Rules, Reasons, and Norms

Selected Essays

Philip Pettit

CLARENDON PRESS · OXFORD

REFERENCES

Pettit, P. (1993). *The Common Mind: An Essay on Psychology, Society and Politics.* New York: Oxford University Press.

—— (2000). 'Rational Choice, Functional Selection and Empty Black Boxes', *Journal of Economic Methodology*, 7: 33–57.

—— (2001). *A Theory of Freedom: From the Psychology to the Politics of Agency.* Cambridge: Polity.

—— (2002). 'Is Criminal Justice Politically Feasible', *Buffalo Criminal Law Review.*

1

Three Aspects of Rational Explanation

Rational explanation, as I understand it here, is the sort of explanation we practise when we try to make intentional sense of a person's attitudes and actions. We may postulate various obstacles to rationality in the course of offering such explanations, but the point of the exercise is generally to present the individual as a more or less rational subject: as a subject who, within the constraints of the obstacles postulated—and they can be quite severe—displays a rational pattern of attitude-formation and decision-making.

In this paper I want to draw attention to three distinct, and progressively more specific, aspects of such rational explanation, as we practise it in everyday life. I do so, because I believe that they are not always prised apart sufficiently. The first aspect of rational explanation is that it is a programming variety of explanation, in a phrase that Frank Jackson and I introduced some years ago (Jackson and Pettit 1988). The second is, in another neologism (Pettit 1986), that it is a normalizing kind of explanation. And the third is that it is a variety of interpretation: if you like, it is a hermeneutic form of explanation.

1. RATIONAL EXPLANATION AS PROGRAMMING EXPLANATION

Rational explanation of action involves the attempt to explain an agent's speech or behaviour by reference to distinctive psychological states: roughly, by reference to states that reflect the information to which the agent gives countenance and the inclination that moves him or her; by reference, as the stock phrase has it, to beliefs and desires. The first thing to be said in characterization of such explanation is that it invokes higher-level causal factors, not factors that operate at the most basic level there is. A similar point holds for the rational explanation of attitudes, of course—the explanation of why an agent comes to think or feel something, for example—but we can take the explanation of action as our main point of reference.

A level is characterized by the causally relevant properties that figure there: the physical level by physical properties, the biological by biological, and so on. Such levels will be distinguished from one another, roughly, by the fact that the properties at any level join forces with one another in a way in which they do not join forces with properties from other levels. And such levels will be designated as higher and lower, depending on which are thought to be causally more basic.

Consider these two properties of a rod: first, its malleability, second, its particular molecular structure. Those properties may each be relevant, by whatever test of relevance, to the rod's bending under a certain pressure. But they are not relevant in the sense of each playing a part in the same process; neither appears earlier than the other in the same process and neither combines with the other, in the way in which the presence of the oxygen combines with the striking of the match to produce combustion. They do not get together in those ways—they do not join forces, as we might put it—in the production of the event to which, nonetheless, they are each causally relevant. They are causally relevant to the event at different levels.

If the malleability of a rod and its molecular structure are properties of different levels, which level is lower, which higher? The judgement will be driven by our assumptions as to whether the causal relevance of the molecular structure mediates the causal relevance of the malleability, or the other way around. Is the malleability relevant to the rod bending in virtue of the fact that it is the molecular structure of the rod that accounts, in context, for the bending? Or the other way around? Clearly, by this test, the molecular structure is causally the more basic level. To be malleable is to have such a molecular structure as will allow bending under such and such a pressure; if the malleability is causally relevant to the bending, its relevance is mediated by the relevance of the structure.

I claimed that the first thing that characterizes rational explanation is that the psychological factors it invokes as causally relevant are higher level. The factors involved are intentional properties, properties of belief and desire, and they represent a different level from that represented, for example, by the properties identified in neurophysiology; they do not join forces with such properties in producing behaviour and yet both sorts of properties are causally relevant, so we judge, to behaviour. As between the psychological and the neurophysiological families of behaviour-relevant properties, which represents the more basic level? If we are to avoid positing special Cartesian forces, then we must say that the neurophysiological level is the more basic. If mother nature has designed us to be such that our psychological states are causally relevant to our doing this or that, if it has designed us to be psychologically organized systems, then it has done so through ensuring that the neurophysiological connections to behaviour sustain the psychological connections: it has done so through designing our neurophysiology to sustain the causal relevance of psychological states in something like the way that the molecular structure of the rod sustains the causal relevance of its malleability.

