29 The Rationality Principle (1967)

In this paper I wish to consider the problem of *explanation in the social sciences*, and briefly to compare and contrast it with the analogous problem in the natural sciences [discussed in selection 12 above]. My thesis is that social explanations are very similar to certain physical explanations, but that the problem of explanation in the social sciences does give rise to problems that are not encountered in the natural sciences.

Let me begin by distinguishing between two kinds of problems of explanation or prediction.

(1) The first kind is the problem of explaining or predicting one or a smallish number of *singular events*. An example from the natural sciences would be, 'When will the next lunar eclipse (or, say, the next two or three lunar eclipses) occur?' An example from the social sciences would be, 'When will there be the next rise in the rate of unemployment in the Midlands, or in Western Ontario?'

(2) The second kind of problem is the problem of explaining, or predicting, a certain *kind or type* of event. An example from the natural sciences would be, 'Why do lunar eclipses occur again and again, and only when there is a full moon?' An example from the social sciences would be, 'Why is there a seasonal increase and decrease of unemployment in the building industry?'

The difference between these two kinds of problem is that the first can be solved *without constructing a model*, while the second is most easily solved *with the help of constructing a model*.

Now it seems to me that in the theoretical social sciences it is hardly ever possible to answer questions of the first kind. The theoretical social sciences operate almost always by the method of constructing *typical* situations or conditions – that is, by the

method of constructing models. (This is connected with the fact that in the social sciences, there is, in Hayek's terminology, less 'explanation in detail' and more 'explanation in principle' than in the physical sciences.)

It is important to realize the close similarity of explanations in the social sciences with explanations of the second kind in the natural sciences. Suppose, in the natural sciences, we wish to explain the repeated occurrence of lunar eclipses. In this case we may construct an actual mechanical model, or refer to a perspective drawing. For our limited purpose, the model may be very rough indeed. It may consist of a fixed lamp: the sun; a little wooden earth rotating in a circle round the sun, and a little moon rotating in a circle round the earth. One thing would be essential however: the planes of the two movements must be so inclined towards each other that we obtain lunar eclipses sometimes, but not always, when the moon is full.

A critical discussion of our rough model must give rise, however, to a new problem, 'How are earth and moon propelled in the real world?'; and with this we may come to Newton's laws of motion. There is no need, however, to introduce initial conditions explicitly into our solution: as far as problems of the second kind are concerned (the explanation of *types* of events) initial conditions may be completely replaced by the construction of the model, which one might say, incorporates *typical* initial conditions. But if we wish to make the model move, or work, or, as we may say, if we wish to '*animate*' the model; that is, if we wish to represent the way in which the various *elements* of the model act upon each other, then we do need *universal laws* (in this case, the consequences of approximating Newton's laws of motion).

So much for the natural sciences. As for the social sciences, I have elsewhere [in the previous selection] proposed that we can construct our models by means of *situational analysis*, which provides us with models (rough and ready models to be sure) of typical social situations. And my thesis is that only in this way can we explain and understand what happens in society: social events.

Now if situational analysis presents us with a model, the question arises: what corresponds here to Newton's universal laws of motion which, as we have said, 'animate' the model of the solar

system? Or in other words, how is the model of a social situation 'animated'?

The usual mistake made here is to assume that in the case of human society, the 'animation' of a social model has to be provided by the human *anima* or *psyche*, and that here, therefore, we have to replace Newton's laws of motion either by laws of human psychology in general, or perhaps by the laws of individual psychology pertaining to the individual characters who are involved as actors in our situation.

But this is a mistake, for more reasons than one. First of all, in our situational analysis itself we *replace* concrete psychological experiences (or desires, hopes, tendencies) by abstract and typical situational elements, such as 'aims' and 'knowledge'. Secondly, it is the central point of situational analysis that we need, in order to 'animate' it, no more than the assumption that the various persons or agents involved act *adequately, or appropriately*; that is to say, in accordance with the situation. Here we must remember, of course, that the situation, as I use the term, already contains all the relevant aims and all the available relevant knowledge, especially that of possible means for realizing these aims.

Thus there is only one animating law involved – the principle of acting appropriately to the situation; clearly an *almost empty* principle. It is known in the literature under the name *'rationality principle'*, a name which has led to countless misunderstandings.

If you look upon the rationality principle from the point of view which I have here adopted, then you will find that it has little or nothing to do with the empirical or psychological assertion that man always, or in the main, or in most cases, acts rationally. Rather, it turns out to be an aspect of, or a consequence of, the methodological postulate that we should pack or cram our whole theoretical effort, our whole explanatory theory, into an analysis of the *situation*: into the *model*.

If we adopt this methodological postulate, then, as a consequence, the animating law will become a kind of zero principle. For the principle may be stated in this way: having constructed our model, our situation, we assume no more than that the actors act within the terms of the model, or that they 'work out' what was *implicit* in the situation. This, incidentally, is what the term 'situational logic' alludes to.

