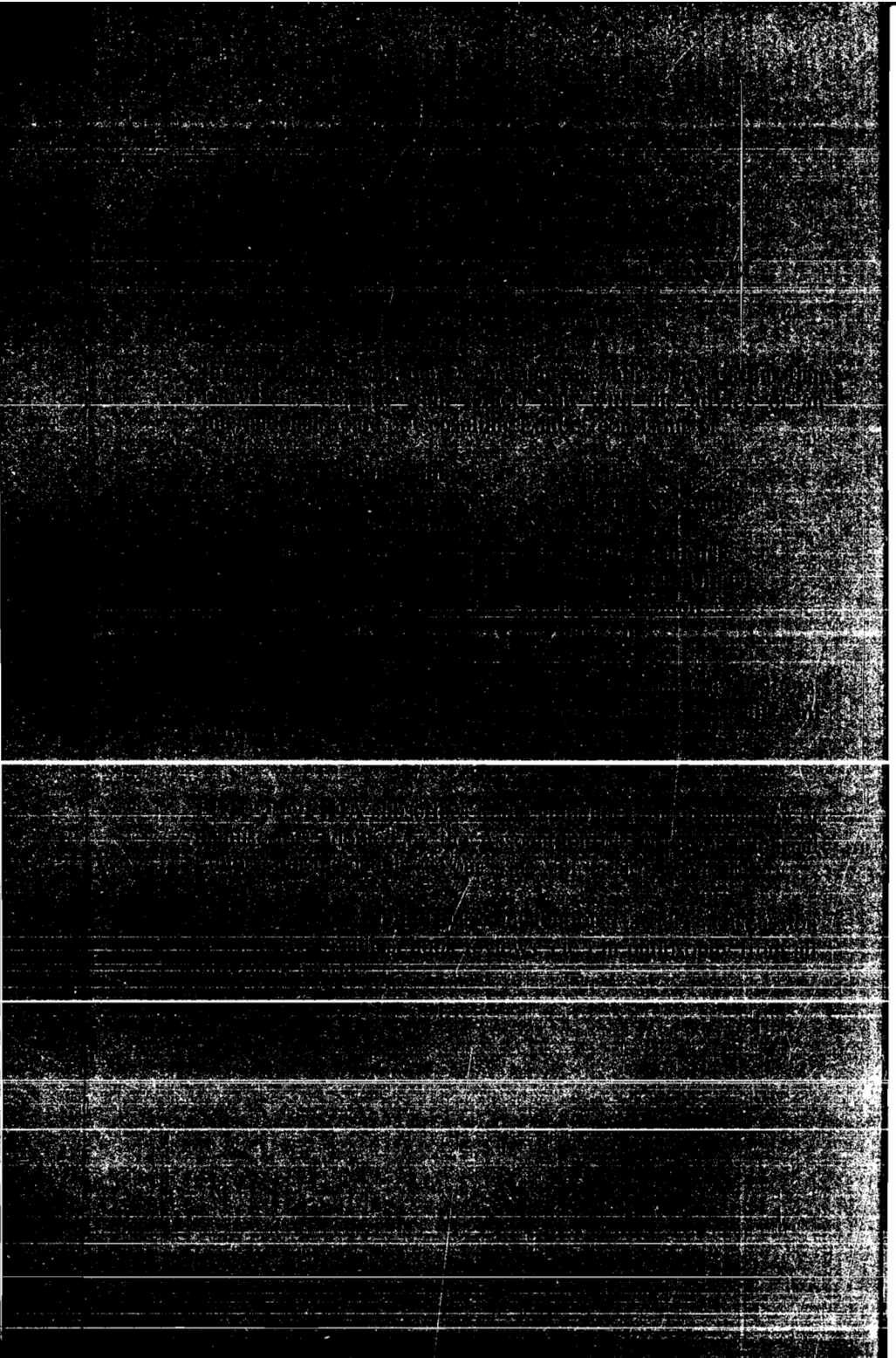

PHILOSOPHIA

PHILOSOPHICAL QUARTERLY OF ISRAEL

VOLUME 23, NOS. 1-4 JULY 1994

BAR-ILAN UNIVERSITY

Lingis, A. (1980). "On Phenomenological Explanation", *Journal of the British Society for Phenomenology*, vol. 11, no.1, January, 54-68.