Let us agree that the psychological properties introduced in rational explanation are higher-level, in particular that they are of a higher level than neurophysiological properties. But how can properties at different levels both be causally relevant to one and the same thing? How can they collaborate causally, as it were, given that they do not join forces: given that they do not collaborate in the familiar diachronic or synchronic fashion? Reflection on this problem leads us to see that rational explanation is a form of programming explanation (see Jackson and Pettit 1988, 1990*a*, *b*; Pettit 1996).

The programming model focuses on the way causal and explanatory relevance, however paraphrased, may be reproduced across different levels. It applies to the intentional and neurophysiological levels but it also applies in many other cases. It helps us to make sense, not just of how beliefs and desires can be causally relevant to something that is produced by neurophysiological antecedents, but also of how malleability can be causally relevant to the bending that is produced, under appropriate pressure, by the molecular structure of the rod; and so on in other cases.

Suppose that there is no doubt about the causal relevance of properties at a given level *L* to the occurrence of an event *E*, of a given type. Suppose that we are interested in how a property, *P*, at a higher level can be simultaneously relevant to *E*. According to the programme model, *P* will be causally relevant to *E* just in case three conditions are fulfilled.

1. The instantiation of *P* non-causally involves the instantiation of certain properties—perhaps these, perhaps those—at the lower level *L*: typically, the instantiation of the *L*-properties will 'realize' *P*, as it is said, given the context.
2. *L*-properties of the sort associated with instantiations of *P*, or at least most of them, are such as generally to be causally relevant—in the circumstances—to the occurrence of an *E*-type event.
3. The *L*-properties associated with the actual instantiation of *P* are causally relevant to the occurrence of *E*.

These conditions are readily illustrated. Intuitively, the malleability of this rod is causally relevant to its bending, and relevant simultaneously with the exact molecular structure. How so? Because the programme model applies. The instantiation of the malleability involves the instantiation of certain molecular-structural properties; the sorts of properties associated with instantiations of malleability are such as generally to be causally relevant to the sort of bending effect in question; and the molecular-structural properties associated with the actual instantiation of malleability are causally relevant to the bending.

A computer program ensures that things are organized in the machine language of the computer—may be in this fashion, may be in that—so that certain results reliably follow on certain inputs. In cases where the programme model applies, even in a simple case like that of the malleable rod, the higher-level property can be cast as programming in a parallel manner for the appearance of a certain effect. The presence of the malleability ensures, non-causally, that things are organized at the molecular level—the level corresponding to the machine language—so that the rod will bend under suitable pressure. Where the molecular structure is described as producing the bending, the malleability can be thought of as programming for the effect produced.

Other examples of the programme model become salient as we recognize suitably corresponding relations across levels in different cases. In every case the relation must be such that the instantiation of the higher-level property ensures or at least probabilifies—in a non-causal way—that there are causally relevant properties present at the lower level. But there may be quite different reasons applicable in the different cases as to why that relation obtains; each case will require its own annotation. The squareness of the peg probabilifies the sort of molecular contact that blocks the peg going through the round hole; the (boiling) temperature of the water in the closed flask probabilifies the presence of a molecule of the right position and momentum to break a molecular bond in the surface and crack the flask; the rise in unemployment probabilifies a shift in motives and opportunities that is likely to increase aggregate crime; and so on across a great variety of possible cases. The probabilification holds for different reasons in the different cases. But the fact that it obtains shows how the programme model may apply in any of the examples, making sense of how the higher-level property can be causally relevant to something that is also traceable to the lower-level properties.

As the programme model applies to these sorts of cases, so it applies too to the way in which intentional and neurophysiological properties produce behaviour. How is a particular psychological set causally relevant to an agent's doing something? In particular, how is it relevant, given that the action is produced without remainder—without leaving anything to be explained—by a certain complex of neurophysiological states? The programme model suggests that the psychological set will be causally relevant so far as its realization in an agent of that kind makes it more probable than it would otherwise have been—it may make it more or less certain—that there will be a neurophysiological configuration of properties present—may be this, may be that—that is sufficient to produce the required behaviour. The psychological set may not produce the behaviour in the same way in which the neurophysiological complex does. But it is nonetheless causally relevant to the appearance of that behaviour. It programmes for the behaviour to the extent that its realization means, more or less certainly, that there will be a suitable neurophysiological producer present.