The adoption of the rationality principle can therefore be regarded as a byproduct of a methodological postulate. It does not play the role of an empirical explanatory theory, of a testable hypothesis. For in this field, the empirical explanatory theories or hypotheses are our various models, our various situational analyses. It is these which may be empirically more or less adequate; which may be discussed and criticized, and whose adequacy may sometimes even be tested. And it is our analysis of a concrete empirical situation which may fail some empirical test, thereby enabling us to learn from our mistakes.

Tests of a model, it has to be admitted, are not easily obtainable and usually not very clearcut. But this difficulty arises even in the physical sciences. It is connected, of course, with the fact that models are always and necessarily rough; that they are always and necessarily schematic oversimplifications. Their roughness entails a comparatively low degree of testability; for it is difficult to decide what is a discrepancy due to the necessary roughness, and what is a discrepancy indicative of a failure, a refutation of the model. Nevertheless, we can sometimes decide by tests which one of two (or more) competing models is the best. And in the social sciences, tests of a situational analysis can sometimes be provided by historical research.

But if the rationality principle does not play the role of an empirical or psychological proposition, and more especially, if it is not treated as subject on its own to any kind of tests: if tests, when available, are used to test a particular model, a particular situational analysis, of which the rationality principle forms a part; then even if a test decides that a certain model is less adequate than another one, since both operate with the rationality principle, we have no occasion to test this principle.

This remark explains, I think, why the rationality principle has been frequently declared to be *a priori* valid. And indeed, if it is not empirically refutable what else could it be but *a priori* valid?

The point is of considerable interest. Those who say that the rationality principle is *a priori* mean, of course, that it is *a priori* valid, or *a priori* true. But it seems to me quite clear that they must be mistaken. For the rationality principle seems to me clearly false – even in its weakest zero formulation which may be put like this:

'Agents always act in a manner appropriate to the situation in which they find themselves.'

I think one can see very easily that this is not so. One has only to observe a flustered driver, desperately trying to park his car when there is no parking space to be found, in order to see that we do not always act in accordance with the rationality principle. Moreover, there are obviously vast personal differences, not only in knowledge and skill – these are part of the situation – but in assessing or understanding a situation; and this means that some people will act appropriately and others not.

But a principle that is not universally true is false. Thus the rationality principle is false. I think there is no way out of this. Consequently I must deny that it is *a priori* valid.

Now if the rationality principle is false, then an explanation which consists of the conjunction of this principle and a model must also be false, even if the particular model in question is true.

But can the model be true? Can any model be true? I do not think so. Any model, whether in physics or in the social sciences, must be an oversimplification. It must omit much, and it must overemphasize much.

My views on the rationality principle have been closely questioned. I have been asked whether there is not some confusion in what I say about the status of the 'principle of acting adequately to the situation' (that is, of my own version of the 'rationality principle'); I was told, quite rightly, that I should make up my mind whether I want it to be a methodological principle, or an empirical conjecture. In the first case it would be clear that, and why, it could not be empirically tested; also why it could not be empirically false (but only part of a successful or unsuccessful methodology). In the second case, it would become part of the various social theories – the animating part of every social model. But then it would have to be part of some empirical theory, and would have to be tested along with the rest of that theory, and rejected if found wanting.

This second case is precisely the one that corresponds to my own view of the status of the rationality principle: I regard the principle of adequacy of action (that is, the rationality principle) as an integral part of every, or nearly every, testable social theory.

Now if a theory is tested, and found faulty, then we have always to decide which of its various constituent parts we shall make accountable for its failure. My thesis is that it is sound methodological policy to decide not to make the rationality principle accountable but the rest of the theory; that is, the model.

In this way it may appear that in our search for better theories we treat the rationality principle as if it were a logical or a metaphysical principle exempt from refutation: as unfalsifiable, or as *a priori* valid. But this appearance is misleading. There are, as I have indicated, good reasons to believe that the rationality principle, even in my minimum formulation, is actually false, though a good approximation to the truth. Thus it cannot be said that I treat it as *a priori* valid.

I hold, however, that it is good policy, a good methodological device, to refrain from blaming the rationality principle for the breakdown of our theory: we learn more if we blame our situational model.

The main argument in favour of this policy is that our model is far more interesting and informative, and far better testable, than the principle of the adequacy of our actions. We do not learn much in learning that this is not strictly true: we know this already. Moreover, in spite of being false, it is as a rule sufficiently near to the truth: if we can refute our theory empirically, then its breakdown will, as a rule, be pretty drastic, and though the falsity of the rationality principle may be a contributing factor, the main responsibility will normally attach to the model. Another point is this: the attempt to replace the rationality principle by another one seems to lead to complete arbitrariness in our model-building. And we must not forget that we can test a theory only as a whole, and that the test consists in finding the better of two competing theories which may have much in common; and most of them have the rationality principle in common.

But did not Churchill say, in *The World Crisis*, that wars are not won but only lost – that, in effect, they are competitions in incompetence? And does not this remark provide us with a kind of model for typical social and historical situations; *a kind of model which emphatically is not animated by the rationality principle of the adequacy of our actions, but by a principle of inadequacy*?