Pettit, P. (1978), "Rational Man Theory", in C. Hookway and P. Pettit (eds.), *Action and Interpretation*. Cambridge University Press, pp. 45-59.

Pettit, P. (1986), "A Priori Principles and action-Explanation". *Analysis*. vol. 46, No. 1, January, 39-42.

Ricoeur, P. (1981), *Essays on Langage, Action and Interpretation*. Cambridge University Press.

TOWARDS INTERPRETATION[1]

PHILIP PETTIT

Dimitri Ginev (1994) argues that the notion of 'normalizing' explanation that I introduced in some earlier papers (Pettit 1986a, 1986b) does not capture the hermeneutic idea of intentional interpretation. I agree; it doesn't. If I suggested otherwise, as Ginev thinks, then I was wrong.

So how do normalizing explanation and intentional interpretation relate to one another? What places do they occupy, respectively, in the large network of concepts of explanation? I shall try to deal with these questions here, though in a relatively unargued mode. Answering the questions requires a big picture and a big picture needs some bold strokes (see too Pettit 1993).

My paper is in three sections. In the first section I present intentional interpretation as a kind of 'programming' explanation; in the second I cast it, more specifically, as a sort of normalizing explanation; and in the third I characterise it, more specifically still, as the sort of programming and normalizing explanation which directs us to the content of an agent's thoughts and deliberations.

1. Interpretation as programming explanation

Intentional interpretation 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; by reference, as the stock phrase has it, to beliefs and desires. The first thing to be said in characterisation of such explanation is that it invokes higher-level causal factors, not factors that operate at the most basic level there is.

What do I mean by a level? How are levels distinguished? And how do they come to be designated as higher and lower?

A level is characterised 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 go together with one another in a way in which they do not go together 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.

Suppose that two properties are both causally relevant, by whatever test of relevance, to the same result. One way of marking off levels is to say that such properties belong to the same level if and only if the causal factors they constitute have to relate as parts of the same causal whole or as stages (or parts of stages) in the same causal chain. They have to collaborate synchronically or diachronically; in a word, they have to join forces. If two properties are relevant to the same event, but do not join forces in this way, then they are of different levels. The properties of being malleable and having such and such a molecular structure are both causally relevant to a rod's bending. But they do not join forces in facilitating that result; neither is caused by the other and neither combines with the other in an act of joint production. Thus they will count as of different levels.

If the malleability of a rod and its molecular structure are properties of different levels, which level is lower, which higher? The judgment 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 which 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 characterises intentional interpretation or 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 organised 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 intentional interpretation or 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 collaborate in the familiar synchronic or diachronic fasion? Reflection on this problem leads us to see that intentional explanation is a form of program explanation (see Jackson and Pettit 1988, 1990a, 1990b and Pettit 1992, 1993).

The program model abstracts from how causal relevance is to be understood—causal relevance is taken to be a matter of intuitive judgment—and focuses on the way 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 program 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 'realise' 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 program 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 organised 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 program 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 organised 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 program model become salient as we recognise 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 which 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 program model may apply in any of the examples, making sense of how the higher-level property can be causally relevant to something which is also traceable to the lower-level properties.

As the program 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 program model suggests that the psychological set will be causally relevant so far as its realisation 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—which 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 programs for the behaviour to the extent that its realisation 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 interpretation is a sort of program explanation. I have presented

arguments elsewhere in defence of that idea. Here I will only say that it is not clear how a higher-level explanation like intentional interpretation can introduce causally relevant properties unless the program 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 1992). If not this, what?

2. Interpretation as normalizing explanation

Whenever the program model applies, whenever there are higher-level properties which 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. Program 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.

All of the non-intentional examples of program 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 (programmed) 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 program model will apply.