I hope that these remarks will help to make vivid the idea that rational explanation is a sort of programming explanation. I have presented arguments elsewhere in defence of that idea. Here I will say only that it is not clear how a higher-level, rational explanation can introduce causally relevant properties unless the programme model applies. There are no alternatives in the literature that would make comparable sense of the way in which properties at higher and lower levels can be simultaneously relevant to a certain effect (see Pettit 1996: ch. 1). If not this, what?

It will be useful, however, to consider an objection. The malleability of a rod may be programmatically and therefore causally relevant to its bending, but we would not ordinarily say that the instantiation of that property—the rod's being or becoming malleable—was a cause of the bending. The state or event in question is not a distinct existence, in Hume's phrase, that contingently gives rise to the bending: rather it is a disposition—a property of being such as to bend under certain pressures—that the bending manifests. Does this mean, by analogy, that, while the programme model allows us to say that beliefs and desires are programmatically and causally relevant in the production of action, the instantiations of such properties—the mental states and events in question—are not causes, in the ordinary sense, of action? If it does mean this, then that is an objection to the model. For in ordinary parlance we regularly say that someone's believing and desiring certain things was the cause of his or her acting in a certain manner.

Happily, the programme model does not have the unwanted implication. A factor that is programmatically relevant to the production of an effect may or may not be a cause of that effect, in our ordinary way of speaking. We may not speak of the malleability of a rod as a cause of its bending. But we do speak of other programming factors as potential causes: we say that the rise in the water's temperature caused the flask to break, for example, and that the increase in unemployment caused the increase in crime. And so, for all that the programme model forces on us, we may say that, when an agent instantiates beliefs and desires that are programmatically relevant to an action, then it is the instantiation of those states—it is the agent's believing and desiring the things in question—that is the cause of the action.

There is a question, of course, as to what is required of a programming property if its instantiation is to deserve to be called a cause of that for which it programmes. Frank Jackson (1998) suggests, for example, that a programming factor can count as a cause in this sense if the programming property is disjunctive—but not wildly disjunctive—and if the lower-level realizers of the property correspond to different disjuncts. He argues that such a disjunctive property may be causally powerful, not just causally programmatic—causally productive, not just causally relevant—and he thinks that belief and desire properties conform to that pattern (Jackson 1995). Beyond noticing the availability of that sort of answer, however, we cannot pursue the question raised any further.

2. RATIONAL EXPLANATION AS NORMALIZING EXPLANATION

Whenever the programme model applies, whenever there are higher-level properties that exercise causal relevance, we will find lawlike regularities in place. I have in mind regularities like that which binds the malleability of the rod to its bending under such and such pressure, or the squareness of the peg to its being blocked from going through a suitably corresponding round hole. Programme explanation will amount to what I have described as normalizing explanation just in case the relevant regularities, or at least some of the relevant regularities, have the status of norms. Otherwise it will be a sort of regularizing explanation (Pettit 1986).

All of the non-intentional examples of programme explanation that were given in the last section involve non-normative regularities and so the explanation in question is of the regularizing kind. Consider the regularities linking malleability and bending, the squareness of the peg and the blocking, the (boiling) temperature of the water and the cracking, the rise in unemployment and the increase in crime. None of these regularities represents a norm for the behaviour of a system, in any plausible sense of 'norm'.

Things are different, however, in other cases. Suppose that we have designed a computer to add any numbers presented to it and to display the sum: we have designed it to function as an adding device. If we have designed the computer properly, then, whenever a set of numbers is registered, the computer will respond by giving us their sum. The presentation of the numbers will be causally relevant to that response, even though the response is produced at a lower level by the machine features of the computer. The presentation of the numbers will program for the appropriate response, ensuring the presence of a machine profile that produces it. The programme model will apply.

This case resembles the other instances of the programme model fairly closely, with one difference. This is that the sort of regularity involved in any of the adding machine's responses will have the status of a norm. Given that we know or assume that this is meant to be an adding device, we can deduce that, if it is given the numbers seven and four as input, then it will display eleven as output. It is a hypothetical imperative for any system that if it is to count as an adder then, for input seven and four, it should produce output eleven. Thus, assuming that the system is an adder, we can say that it is a norm for the system that, for that input, it should produce that output.

