CHAPTER EIGHT

COGNITIVE PSYCHOLOGY

I

Is there something wrong in principle with the major direction of cognitive psychology these days? A blindness which comes from a too great confidence in its rationalist and mechanist assumptions?

Take our everyday performance, like catching a ball, or carrying on a conversation. The current mainstream in cognitive psychology sees as its task to explain these by some underlying process which resembles a computation. When we reflect, we are struck by the skill we exhibit in these performances: knowing just where to reach to intercept the ball, knowing just where and how to stand, what tone to adopt, what nuance of phrasing to use, to respond appropriately to what our interlocutor has said. To explain the performance would then be to give an account of how we compute these responses, how we take in the data, process them, and work out what moves to make, given our goals.

To reach an answer by computation is to work it out in a series of explicit steps. The problem is defined, if necessary broken up into subproblems, and then resolved by applying procedures which are justified by the definition. We resort to computation sometimes when we cannot get the answer we want any other way; and sometimes when we want to show that this is the right answer. Explicit procedures can be crucial to a justification of our result.

But in the case of skilled performances like the above, we are not aware of any computation. That is not what we are *doing*, in the sense of an activity that we are engaged in and could be made to avow and take responsibility for, granted undistorted self-knowledge. The computation would have to be an underlying process, on a par with – or, indeed, identical with – the electrical discharges in brain and nervous system.

The nature of the activity as *we* carry it on is in some respects antithetical to a computation. When we catch the ball, or respond appropriately to our neighbour's conversational opening over the back fence, we make no explicit definition of the problem. Indeed, we would be very hard put to it make one, even if we set ourselves the task, and might find it beyond our powers. There is correspondingly no breakdown into subproblems, or application of procedures. We operate here, as in most contexts, with the task implicit, that is, not expressly formulated. That is part of what people mean when they say that we are applying know-how here, not explicit knowledge.

Our awareness of our activity shows that computation is something we do sometimes, not all the time. But more, we can see that it is quite beyond our powers to do it all the time. It is not just that there are some performances where an explicit definition of the problem seems beyond our capacity. The fact also is that every activity of computation deploys skilled performances which are themselves not explicitly thematized, and cannot be right now without disrupting the computation in train. As I define my problem in some explicit formulation, I draw on my capacity to use language, build declarative sentences, zero in on salient issues, and others again, which I have to leave tacit for the moment while concentrating on the matter at hand. In different ways, Wittgenstein and Polanyi have made us aware of this inescapable horizon of the implicit surrounding activity, which the latter discusses in terms of 'tacit knowing'.

On the other hand, it is clear that we can only match these performances on a machine by defining the problems explicitly and building the machine to compute. There is nothing comparable to tacit knowledge in a machine. The fact that we have in the last half century developed the theory of such machines, and then made considerable progress in building them, has been of great moment for psychology. It has given rise to a new explanatory paradigm, which seems to offer the hope of a materialist theory of behaviour, which would not be as idiotically reductive as classical behaviourism.

It is the prestige of this paradigm – and the strength of the underlying commitment to a mechanistic materialism – which powers cognitive psychology; that, and the continuing influence of the epistemological tradition of rational reconstruction. We are after all material objects, susceptible like all others to some mechanistic explanation: so runs the reasoning. We are moreover material objects which bring off these extraordinary performances of ball catching and conversation. Where more plausible to look for an explanation than in that other range of things which we design to realize (supposedly) comparable performances, viz., computing machines?

So cognitivists feel justified in ignoring the deliverances of self-understanding of agents, which cannot but draw a distinction between computations and the exercise of tacit skills, and plumping for the machine paradigm.

Is this move justified? This is what I would like to explore in this paper. More generally, I want to ask whether features which are crucial to our self-understanding as agents can be accorded no place in our explanatory theory. Is this extrusion a justified move to a properly scientific theory, or is it rather a way of side-stepping the important explanatory issues? Of 'changing the subject', as Davidson puts it?

The implicit/explicit distinction is one such important feature. Before grappling with the main question, I would like briefly to introduce another. This resides in the fact that human beings are self-interpreting animals. This means among other things that there is no adequate description of how it is with a human being in respect of his existence as a person which does not incorporate his self-understanding, that is, the descriptions which he or she is inclined to give of his emotions, aspirations, desires, aversions, admiration, etc. What we are at any moment is, one might say, partly constituted by our self-understanding.