This case resembles the other instances of the program 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 as a norm for the system and, more generally, the organism. The regularity isn't 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 organised 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 1993, Chapter 5). To know that a certain antecedent not only programs for a result, but programs for it normatively, is to acquire a distinctive sort of information on the genesis of the event. It is to learn that the antecedent programmer gave rise in context to the result, as in any other case, but it is also to learn that consistently with the system's satisfying the role for which it is designed or selected—consistently, for example, with its being an adding device—that antecedent condition was required in that context to give rise to that result; it could have failed to do so only through malfunction.

The normalizing explanation not only tells us what any program 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 organised in this world so that the realisation 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 organised 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 intentional interpretation is not just a form of program 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.

As I have not argued that interpretation is a form of program explanation, so I will not repeat my arguments here for holding that it is a form of normalizing explanation (see Pettit 1986a, 1986b). Suffice it to mention that the picture of interpretation as normalizing explanation fits with a variety of views current in philosophy; it is not based in any particularly sectarian commitment. A range of views emphasise the extent to which intentional interpretation is directed and driven by the attempt to represent the behaviour explained, given background and context, as a more or less rational mode of

comportment. Any such view would give us reason for being hospitable to the thought that in intentional interpretation, we not only trace an agent's responses to certain, programming antecedents, we often trace it to antecedents whose realisation means that the responses are rationally required of the agent.

3. Interpretation as interpretation

Is every sort of normalizing explanation going to count as interpretation? Of course not. The explanations which we invoke for the adding machine's responses need not count as interpretations. And certainly the explanations which we give for the responses of the human body in a sauna or in a cold shower will not count as interpetative. But is an intentional, normalizing explanation of a human being's responses bound to count as interpretation? 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 which 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 does, and indeed every revision of probability he undergoes, 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 (Pettit 1991). We find that in every thing the person does, he maximises expected utility: 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 reponses. 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 an adequate intentional interpretation. Consistently with displaying the patterns that invite the decision-theoretic explanations, the agent could be a creature which does not go through any conscious ratiocination. He might be a sort of automaton who enjoys such a superb design that exposed to appropriate evidence, he revises his degrees of belief in the rational way and, presented with any range of options, he forms degrees of desire, and chooses according to strength of desire, in the rational way. He might never have to think about the import of the new evidence put before him, weighing its significance in the balance with more familiar facts. And he might never have to deliberate about the options which he faces, trying to determine their relative attractions and trying to establish which is the most desirable.

Short of introducing the possibility that there is no way that the Bayesian subject thinks, we can already see that decision-theoretic explanation is silent on how things present themselves within the forum of the agent's attention. The explanation does not suggest, on any plausible reading, that the agent thinks explicitly in terms of his own probabilities and utilities; it is not clear how he would even know what these are, given the detail involved (Harman 1986, Chapter 9; Pettit 1991). And the explanation leaves it entirely dark as to how the agent reasons otherwise. Suppose that his degrees of belief that p and that q lead him, 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 him, rationally, to form a certain degree of desire that u. How is the agent supposed to think as he reasons his way, however implicitly, to the conclusion that leaves him with the appropriate degree of belief that r or degree of desire that u? Presumably the agent holds the objects of his grounding beliefs—'p' and 'q'—before his mind. But how does he register the partiality of his beliefs in these objects? Presumably, again, he holds the objects of his grounding desires—'s' and 't'—before his mind. But how does he register the fact that he desires those propositions rather than believing them? We are left entirely in the dark.

The pattern of decision-theoretic explanation which 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 compelling when we recognise that there is a more common-or-garden sort of intentional 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 (Pettit and Smith 1990).

Consider a case where an agent 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 his 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 an *entree* to his cast of mind.

That an agent has a certain degree of belief in something may be a brute fact about him; or it may reflect a sensitivity to the similarity between the matter in question and some more familiar sort of happening; or it may come of a judicious weighing of a variety of signals, in the style of Sherlock Holmes; or whatever. Similarly, that an agent has a certain degree of desire for a certain prospect may be a brute fact, reflecting a primitive urge; or it may come of his recognising that it represents the promise of fun and amusement; or it may be generated by his seeing it as the only way he can remain faithful to his duty; and so on. In invoking degrees of probability and utility, however comprehensive, the decision-theoretic story leaves us in the dark about these matters. By contrast, the more regular sort of

narrative would open up the agent's mind to us. It would let us in on the secret of how he sees things and why, in the light of that perception, his reponse is a more or less rational one to evince.