There is no mystery in how a regularity, in particular a programmed regularity, can have the status of a norm. As we have imagined this happening with an artificially designed system, so we can envisage it coming about with any system that is the product of design or selection. A regularity will count as a norm for a system just in case the satisfaction of that regularity is required for the system to succeed in the role for which it has been designed or selected.

An example from the realm of natural selection will help to make the point. We assume that the temperature-control system in the human body has been selected—or the associated genetic profile has been selected—for the effect it has in maintaining a certain temperature within the body. That being so, we must see the regularity whereby it produces perspiration in a sauna-like atmosphere as a norm for the system and, more generally, the organism. The regularity is not just something that happens to obtain. It is something that more or less has to obtain if the system is to be successful in the role for which it has been shaped.

That a programmed regularity is a norm is not of ontological significance. It means that the system in question is the product of design or selection, it is true, but it does not entail any further difference between that system and other less normatively directed organisms. Normatively organized systems, in the sense introduced here, are as much a part of the natural world, and are just as subject to the regime of natural laws as any rock or cloud or mountain.

But, if the normative status of programmed regularities is not of ontological significance, it may be very important from a heuristic point of view. The reason should be clear. We can have evidence that a system is designed or selected to fit a certain sort of role, and we may be able to work out the regularities that should be normative for such a system, independently of identifying empirically the regularities that it actually satisfies in its behaviour. Knowledge of the designer responsible, or of the designer's purposes, or just a little experience of the system itself, may convince us that this device is meant to add. And that being so, we are in a position to predict a whole range of responses, at least when the system does not go on the blink. We can occupy a vantage point on the performance of the system that is going to be difficult to attain with any agent that is not normatively directed in this way.

The normative status of certain programmed regularities may be not just heuristically significant—not just significant in the generation of knowledge—but also significant from an explanatory point of view. To get an explanation of the kind that is relevant here is always, I take it, to get information on the causal history of the event or condition explained (see Lewis 1986: essay 22; Pettit 1996: ch. 5). To know that a certain antecedent not only programmes for a result, but programmes for it normatively, is to acquire a distinctive sort of information on the genesis of the event. It is to learn that the programming factor gave rise in context to the result, as in any other case. But it is also to learn that, given the role for which the system is designed or selected—given, for example, that the system is an adding device—it was inevitable that that antecedent factor should give rise to that result; it could have failed to do so only through malfunction.

The normalizing explanation not only tells us what any programme explanation tells us, in other words; it also directs us to a certain sort of modal or counterfactual information about the genesis of the matter explained. It lets us see that, in any possible world where the system is to satisfy its role—subject perhaps to certain constraints—things will have to be such that, absent malfunction, the antecedent state gives rise to the result in question. Not only are things organized in this world so that the realization of the programming state more or less ensures that there will be a lower-level state available to produce the result. Things have to be organized in that way in any world where the system satisfies the role for which it is designed or selected.

So much on normalizing explanation in general. What I now suggest is that rational explanation is not just a form of programming explanation, it is also a form of normalizing explanation. In dealing with one another, we put in place an assumption that, absent malfunction and other ills, we are creatures who satisfy the role of rational agents: we are more or less rational in our responses to evidence and more or less rational in moving from what we believe and value to what we do (see Cherniak 1986). The regularities that govern our adjustments in these respects are norms of rationality: they are regularities that any rational creature will have to respect, as the principles of addition are regularities that any adding machine will have to honour. We may believe that we satisfy the role of rational creatures as a result of natural selection, or cultural influence, or divine design, or a mix of these influences. The grounding does not matter. The important thing is that we expect one another—and, if we are to relate as human beings, we probably must expect one another—to conform to that role and to the associated regularities.

The expectation of rationality—strictly, rationality-absent-malfunction-or-disturbance-or . . .—enables us to generate predictions of another agent's behaviour that would otherwise be difficult to generate. This is the heuristic aspect of our seeing intentional regularities as norms. Furthermore, the expectation means that we each find a special explanatory significance, a significance lacking in regularizing explanation, in the fact of being able to trace another's response to an intentional, programming antecedent; we see the response as one that is required in any rational agent who displays the antecedent state. This is the explanatory significance of our representing the regularities as norms.

