good; for another, there may be no "best" option at all. One might retort that these are nonstandard cases that, like the problems underlying conditions (2) and (3), only arise in rather perverse situations. This reply is not valid. There exists a strong general argument to the effect that uniquely maximizing behavior is in general not possible.

Consider first the nonunicity of optimal choice, arising because the agent is indifferent between several options than which none better. There is then no room left for rational choice; yet typically the agent will be able at least to "pick" one of the options.[11] A fully satisfactory theory would then offer a causal supplement to the rational-choice explanation by indicating how perceptual salience or some other value-neutral feature of the situation led to the "picking" of one option rather than another. Or, alternatively, one might redefine the choice situation by bunching the top-ranked alternatives into a single option. If I am indifferent between a red umbrella and a blue umbrella but prefer both to a raincoat, the choice becomes determinate once we have bunched the first two options as "an umbrella." This way out, however, may be unavailable if the top-ranked alternatives differ along more than one dimension, since then the indifference vis-à-vis the options could be due to offsetting virtues rather than to value-neutrality.

The presence of multiple optima can create a good deal of embarrassment. General-equilibrium theory, for instance, is not really able to cope with this problem. In the simplest version of this theory, all optima in production and consumption are assumed to be unique. Given some additional assumptions, one can then show that there is a set of prices that will allow all markets to clear when agents optimize. In the more complex version, multiple optima

are allowed. The equilibrium concept is correspondingly modified to mean the existence of a set of prices and a set of optimizing acts that allow market clearing.[12] The difficulty is not that the choice of these acts rather than other optimizing acts would be a pure accident. Rather it is that the indeterminacy is essential for the existence proof to go through. In the actual world, there is no indeterminacy. One optimum will always be chosen. Clearly, if one had a theory that explained which of the maximally good options is chosen (or picked), it would be an improvement over a theory which leaves this indeterminate. Yet it would destroy the existence proof by introducing a discontinuity in the reaction functions.

In game theory, multiple optima abound. In the wide class of noncooperative games that have an equilibrium point, many have equilibria that consist of mixed strategies. At any equilibrium point of mixed strategies any player has many optimal strategies, given that all the others choose their equilibrium strategies. In fact, any pure strategy or linear combination of pure strategies is as good as any other. Why, then, should a player choose the equilibrium strategy? John Harsanyi argues that the lack of any good answer to this question is a basic flaw in game theory as traditionally conceived. He proposes a substitute solution concept, according to which only "centroid" or equiprobabilistic mixed strategies are allowed. This corresponds to the idea that when there are several optima, one is chosen at random by "what amounts to an unconscious chance mechanism inside [the player's] nervous system."[13] This, of course, is essentially a causal concept.

Consider now the nonexistence of optimal behavior, which can arise in strategic as well as nonstrategic situations. A simple case obtains when an agent has incomplete preferences, so that for at least one pair of alternatives x and y it is neither true that he weakly prefers x to y nor that he weakly prefers y to x. If a pair of such noncomparable options are on the top of the agent's preference ranking, in the sense that for each of them it is true that there is none better, it will not be true that there is at least one alternative that is at least as good as all others. In actual cases it may seem hard to distinguish between incomparability and indifference, but the following test should help us. If there is an alternative (perhaps outside the feasible set) that is preferred to x, then it should also be preferred to y if the relation is one of indifference,[14] but this implication does not hold in cases of noncomparability.

As suggested by Sen and Williams, noncomparability may be especially important when our rankings are sensitive to the welfare of other people.[15] Assume that I have the choice between giving 10 dollars to one of my children and giving them to another. I may well find myself unable to decide and, moreover, find that I am equally unable to choose between giving 11 dollars

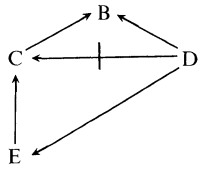
to the first and 10 dollars to the second, although I would rather give 11 than 10 to the first. This would indicate that I simply am unable to assess the welfare they would derive from the money in a sufficiently precise way to allow me to make up my mind. Yet decisions will usually be made (although in this case paralysis of action is perhaps more plausible than in some other cases[16]), so for their explanation we must look beyond rational-choice theory.