The drift of these considerations will be obvious. If we are to speak of interpretation, then it seems appropriate to reserve the word for the sort of intentional explanation which not only serves to provide a programming and normalizing gloss on an agent's responses but which also enables us to get in on the agent's way of seeing things: to access his mode of reasoning and deliberation. There is a lot more to say about the nature of interpretation in this sense: about its presuppositions and about its methods (Pettit 1993, Chapter 5). But we need not go into those matters now. The important point is that there is reason to treat interpretation as involving more than program explanation and more, even, than explanation of a normalizing variety. Useful though those categories are, they are not sufficient in themselves to differentiate this elusive form of understanding.

RESEARCH SCHOOL OF SOCIAL SCIENCES
AUSTRALIAN NATIONAL UNIVERSITY
CANBERRA, ACT 2601
AUSTRALIA

REFERENCES

Cherniak, Christopher (1986) *Minimal Rationality*, M.I.T. Press, Cambridge, Mass..

Ginev, Dimitri (1994) 'Beyond Normalizing Explanation', *Philosophia*, 23, 1994 (*this issue*).

Harman, Gilbert (1986) *Change in View*, M.I.T. Press, Cambridge, Mass..

Jackson, Frank and Philip Pettit (1988) 'Functionalism and Broad Content' *Mind*, 97, 1988, 381-400.

Jackson, Frank and Philip Pettit (1990a) 'Program Explanation: A General Perspective', *Analysis*, 50, 107-17.

Jackson, Frank and Philip Pettit (1990b) 'Causation in the Philosophy of Mind', *Philosophy and Phenomenological Research*, 50, 1990, 195-214.

Lewis, David (1986) *Philosophical Papers*, Vol 2, Oxford University Press, New York.

Pettit, Philip (1986a) 'A priori Principles and Action-Explanation', *Analysis*, 46, 39-42.

Pettit, Philip (1986b) 'Broad-minded Explanation and Psychology', in Philip Pettit and John McDowell, eds, *Subject, Thought, and Context*, Oxford University Press, Oxford, 17-58.

Pettit, Philip (1991) 'Decision Theory and Folk Psychology', in Michael Bacharach and Susan Hurley, eds, *Foundations of Decision Theory*, Blackwell, Oxford.

Pettit, Philip (1992) 'The Nature of Naturalism', *Proceedings of the Aristotelian Society*, 63.

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

Pettit, Philip and Michael Smith (1990) 'Backgrounding Desire', *Philosophical Review*, 99, 565-92.

NOTE

[1] I am grateful to Ian Ravenscroft and Michael Smith for helpful comments.

TOE WIGGLING AND STARTING CARS: A RE-EXAMINATION OF TRYING

O. H. GREEN

1. Introduction

In manuscript 166 Wittgenstein remarks, "We are inclined to look for an activity when we are to give an account of the meaning of a verb and if some activity is closely connected with it we tend to think that the verb stands for this activity." This, of course, can lead to trouble. To understand trying, we naturally consider failed actions. In many cases, an activity associated with trying is evident. A woman tries to start her car: the car does not start, but she does turn the key. Not all cases are like that. Instructed to wiggle his toes a newly paralytic man fails to get them to move even a little bit. This case typically sends action theorists scurrying in search of an action or action substitute. This is a wild goose chase. I don't mean to suggest that these action theorists are simply victims of linguistic bewitchment. At issue in their pursuit is the nature of trying, intending, and volition.

2. Trying and Failed Action

While attempts may issue in success, cases in which one tries and fails are of special interest to philosophers concerned to understand trying. This is not because we tend to talk of trying in cases involving at least some likelihood of failure; that is, after all, only a fact about talk of trying. These cases are significant because they provide a contrast not only with cases in which one does what one tries to do but, even more importantly, with cases of simple inaction as well. The case of the woman who tries to start her car is clearly different from one in which she does not even approach the car. She would have