This is another feature unmatched by machines; or we find in them only the weakest analogies. A computer may indeed be monitoring some of its own operations, and this may seem an analogue to self-understanding. But in this case, there is a clear distinction between description and object. The operations are there independent of the monitoring, even if this may bring about other operations which interact with them.

But in our case, our self-understanding shapes how we feel, for example, in such a way that there is no answer to the question 'What is our state of feeling?' independent of our self-description. The analytical distinction description/object cannot be made.

To look at some examples: I love A, I admire B, I am indignant at the behaviour of C. My love for A is, let us say, not just a momentary élan; it is bound up with the sense that our lives are permanently or definitively linked, that being with her is an essential part of being who I am. Now these last clauses constitute a description of how I feel. They are not just predictions or counterfactuals based on what I feel. I am not just predicting, for example, that were we to separate, I should feel terrible, or be at a loss. What I am doing is describing the quality of my emotion, which is quite different in what it is and how it feels from other kinds of attachment which lack this defining character.

And the quality of the emotion is essentially given by this description; that is, having this emotion is defined in terms of being inclined to give this kind of description. This is not to say that there are not cases where one might love in this way and not be ready to describe one's feelings in this way. One might only later come to recognize that this sense of lives being joined was the essential character of one's feeling. But in attributing the feeling to oneself even before one was ready to speak in this way, one is still saying something about one's self-understanding. When I say that I loved her in this way last year, before I came to understand properly how our lives are bound up together, I am grounding myself on the sense I had then of her importance to me, and I am purporting to give a more adequate characterization of that sense. Presumably, I was even then making plans which involved a life-time together, or committing myself to some long-term path, and it is in virtue of *that* that I can say now: 'I loved her in this way then.'

Put another way, I could not attribute this kind of love to an agent who was incapable of having in any form whatever the sense of being bound to someone for life. That is why we cannot attribute many human emotions to animals. Some animals do in fact mate for life; but they cannot have the kind of love we are talking about, because this requires the sense that it is for life, and therefore the possibility of making a distinction between the passing and the permanent.

Thus even before we are fully conscious of it, this emotion is characterized essentially by our self-understanding, by the sense we have of the meaning of its object to us. Similar points could be made in relation to admiration and indignation. We admire someone whom we think is great, or exceptional, or exhibits some virtue to a high degree. This emotion is defined by this kind of understanding of its object. And once again this does not prevent us ascribing admiration to people who do not recognize their favourable judgements of those they admire. I recognize now that I not only felt well-disposed towards B, but I also admired him. I did not want to admit it at the time, because I have trouble avowing that I grudgingly recognize a virtue in his way of being. But I acknowledge that I have all along, and *therefore* that I admired him then, albeit without recognizing it. I can only attribute the admiration retrospectively because I attribute the virtue-judgement retrospectively as well. I see it there, in the things I thought and said and did, even though I did not allow it its right name. A parallel point could be made about the judgement encapsulated in indignation, that the object of our feeling has done some flagrant wrong.

In this way our feelings are constituted by self-understandings; so that, as I said above, the properly human feelings cannot be attributed to animals; and some feelings are specific to certain cultures. But all this occurs in a way which defeats any attempt to distinguish description and object. If one searches for some core of feeling which might exist independently of the sense of its object which constitutes it, one searches in vain. More, the very nature of human emotion has eluded one. An emotion is essentially constituted by our sense of its object, that is, by what we are inclined to say about its significance for us. That is what is contained in the slogan that human beings are self-interpreting animals: there is no such thing as what they are, independently of how they understand themselves. To use Bert Dreyfus' evocative term, they are interpretation all the way down.

This is a second feature of ourselves, as we understand our activity and feeling, which has no machine analogue. And in fact the two features are linked.

I was arguing above that we can have a certain emotion before we are ready to apply what we can later recognize as the essential description. We can make these later attributions in virtue of what descriptions we *were* ready to make, which we retrospectively understand as expressing the sense of things which is properly encapsulated in the essential description. Our emotions can be better or less well understood by ourselves, can be more or less explicitly formulated. We might want to say: 'Yes, I loved her in this way before, but it wasn't explicitly formulated for me as it is now; it was still something implicit, unsaid, unrecognized.'