Preferences can be defined over outcomes or over actions. I shall assume that the latter are derived from the former, so that one prefers one action over another because one prefers the outcome it brings about.[17] I have just discussed the case in which the preferences among actions are incomplete because the corresponding outcome-preferences are. Action-preferences may, however, be incomplete even when the outcome-preferences are complete, viz. if one is in the presence of uncertainty. Observe first that in condition (1) the notion of "best" is to be taken in a subjective sense — "best" relative to the beliefs of the agent. This includes the case of probabilistic beliefs, in which to act rationally means to maximize expected utility. Sometimes, however, it is not possible to establish subjective probabilities on which one can rationally rely when making up one's mind. In decisions concerning nuclear energy, for instance, it seems pointless to ask for the subjective probability attached to the event that a given democratic country some time in the next millennium will turn into a military dictatorship that could use the reactor plutonium to make bombs.[18] And I believe the same problem arises in may cases of short-term planning as well. In decision-making under uncertainty it is only under very special conditions that we can pick out the top-ranked action. Specifically, this requires that there is one option such that its worst-consequence is better than the best-consequence of any other option.[19] Failing this, rationality is no guide to action, and a fortiori not a guide to explanation of action.

Nonexistence of optimal choice may also stem from the strategic nature of the situation. There are two cases: either there is no equilibrium point, or there are several equilibria none of which can be singled out as the solution.[20] The first can arise when the set of alternatives is unbounded or open. In the game "Pick a number — and the player who has picked the largest number wins" there is no equilibrium set of strategies because the strategy set is unbounded. Hyperinflation sometimes looks a bit like this game. In the game "Pick a number strictly smaller than 1 — and the player who has picked the largest number wins" there is no equilibrium point because the set is open. One may illustrate this with a variant of the game of "Chicken," in which the point is to drive at top speed toward a wall and then stop as close to it as possible.

More central, probably, are games that do have equilibria but no unique solution. The standard version of "Chicken" illustrates this concept. Here two players are driving straight toward each other, and the point is not to be the first to swerve. There are two equilibria, in each of which one driver swerves and the other does not, but there is no way in which rationality alone will help the players converge toward the one or the other. An example of this interaction structure could be some forms of technical innovation, characterized by "Winner takes all."[21] The individual firm will have little incentive to invest in R&D if other firms invest heavily and a strong incentive to do so if others do not. I want to insist that such cases illustrate the nonexistence of optimal choice rather than its nonunicity. When there are multiple equilibria, individual agents cannot toss a coin between the various equilibrium strategies attached to them. True, by coordinating their actions they might toss a coin between the full equilibrium strategy sets, but in that case we have left the domain of individual rationality with which we are concerned here.

The following diagram summarizes what has been said so far about nonunicity and nonexistence of optimal choice.



I have been arguing for the following phenomena: (i) nonunicity of optimal behavior, given D and C; (ii) nonexistence of optimal behavior, given D and C; (iii) nonexistence of optimal beliefs, given E. Here, D and E have been considered as given. I said above that I did not want to enter into the speculative question whether D could also be subject to rationality criteria, but we surely have to ask this question concerning E. How much evidence is it rational to collect before forming the belief on the basis of which one decides to act? Every decision to act can be seen as accompanied by a *shadow decision* — the decision about when to stop collecting information. The former can be no more rational than the latter, on which it is based, although it may well be less rational if some other things go wrong in the sequence.

In most cases it will be equally irrational to spend no time on collecting evidence and to spend most of one's time doing so. In between there is some optimal amount of time that should be spent on information-gathering. This, however, is true only in the objective sense that an observer who knew everything about the situation could assess the value of gathering information

and find the point at which the marginal value of information equals marginal costs. But of course the agent who is groping toward a decision does not have the information needed to make an optimal decision with respect to information-collecting.[22] He knows, from first principles, that information is costly and that there is a tradeoff between collecting information and using it, but he does not know what that tradeoff is.