The answer is that Churchill's dictum means that most war

leaders are inadequate to their task, that they do not see the situation as it is, rather than that their actions cannot be understood (in good approximation at least) as adequate for the situation *as they see it*.

In order to understand their (inadequate) actions, we have therefore to reconstruct a wider view of the situation than their own. This must be done in such a way that we can see how and why the situation as they saw it (with their limited experience, their limited or overblown aims, their limited or overexcited imagination) led them to act as they did; that is to say, adequately for their inadequate view of the situational structure. Churchill himself uses this method of interpretation with great success, for example in his careful analysis of the failure of the Auchinleck/Ritchie team (in volume IV of *The Second World War*).

It is interesting to see that we employ the rationality principle to the limit of what is possible whenever we try to understand an action, even the action of a madman. We try to explain a madman's actions, as far as possible, by his aims (which may be monomaniac) and by the 'information' on which he acts, that is to say, by his convictions (which may be obsessions, that is, false theories so tenaciously held that they become practically incorrigible). In so explaining the actions of a madman we explain them in terms of our wider knowledge of a problem situation which comprises his own, narrower, view of his problem situation; and understanding his actions means seeing their adequacy according to his view – his madly mistaken view – of the problem situation.

We may in this way even try to explain how he arrived at his madly mistaken view: how certain experiences shattered his originally sane view of the world and led him to adopt another – the most rational view he could develop in accordance with the information at his disposal, so far as he found it credible; and how he had to make this new view *incorrigible*, precisely because it would break down at once under the pressure of refuting instances which would leave him (so far as he could see) stranded without any interpretation of his world: a situation to be avoided at all costs, from a rational point of view, since it would make all rational action impossible.

Freud has often been described as the discoverer of human irrationality; but this is a misinterpretation, and a very superficial

one to boot. Freud's theory of the typical origin of a neurosis falls entirely into our scheme: of explanations with the help of a situational model *plus* the rationality principle. For he explains a neurosis as an attitude adopted (in early childhood) because it was the best available way out of a situation which the agent (the child, the patient) was unable to understand and cope with. Thus the adoption of a neurosis becomes a rational act – as rational, say, as the act of a man who, jumping back when confronted by the danger of being run over by a car, collides with a bicyclist. It is rational in the sense that the agent chose what appeared to him the immediately or obviously preferable or perhaps just the lesser evil – the less intolerable of two possibilities.

I shall say no more here about Freud's method of therapy than that it is even more rationalistic than his method of diagnosis or explanation; for it is based on the assumption that once a man fully understands what befell him as a child, his neurosis will pass away.

But if we thus explain everything in terms of the rationality principle, does it not become tautological? By no means; for a tautology is obviously true, whilst we make use of the rationality principle merely as a good approximation to the truth, recognizing that it is not true, but false.

But if this is so, what becomes of the distinction between rationality and irrationality? Between mental health and mental disease?

This is an important question. The main distinction, I suggest, is that a healthy person's beliefs are not incorrigible: a healthy person shows a certain readiness to correct his beliefs. He may do so only reluctantly, yet he is nevertheless ready to correct his views under the pressure of events, of the opinions held by others, and of critical arguments.

If this is so then we can say that the mentality of the man with definitely fixed views, the 'committed' man, is akin to that of the madman. It may be that all his fixed opinions are 'adequate' in the sense that they happen to coincide with the best opinion available at the time. But in so far as he is committed, he is not rational: he will resist any change, any correction; and since he cannot be in possession of the full truth (nobody is) he will resist rational

correction of even wildly mistaken beliefs. And he will resist, even if their correction is widely accepted during his lifetime.

Thus when those who praise commitment and irrational faith describe themselves as irrationalists (or post-rationalists) I agree with them. *They are irrationalists*, even if they are capable of reasoning. For they take pride in rendering themselves incapable of breaking out of their shell; they make themselves prisoners of their manias. They make themselves spiritually unfree, by an action whose adoption we may explain (following the psychiatrists) as one that is rationally understandable; understandable, for example, as an action they commit owing to fear – fear of being compelled, by criticism, to surrender a view which they dare not give up since they make it (or believe they must make it) the basis of their whole life. (Commitment – even 'free commitment' – and fanaticism, which, we know, can border on madness, are thus related in the most dangerous manner.)

To sum up: we should distinguish between rationality as a personal attitude (which, in principle, all sane men are capable of sharing) and the rationality principle.

Rationality as a personal attitude is the attitude of readiness to correct one's beliefs. In its intellectually most highly developed form it is the readiness to discuss one's beliefs critically, and to correct them in the light of critical discussions with other people.

The 'rationality principle' on the other hand has nothing to do with the assumption that men are rational in this sense – that they always adopt a rational attitude. It is, rather, a minimum principle (since it assumes no more than the adequacy of our actions to our problem situations as we see them) which animates all, or almost all, our explanatory situational models, and, although we know it not to be true, we have some reason to regard as a good approximation. Its adoption reduces considerably the arbitrariness of our models; an arbitrariness which becomes capricious indeed if we try to proceed without this principle.