Just as I have not formally argued, in this paper, that rational explanation is a form of programming explanation, so also I will not repeat here my arguments (Pettit 1996) for holding that it is a form of normalizing explanation. Suffice it to mention that the picture of rational explanation as normalizing fits with a variety of views current in philosophy; it is not based in any particularly sectarian commitment. A range of views emphasize the extent to which rational explanation is directed and driven by the attempt to represent the behaviour or attitudes explained, given background and context, as in some way normatively appropriate responses. Any such view would give us reason for being hospitable to the thought that, in rational explanation, we not only trace an agent's responses to certain, programming antecedents; we often trace it to antecedents whose realization means that the responses were required or expected of the agent.

3. RATIONAL EXPLANATION AS INTERPRETATIVE EXPLANATION

It is natural to describe a form of explanation as interpretative when it reveals that the subject of explanation saw things in a certain way, thought of them in a certain way, and acted on the basis of such an assigned meaning: acted on the basis of such an interpretation of the situation. This characterization is rough and intuitive, of course, but it is clear that not every explanation, not even every explanation of a programming and normalizing character, need be interpretative in this sense. The explanations that we give for the responses of the human body in a sauna or in a cold shower will not count as interpretative. And neither will the explanations that we invoke for the adding machine's responses: the adding machine does not do any interpreting—not at least in any intuitive sense—of the inputs to which it offers those responses.

But is the intentional, normalizing explanation of a human being's responses bound to count as interpretative? Again, and surprisingly, no. Consider a human being to whom we apply, successfully, the apparatus of Bayesian decision theory. We find a pattern in the person's responses that allows us to assign a probability function—this determines degrees of belief—and a utility function—this determines degrees of desire—and to see everything he or she does, and indeed every revision of probability that occurs in the person, as rational in Bayesian terms. The utility function gives a utility figure to every prospect and the probability function offers us suitably corresponding measures of probability; different versions of Bayesian theory require different measures (Eells 1982). We find that in every thing the person does, then, expected utility is maximized: the utility of the option chosen, computed as the sum of the utilities of its probabilistically weighted possible outcomes, is always greater than the utility of any alternative.

If we were able to make decision-theoretic sense of an agent in this way, then we would surely have programming and normalizing explanations of his or her responses. We would be able to subsume those responses under regularities that count as norms for a decision-theoretically rational subject. We would be able to see each of the responses as being programmed for by the state of the agent's utility and probability functions and we would be able to see the sort of programming involved as normatively required in any suitably rational agent.

But, though we would be in a position to offer a programming and normalizing explanation of the person's responses, there is still an important sense in which we might fail to provide anything worthy of being called an interpretative explanation. Consistently with displaying the patterns that invite the decision-theoretic explanations, the agent could be a creature that does not go through any conscious ratiocination. The agent might be a sort of automaton, which enjoys such a superb design that, exposed to appropriate evidence, it revises its degrees of belief in the rational way and, presented with any range of options, forms degrees of desire, and chooses according to strength of desire, in the rational way. It might never have to think about the import of the new evidence put before it, weighing the significance of that evidence against more familiar facts. And it might never have to deliberate about the options that it faces, trying to determine their relative attractions and trying to establish which is the most desirable. Its revisions of belief, and its decisions about what to do, might just happen, without anything that approximates an interpretation of its situation. They might happen without any consciousness and without any responsibility.

But suppose that a decision-theoretic subject cannot be a mere automaton: that it must work with some pattern of interpretation of its environment. Even in that case, we have to admit that the decision-theoretic explanation itself does not give us any insight into the subject's interpretation. It is entirely silent on how things are supposed to present themselves within the forum of the agent's attention and on how the agent is supposed to think and reason about them. The explanation of a change of belief or a choice of action does not suggest, in either case, that the system imagined thinks explicitly in terms of its own probabilities and utilities; it is not clear how it would even know what these are, given the detail involved (Harman 1986: ch. 9). And the explanation leaves it entirely open as to how the agent reasons otherwise (Pettit 1991).