But this transposition from the implicit to the explicit is an important one. The emotion itself changes. An emotion clarified is in some way an emotion transformed. This is a corollary of the fact that emotions are constituted by self-understandings. And it will typically play a crucial part in our explaining someone's behaviour that he did not explicitly understand what he was feeling, or perhaps that at the crucial moment he began to understand explicitly.

In other words, because of our nature as self-interpreting animals, the quality of our self-understanding plays an important role, and the distinction implicit/explicit has a crucial explanatory relevance. That is, it has relevance in the understanding we have of ourselves as agents and subjects of feeling.

Now there is no analogue of this in computing machines. The connected features of self-interpretation and a partly implicit sense of things have no place there. Should this worry us in adopting such machines as paradigms for the explanation of human performance?

II

No, say the cognitivists; why should this worry us? It would not be the first time that the way things look to the uninstructed eye, or to ordinary consciousness, or to common sense, turns out to be misleading. The progress of science is littered with such over-rulings of appearance. Right back at the beginning, we had to disregard the fact that the sun seems to go around the earth, that moving objects feel as though they stop when we cease to exert effort. Now we recognize that the four-dimensional space–time continuum of ordinary awareness is crucially different from the one invoked in physics. Closer to home, we learn that the pain in my arm really comes from a malfunction of the heart; that the pain I feel in the area of the heart in my hypochondriac panic really comes from my pectoral muscles. Why should the case be any different with acting, thinking, feeling?

The answer to this (would-be rhetorical) question might be that the supposedly phenomenal features of action and understanding are a crucial part of the explanandum.

An objection that comes to mind right away to the proposal that we explain human skilled performance in terms of underlying processes resembling those of computing machines is this: the two kinds of process differ in what looks like a crucial respect.

We do attribute some of the same terms to both humans and machines. We speak of both as 'calculating', or 'deducing', and so on for a long list of mental performance terms. But the attribution does not carry the same force in the two cases, because we cannot really attribute action to a machine in the full-blooded sense.

Why do we want to say that a machine computes, or for that matter that a machine moves gravel, or stacks bottles? Partly because the machine operates in such a way as to get these tasks done in the proper circumstances. But also, and more strongly, in the case for example of computers, because the way the machine gets these tasks done has a certain structural resemblance to the way we do them. Characteristically, the machine's operation involves breaking down the task into sub-tasks, the fulfilment of which can be seen as logical stages to the completion of the computation; and this breaking down into sub-tasks is essential to what we call computation when *we* compute – you would not say someone was computing, if he gave the answer straight off without any analytical reflection.

More generally, to borrow Fodor's formulation, we can see a physical system as a computational device, if we can map different physical stages of that system on to formulae of some language of computation, in such a way that the causal relations between the physical states match the logical relations among the formulae.[1] 'The idea is that, in the case of organisms as in the case of real computers, if we get the right way of assigning formulae to the states it will be feasible to interpret the sequence of events that *causes* the output as a computational *derivation* of the output.'[2]

Thirdly, we say that a machine 'does' something when we have designed it to accomplish the task. All three factors apply in the case of computers; at least two in the case of bottle-stackers. But there could be objects which we would describe as *phi*-ing just because they were very useful at accomplishing the task of getting something *phi*-ed, even though they were discovered in nature and not manufactured.

But it is clear from this that the attribution of an action-term to such artefacts or useful objects is relative to our interests and projects. A machine *phi*s because we have manufactured it to *phi*, or we use it to *phi*, or we are interested in it in respect of the *phi*-ing it gets done. If we ask why we want to say that it is *phi*-ing and not *psi*-ing, where 'things being *psi*-ed' is a description of some other outcome of the machine's operation (our computer also hums, heats up the room, keeps George awake, increases our electricity bill), the answer is that *psi*-ing is not what we use it for, or what we built it for.