> It is like going into a big forest to pick mushrooms. One may explore the possibilities in one limited region, but at some point one must stop the exploration and start picking because further exploration as to the possibilities of finding more and better mushrooms would defeat the purpose of the outing. One must decide on an intuitive basis, i.e., without actually investigating whether further exploration would have yielded better results.[23]

This argument does not imply that any decision about when to stop information-gathering is arbitrary. There will usually be many specific pieces of information that one knows it is worthwhile acquiring. One knows that in order to build a bridge there are some things one must know. These form a lower bound on information-collection. An obvious upper bound is that one must not spend so much time gathering the information that it becomes pointless. If one wants to predict the next day's weather, one cannot spend more than 24 hours gathering evidence. Sometimes the gap between the upper and the lower bound can be narrowed down considerably, notably in highly stereotyped situations like medical diagnostics. One then has a basis for estimating, with good approximation, the expected value of more information. In many everyday decisions, however, not to speak of military or business decisions, a combination of factors conspire to pull the lower and upper bounds apart from one another. The situation is novel, so that past experience is of limited help. It is changing rapidly, so that information runs the risk of becoming obsolete. If the decision is urgent and important, one may expect both the benefits and the opportunity costs of information-collecting to be high, but this is not to say that one can estimate the relevant marginal equalities.

The upper and lower bounds on information-collection are determined in part by the nature of the problem, in part by one's preferences. When one builds a bridge with profit as the objective and safety as the constraint, the bounds will differ from those when safety is the objective and profit is the constraint. There is nothing wrong, therefore, in the presence of a causal link between D and E, as drawn in the diagram. Note, however, that desires can determine the collection of information in another way, more related to wishful thinking. (Wishful thinking in the diagram is indicated by the line from D to C — blocked in order to indicate that this is not a proper causal

influence.) One may stop collecting evidence at the point where the sum total of the evidence collected until then favors the belief that one would want to be true. Sometimes this is clearly irrational, viz. if one is led to stop collecting evidence before the lower bound has been reached. But what if the wish for a certain belief to hold true leads one to collect an amount of evidence well between the lower and upper bounds? Imagine a general who is gathering information about the position of enemy troops. The information is potentially invaluable, but waiting to gather it exposes him to grave risks. He decides to attack when *and because* the net balance of information so far leads him (rationally) to believe that the enemy is highly vulnerable. I am not sure about this case, but I submit that his procedure is not irrational. The wish in this case functions merely as a heuristic device that allows him to make a decision. There is no reason to think that the causal influence of the wish tends to make the decision worse than it would have been had a different device been used.

In short, the only condition one can impose on E is rather vague:

(N) one should collect an amount of evidence that lies between the upper and lower bounds that are defined by the problem situation, including D.

Correspondingly, we may impose the following condition on the relation between evidence, belief, and desires:

(7) the relation between C, D, and E must satisfy (1b), (2b), (3b), and (N).

This concludes my account of rational-choice explanation.

IV. Summary

Ideally, a fully satisfactory rational-choice explanation of an action would have the following structure. It would show that the action is the (unique) best way of satisfying the full set of the agent's desires, given the (uniquely) best beliefs the agent could form, relative to the (uniquely determined) optimal amount of evidence. We may refer to this as the *optimality part* of the explanation. In addition, the explanation would show that the action was caused (in the right way) by the desires and beliefs, and that the beliefs were caused (in the right way) by consideration of the evidence. We may refer to this as the *causal part* of the explanation. These two parts together yield a

first-best rational-choice explanation of the action. The optimality part by itself yields a second-best explanation, which, however, for practical purposes may have to suffice, given the difficulty of access to the psychic causality of the agent.