Suppose that its degrees of belief that *p* and that *q* lead it, rationally, to form a certain degree of belief that *r*. Or suppose that those degrees of belief combine with certain degrees of desire that *s* and that *t* to lead it, rationally, to form a certain degree of desire that *u*. How is the agent supposed to think as it reasons its way, however implicitly, to the conclusion that leaves it with the appropriate degree of belief that *r* or degree of desire that *u*? We may assume that the creature holds the objects of its grounding beliefs—'*p*' and '*q*'—before its mind. But how does it register the partiality of its beliefs in these objects? We may assume, again, that it holds the objects of its grounding desires—'*s*' and '*t*' before its mind. But how does it register the fact that it desires those propositions rather than believing them? There is no suggestion in decision theory that the agent thinks in terms of what is probable and what is desirable; probability is associated with degrees of belief in non-probability contents, desirability with degrees of desire for non-desirability contents. And so it is entirely obscure how the subject is supposed to reason and think, if indeed it does reason and think.

The pattern of decision-theoretic explanation that we have been discussing is certainly a programming and normalizing form of explanation, then, but it hardly deserves the name of interpretation. The point becomes striking when we recognize that there is a more common-or-garden sort of rational explanation that is not silent in the decision-theoretic manner on the way a person thinks and reasons. Not only does it seek to subsume our responses under appropriate norms, it also points us to how things present themselves from the agent's point of view: how they are interpreted by the agent (Pettit and Smith 1990; Pettit 1996: ch. 5).

Consider a case where someone walks up to a beggar by the roadside and puts some money in his cap. The decision-theoretic mode of explanation would direct us to the agent's utilities for the different possible outcomes—probabilistically weighted—of that option and would present the option as superior in such terms to the alternatives. But it would not give us any idea as to how the agent is thinking; indeed, as have seen, it would be compatible with the complete absence of thought. The more regular sort of intentional explanation would score over the decision-theoretic story in this regard. It might say, for example, that the agent took pity on the beggar and gave him the first coin that came to hand; or that the agent was following the principle of always giving beggars a certain amount; or that the agent conceived it to be a duty to help a beggar a day and this was the lucky one; or whatever. But, in any case, it would draw attention to the sorts of things that imposed themselves, more or less consciously, on the agent's attention. It would give us a sense, as we say, of how the agent interpreted the situation.

This common-or-garden variety of rational explanation, then, is quite distinct from the austere, decision-theoretic kind. It invokes psychological states that programme and normalize the responses explained, as decision-theoretic explanation does. But it also lets us see the structure of the subject's thought, as we might put it. The human subject is not just an arena within which degrees of belief in, and desire for, certain contents rationally come and go, and rationally congeal, as occasion requires, in the formation of decisions. Ordinary people make judgements about those contents, in particular judgements about their degrees of probability and desirability. And ordinary people may make efforts to ensure that they conform to the requirements of such self-represented probabilities and desirabilities in the attitudes formed and in the actions taken; they may try to ensure that they do not commit mistakes like the gambler's fallacy, or display failures such as weakness of will. The common-or-garden variety of explanation focuses on this process of reasoning in making sense of how a person thinks and acts. The austere decision-theoretic variety ignores the process; it treats the human subject as a black box.

When rational explanation assumes an interpretative or hermeneutic form, and not just a programming and normalizing one, then it casts the person as a reasoning or ratiocinative subject, not merely as a rational system. The rational system—the ideal subject of decision theory—may realize its rationality on the basis of a purely sub-personal mode of organization and attunement; it need not have what we would describe as a mental life. The ratiocinative system—the sort of system that our species implicitly or explicitly typifies—may be a more or less rational system in this sense but it is also something else besides; it is a rational system that attains rationality, to the extent that it does, on the basis of attention to reasons and to what reasons require. Rational explanation goes interpretative when it characterizes such a life of reasons in the person whose attitudes and actions it explains.

Rational explanation, qua programming, directs us to regularities in the way in which certain higher-level factors occasion thoughts and actions. Rational explanation, qua normalizing, represents those regularities as norms for the subject in question: ideals that it has to satisfy, though perhaps only within certain constraints and up to certain limits, on pain of not counting as a rational system. And rational explanation, qua interpretative, represents those norms as ideals that the subject tries or can try to fulfil, in the manner of a reasoning agent; it offers an insight into how the subject achieves whatever rationality he or she displays.