Of course we normally would say quite unproblematically that the machine hums, heats the room, and so on; but where we want to make a distinction between what it is really engaged in, as against just incidentally bringing about (it is a computer, dammit, not a room-heater), we do so by reference to our interests, projects, or designs. A changed economic picture, or the demands of a new technology, could make it the case that the *psi*-ing was suddenly a very important function, and then we might think of the same machine as a *psi*-er and as *psi*-ing (provided it also was an efficient device for this end). Indeed, we could imagine two groups, with quite different demands, sharing time on the same device for quite different purposes. The computer also makes clicks in strange patterns, very much valued by some eccentric group of meditation adepts. For them, the machine is a 'mantric clicker', while for us it is computing payrolls, or chi-squares.

But what is it *really* doing? There is no answer to the question for a machine. We tend to think in this case that it is really computing, because

[1] J. A. Fodor, *The Language of Thought* (Hassocks, Sussex, 1975), p. 73. [2] *Ibid.*

we see it as made for this purpose, and only by accident serving the purpose of helping meditation. But this is a contingent, external fact, one external, that is, to the machine's make-up and function. It could have been designed by some mad yogi with a degree in electronic engineering, and just happen to serve as a computer. Or it could just have come into existence by some cosmic accident: a bolt of many-tongued lightning fused all this metal into just the structure needed to fulfil both these functions.

So attributions of action-terms to such devices are relative to our interests and purposes. As Fodor puts it: 'it is *feasible* to think of such a system as a computer just insofar as it is possible to devise some mapping which pairs physical states of the device with formulae in a computing language in such a fashion as to preserve the desired semantic relations among the formulae'.[3] And he adds later: 'Patently, there are indefinitely many ways of pairing states of the machine with formulae in a language which will preserve [the right] sort of relation.'

But the same is not true of ourselves. There is an answer to the question, What is he doing? or What am I doing? – when it is not taken in the bland form such that any true description of an outcome eventuating in the course of my action can provide an answer – which is not simply relative to the interests and purposes of the observer. For action is directed activity. An action is such that a certain outcome is privileged, that which the agent is seeking to encompass in the action.

This purpose may be unconscious, as when my awareness of certain desires is repressed; it may be partly unformulated, as when I walk in such a way as to avoid the holes in the pavement while concentrating on something else; it may be at the margins of attention, as when I doodle while talking on the phone. But in all these cases, our willingness to talk about action depends on our seeing the activity as directed by the agent, on their being such a privileged outcome, which the agent is trying to encompass. This is the basis of the distinction between action and non-action (e.g., events in inanimate objects, or reflex-type events in ourselves, or lapses, breakdowns, etc.).

So in contradistinction to machines, we attribute action to ourselves in a strong sense, a sense in which there is an answer to the question, What is he doing? which is not merely relative to the interests and purposes of an observer. Of course, there are issues between different action-descriptions which may be settled by the interests of the observer. For any action may

[3] *Ibid.*, my emphasis.

bear a number of descriptions. Notoriously, there are further and more immediate.purposes, broader and narrower contexts of relevance. So, we can say severely, 'I know you just wanted to do the best by him, but did you physically prevent him leaving the house?', or 'I know you only meant to scare him, but did you shoot the dog?' Here we have classic examples of the distinction between the description which is salient for the agent, and that which is crucial for someone assessing his conduct.

But however great the interest-relative variability in the description of what I do, a distinction can be drawn among the outcomes that eventuate in my action between what I *do* under whatever description, and the things that cannot be attributed to me at all in any full-blooded sense. This distinction is not observer-, or user-, or designer-relative; and that is the difference with machines.

Thus there are descriptions of things which get done when I act which I can repudiate as action-descriptions: for example, that I move molecules of air when I talk, or even give clicks with my teeth which are highly prized by the eccentric meditation circle. We can imagine that they hire me to come and give lectures in philosophy, and I am puzzled why they keep inviting me back, because they do not seem interested in what I say, and indeed, sink into a deep trance when I talk. There is some sense in which 'putting them to sleep' is an action-description applying to me; but we recognize that this applies in a quite different way than, for instance, the description 'lecturing on philosophy'; and hence we have a barrage of reservation terms, like 'unwittingly', 'inadvertently', 'by accident', 'by mistake', and so on.

Now *this* distinction, between what I am full-bloodedly doing, and what is coming about inadvertently, is not relative to observer's or designer's interests and purposes. Unlike the case of the artefact, it remains true of me that what I am doing in the full-blooded sense is lecturing on philosophy, and not mantric clicking; even though I may be much more useful as a device to accomplish the second end than the first, may do it more efficiently, and so on; or even though everyone else becomes interested in mantric clicking, and no one even knows what philosophy is any more besides me.