It follows from Section III that even the second-best explanation runs into serious difficulties. It rests on three uniqueness postulates: unique determination of the optimal evidence, of the optimal beliefs given the evidence, and of the optimal action given the beliefs and the desires. Each of the links in the chain has been challenged, in the sense that both the unicity and the very existence of optimality have been shown to be problematic in certain cases. The most serious challenge arises at the level of information-gathering, since it will only exceptionally be possible for the agent to determine the marginal cost and benefit of more information. The challenge at the next level arises in cases of uncertainty, i.e., when the evidence does not allow any belief, even a probabilistic one, to be formed. Finally, the link from mental states to action was shown to be problematic, both with respect to unicity and with respect to existence.

Given that one or more of these links fails to yield a unique optimum, the explanation cannot take the form of determination; rather it must consist in eliminating some of the abstractly possible actions. At each level, it is possible to eliminate some of the options in the feasible set. The nature of the problem sets upper and lower bounds on the amount of information one should collect. In cases of uncertainty one should at least not choose an action that has worse best-consequences than the worse worst-consequences of some other action. In cases of indifference or noncomparability, one should not choose an option to which some other alternative is strictly preferred. In games without solutions it is less clear what options are eliminated.

Under the same assumption, the rational-choice explanation must be supplemented by a causal account. At the level of information-gathering one may refer to the fact that people have different aspiration levels. Some people spend ten minutes, others two hours looking for the best place for mushrooms. In decision-making under uncertainty one may invoke such psychological features as optimism or pessimism to explain why people choose maximax or maximin strategies. When the indeterminacy occurs at the level of action, the explanation may involve perceptual salience (in the case of indifference or noncomparability) or a desire for security (if the maximin behavior is chosen in games without solution).

Hence rational-choice explanation may fail because the situation does not allow a unique behavioral prediction from the hypothesis that agents behave rationally. But we should not forget that it sometimes fails simply because

people act irrationally. They yield to wishful thinking, in the sense of letting their desires determine their beliefs or the amount of evidence they collect before forming their beliefs (assuming that the result is below the lower bound). Or they succumb to weakness of will, in the sense of acting for the sake of a desire which they themselves value less highly than the remaining set of desires. Finally, their intentions and beliefs may be subject to various inconsistencies that are also incompatible with rational choice.

Let me point to a final consequence of this analysis. It has shown that there are many dimensions of latitude in the notion of rationality. Correspondingly, we get more degrees of freedom in our interpretation of other people. In trying to understand each other, we are guided and constrained by the idea that by and large others are as rational as ourselves. The slack in the concept of rationality implies that we are able to understand more, although it also implies that our understanding will be more diffuse.[24]

Notes

I thank Marcelo Dascal, Dagfinn Føllesdal, and Michael Root for their comments on an earlier version of this paper. It also appears in: *Actions and Events: Perspectives on the Philosophy of Donald Davidson,* ed. E. Le Pore and B. McLaughlin; Oxford: Blackwell, 1986.

1. Here, as elsewhere, my debt to Donald Davidson's work will be obvious.
2. Donald Davidson, *Essays on Actions and Events,* Oxford: Oxford University Press, 1979, p. 79.
3. Ibid., Chap. 11.
4. It is inconsistent because there is no possible world in which the belief is both true *and believed* (J. Hintikka, *Knowledge and Belief,* Ithaca, N.Y.: Cornell University Press, 1961).
5. Cf. Chap. I of my *Sour Grapes,* Cambridge: Cambridge University Press, 1983, for more details.
6. This is contested by David Pears, *Motivated Irrationality,* Oxford: Oxford University Press, 1984, Chap. 5. He argues that motivated, irrational belief formation always takes the form of a failure to correct an irrational belief, not the positive form of directly producing it; hence there is never any superfluous irrationality. I disagree, but the point is not essential to my argument, since there are other ways in which a belief might be caused by something other than the available evidence. A person might be hypnotized into forming a belief for which he also has good evidence without having formed the belief prior to the hypnosis, since we do not usually put together the pieces of information in our mind unless there is a need to do so.
7. Richard Nisbett and Lee Ross, *Human Inference: Strategies and Shortcomings of Social Judgment,* Englewood Cliffs, N.J.: Prentice Hall, 1980, pp. 267–268.
8. See my *Sour Grapes,* Chap. I.3.
9. Davidson, op. cit. (note 2), Chap. 2.
10. R. Ashby, *Introduction to Cybernetics,* London: Chapman and Hall, 1971, p. 130.
11. E. Ullmann-Margalit and S. Morgenbesser, "Picking and Choosing," *Social Research* 44 (1977): 757–785.
12. See Gerard Debreu, *Theory of Value,* New York: Wiley, 1959, and many later expositions.