Or, at any rate, that is what rational, interpretative explanation ordinarily does. One element that needs to be added to this elegant picture brings out a further strength in such explanation. This is that, even when people fail to live up to relevant norms, it may still be possible to provide an interpretative explanation of why they act as they do. Consider the way they reason when they fall prey to the gambler's fallacy and assume that, given a sequence of five heads, the chance of a head on the next toss of a fair coin has to be less than a half. Or think about how people reason when a certain myopia or weakness of will makes them fall short of their own standards. In such cases they do not live up to relevant norms but they can at least be represented as attempting to live up to such norms. And, that being so, we can still look sensibly for interpretative explanations; we do not have to see them as going on the blink and behaving in an interpretatively opaque way.

Why do we seek interpretative explanation in our day-to-day dealings with one another? What function does it serve that would not be served by the non-interpretative sort of explanation that is provided by decision theory?

The answer is that we need interpretative explanation in order to be able to converse with one another (Pettit 1996; Pettit and Smith 1996: postscript). If it is going to be worthwhile talking to another person, then that person must be capable, not just of being more or less rational, but of registering and generally responding to reasons: registering that this or that piece of evidence makes it probable that such and such, for example, or registering that this or that value makes it desirable to take one or another course of action. If I do not see an interlocutor as responsive to such considerations—if I see the interlocutor just as a decision-theoretic automaton, for example—then there will be no point in engaging with the person; I might as well be talking to the wall. But when I see an interlocutor as responsive to reasons—in particular, when I explain the things the interlocutor thinks and does as responses to reasons—then I make sense of the person, precisely, in the interpretative manner.

Daniel Dennett (1979) talks of the intentional stance as the perspective we adopt when we see another creature as a more or less rational system, say as a system that makes rough decision-theoretic sense. The stance that we adopt when we see another creature as a more or less reasoning system, as a system whose thoughts and deeds are the product of interpretation, may be described by analogy as the conversational stance. We resort to interpretation to the extent that we adopt that stance, pursuing or at least envisaging conversation with the subjects of our explanations. We resort to interpretation when we try to meet other minds and not just to observe them.

REFERENCES

Cherniak, Christopher (1986). *Minimal Rationality.* Cambridge, Mass: MIT Press.

Dennett, Daniel (1979). *Brainstorms: Philosophical Essays on Mind and Psychology.* Brighton: Harvester.

Eells, Ellery (1982). *Rational Decision and Causality.* Cambridge: Cambridge University Press.

Harman, Gilbert (1986). *Change in View.* Cambridge, Mass.: MIT Press.

Jackson, Frank (1995). 'Mental Properties, Essentialism and Causation', *Proceedings of the Aristotelian Society*, 95: 253–68.

—— (1998). 'Colour, Disjunctions, Programming', *Analysis*, 58: 86–8.

—— and Pettit, Philip (1988). 'Functionalism and Broad Content', *Mind*, 97: 381–400 (reprinted in Jackson, Pettit, and Smith, *Mind, Morality and Explanation: Selected Collaborations.* Oxford: Oxford University Press, forthcoming).

—— —— (1990*a*). 'Program Explanation: A General Perspective', *Analysis*, 50: 107–17 (reprinted in Jackson, Pettit, and Smith, forthcoming).

—— —— (1990*b*). 'Causation in the Philosophy of Mind', *Philosophy and Phenomenological Research*, 50: 195–214 (reprinted in Jackson, Pettit, and Smith, forthcoming).

Lewis, David (1986). *Philosophical Papers*, ii. New York: Oxford University Press.

Pettit Philip (1986). 'Broad-Minded Explanation and Psychology', in Philip Pettit and John McDowell (eds.), *Subject, Thought, and Context.* Oxford: Oxford University Press.

—— (1991). 'Decision Theory and Folk Psychology', in Michael Bacharach and Susan Hurley (eds.), *Foundations of Decision Theory: Issues and Advances.* Oxford: Blackwell (this volume, Pt. II, Ch. 2).

—— (1996). *The Common Mind: An Essay on Psychology, Society and Politics*, paperback edn. with new postscript. New York: Oxford University Press.

—— and Smith, Michael (1990). 'Backgrounding Desire', *Philosophical Review*, 99: 565–92 (reprinted in Jackson, Pettit, and Smith, forthcoming).

—— —— (1996). 'Freedom in Belief and Desire', *Journal of Philosophy*, 93: 429–49 (reprinted in Jackson, Pettit, and Smith, forthcoming).