Nor can we account for this difference by casting me in the role of crucial observer, and saying that the crucial description is the one relative to my interests. For this neglects the crucial difference, that with the artefact the observer's interests are distinguishable from the machine. So that it makes sense to speak of a machine as surviving with its functioning intact even when no one is interested any more in its original purpose, and it serves quite another one, or none at all. But an action is essentially constituted by its purpose. This is a corollary of the point above, that men are self-interpreting animals. The attempt to make a comparable distinction to the one we make with artefacts, between external movement and some separable inner act of will, breaks down, as is now notorious; for the inner act shrinks to vanishing point. Our ordinary conception of an act of will is parasitic on our ordinary understanding of action.

So mental performance terms, like 'calculating', have a different sense when attributed to artefacts than when attributed to humans. In the latter case, we mean to describe actions in the strong sense, in a way which is not merely observer- or user- or designer-relative. Let me say quickly, as a sort of parenthesis, that this represents as yet no decisive objection against cognitivism; it just puts the issue about it in clearer perspective. It is a point about the logic of our action-attributions. It does not show by itself that what goes on when people calculate is something very different than what goes on in computers. For all we can say at this stage, a computer-type, observer-relative 'calculation' may underlie every act of calculating; and it may provide the best explanation for our performance.

The point is only that our language of action attributes something quite different to us agents, viz., action in the strong sense; something for which there is no basis whatever in machines, or in the functioning of the organism understood analogously to that of a machine; and indeed, for which one cannot easily conceive of any basis being found in a machine.

I have made the point in terms of action, but the same point goes for other 'functional' states of machines in contrast to ourselves. We might try to find states of machines which parallel our desires and emotions. A machine might be said to 'want to go' when it is all primed, and started, and only being held back by a brake, say. But it is clear that an analogous distinction applies here to the one in the case of action. What the machine 'desires' is determined by the observer's interest or fiat, or that of the user or designer; while this is not so for the human agent. Actually, the temptation does not even exist here, as it does in the case of action, to apply such terms to machines, except as a kind of anthropomorphic joke.

This is because the crucial difference is even more evident here than in the case of action. For the clear upshot of the above discussion is that human and animal agents are beings for whom the question arises of what significance things have for them. I am using the term 'significance' here as a general term of art to designate what provides our non-observer-relative answers to such questions as: What is he doing? What is she feeling? What do they want?

Ascribing action in the strong sense to some being is treating that being as a subject of significance. The full-blooded action-description gives us the action as purposed by the agent. We define the action by the significance it had for the agent (albeit sometimes unconsciously), and this is not just one of many descriptions from different observers' standpoints, but is intrinsic to the action *qua* action. So we can only attribute action to beings we see as subjects of significance, beings for whom things can have significance in a non-observer-relative way.

We have to add this last rider, because there is, of course, another, weaker sense in which we can speak of things having significance for inanimate beings: something can be dangerous for my car, or good for my typewriter. But these significances are only predicable in the light of extrinsic, observer-relative or user-relative purposes. By contrast, the significances we attribute to agents in our language of action and desire are their own. It is just the principal feature of agents that we can speak about the meanings things have for them in this non-relative way, that, in other words, things *matter* for them.

Let us call this essential feature of agents the 'significance feature'. Then the crucial difference between men and machines is that the former have it while the latter lack it.

This difference is less immediately evident to us in wielding our action-descriptions, or at least some of them. For action-descriptions focus our attention on what gets done; that action is directed by the agent is usually subsidiary to our main point of concern. Thus we have no trouble applying action-terms in a weaker sense to inanimate things. But desire- or feeling-descriptions focus our attention directly on the significance things have for the agent. That is why there is something strained or metaphoric in applying these to machines.

The strain gets even greater when we come to emotion terms. We might speak of our car as 'raring to go', because at least 'going' is something it is capable of, albeit in a weak, user-relative sense. But when we get to an emotion term like 'shame', we could not have even the remotest temptation to apply it to the inanimate.