13. John Harsanyi, *Rational Behavior and Bargaining Equilibrium in Games and Social Situations,* Cambridge: Cambridge University Press, 1977, p. 114.
14. This follows if we make the assumption of consistent preferences (K. Suzumura, *Rational Choice, Collective Decisions and Social Welfare,* Cambridge: Cambridge University Press, 1984), a somewhat weaker requirement than transitivity.
15. A. Sen and B. Williams, Introduction to *Utilitarianism and Beyond,* ed. A. Sen and B. Williams, Cambridge: Cambridge University Press, 1982, p. 17.
16. The alternatives are *x*: give to one child, *y:* give to the other child, and *z:* give to neither. It may happen that *because* I neither weakly prefer *x* to *y* nor *y* to *x,* I strongly prefer *z* to both, perhaps because it would create family trouble if I selected one child without being able to justify my choice in terms of welfare. Yet in the absence of *x* (or *y*), I would strongly prefer *y* (or *x*) to *z*.
17. I do not, of course, deny that actions may be valued for themselves. The assumption is made only for the sake of simplifying the discussion.
18. My *Explaining Technical Change,* Cambridge: Cambridge University Press, 1983, Appendix I, offers a further discussion.
19. For the proof that in decision-making under uncertainty one can rationally only take account of the best and the worst consequences of each action, see K. Arrow and L. Hurwitz, "An Optimality Criterion for Decision-Making under Uncertainty," in: *Uncertainty and Expectation in Economics,* ed. C.F. Carter and J.L. Ford, Clifton, N.Y.: Kelley, 1972. The proof turns upon the idea that rational choice should remain invariant under an arbitrary reclassification of "states of nature."
20. Recent work has raised a third possibility: even if there is only one equilibrium point in the game, there may be several strategy sets that are "rationalizable" [B.D. Bernheim, "Rational Strategic Behavior," *Econometrica* 52 (1984): 1007–1028; D.G. Pearce, "Rationizable Strategic Behavior and the Problem of Perfection," *Econometrica* 52 (1984): 1029–1050].
21. See my *Explaining Technical Change* (note 18), pp. 109ff., drawing on P. Dasgupta and J. Stiglitz, "Uncertainty, Industrial Structure and Speed of R&D," *Bell Journal of Economics* 11 (1980): 1.28.
22. Since the point is crucial, let me clarify it by means of an analogous example. In the theory of induced factor-bias in technical change, the argument was put forward that firms optimize with respect to an "innovation possibility frontier" [C. Kennedy, "Induced Bias in Innovation and the Theory of Distribution," *Economic Journal* 74 (1964): 541–547]. Although one may agree, at least for the sake of argument, that an omniscient observer would know which innovations are possible at a given time, it is impossible to see how this would help explaining the behavior of the firms, since there is no way in which they could acquire the same knowledge. Rational-choice explanations turn upon what the agents *believe* to be the best action, not on an objective conception of the best. Any theory that neglects this constraint lacks microfoundations [W. Nordhaus, "Some Sceptical Thoughts on the Theory of Induced Innovations," *Quarterly Journal of Economics* 87 (1973): 208–219].
23. Leif Johansen, *Lectures on Macroeconomic Planning,* Amsterdam: North-Holland, 1977, p. 144. Ultimately the argument derives from Herbert Simon. For a strikingly provocative discussion, see also S. Winter, "Economic 'Natural Selection' and the Theory of the Firm," *Yale Economic Essays* 4 (1964): 225–272.
24. Davidson, op. cit. (note 2), Chap. 11.