'Shame' is in fact intrinsically bound up with the significance feature – one might say, doubly bound up. It is not just that to attribute shame is to say that the situation has a certain significance for the agent: it is humiliating, or reflects badly on him, or something of the kind. It is also that the significance or import of the situation is one which only makes sense in relation to beings with the significance feature.

This contrasts with an import like danger. My car can be in danger, if there is a rock about to fall on it, for instance. This is, of course, a user-relative attribution: the danger is only to it in its function as car; *qua* collection of metal and glass bits, the rock represents no danger. But at least the attribution user-relatively makes sense. A car *qua* car can be in danger.

But 'shame' points to a different kind of import. Someone can only experience shame who has a sense of himself as an agent. For I feel ashamed of what I am like/how I appear as an *agent* among other agents, a subject of significance among others. It may seem sometimes that the immediate object of my shame is some physical property that a non-agent could bear. I may be ashamed of my small stature, or outsize hands. But these can only be objects of shame because of their significance: small stature means being overshadowed, failing to have a commanding presence among others; outsize hands embody indelicacy, lack of refinement, are proper to peasants.

The import of danger can be physical destruction, and this can happen to a car *qua* car. But the import of shame touches us essentially as subjects of significance. It makes no sense to apply it to any but agents (and not even to all of them; not to animals, for instance).

The significance feature is crucially bound up in our characterization of ourselves as agents. It underlies our attributing action to ourselves in a strong sense, as well as our attributions of desire and feeling; and reference to it is essentially involved in the definition of our emotions. With these, it is not just a matter of our attributing them to ourselves in a stronger sense than to inanimate objects; these concepts cannot get a grip on non-agents, even in a metaphorical manner. They only make sense in relation to us. In a world without self-aware agents, they could have no senseful application whatever.

The significance feature underlies the two features I singled out in the first section. We have these two, interpretation and the implicit/explicit distinction, because we are agents with a linguistic capacity, a capacity to formulate the significance things have for us.

But to formulate the significance of something, to make it explicit, is to alter it, as we saw above. This is because we are dealing with agents, subjects of non-observer-relative significance. My making explicit the danger my car is in does not alter the import of the situation for it; but my coming to see clearly the import of my situation for me can be *ipso facto* an alteration of its significance for me. Our being agents is a condition of our self-interpretations being constitutive of what we are; and it is because these interpretations can be explicitly formulated that the distinction implicit/explicit plays a crucial role for us.

These three features are closely connected, and are essential to us in our understanding of ourselves as agents.

III

Let us return to the main issue. Should the fact that our ordinary self-understanding attributes to us features which have no place in the computing machine paradigm make us wary of this in explaining human performance? Or can we dismiss these features as misleading surface appearance, on all fours with the sun's apparent movement around the earth?

At least a prima facie objection arises to just dismissing them. Are they not an essential part of what we have to explain? This objection could be spelled out in the following way. We are asked to believe that some behaviour of ours in computing, or some behaviour which involves no computing but involves skilled selection of response, is to be explained on the same principles as those accounting for the operation of a computing machine. This is pressed as an overwhelmingly plausible line of approach, given the similarity in outcome given the (physically defined) input.[4] It appears plausible, in other words, because we seem to be able to apply terms like 'computing', 'figuring out the answer', 'finding the solution', to the machine which we also apply to ourselves. If they do the same things as us, perhaps they can show us how to explain what we do.

But, the objection goes, they do not do the same thing as us, or only within the range of the analogy between weak and strong action-attributions. They do something we can call 'computing' in a weak, observer-relative sense of this term; which relative to another observer might be described as 'mantric-clicking'. We do something we call 'computing' in the strong sense, not observer-relative. How can we be so sure that an underlying process describable by the weak sense explains the overt action described in the strong sense? These are after all, *very different*, distinguished by everything that divides things possessing the significance feature from things without it. 'Computing' engines present some analogies to computing people, but they offer as yet no hint of how one might account for this salient feature of the latter, that they are agents,

[4] This claim involves a big promissory note, because there are all sorts of performances by us we have not even begun to match on machines, but I will not take this up here. For cogent objections, see H. L. Dreyfus, *What Computers Can't Do* (New York, 1979).

and act. Indeed, it has been essential to their utility that we can understand and operate these machines without reference to the significance feature.

Those who are nevertheless sure of the machine paradigm must be grounding their confidence on the belief that we can somehow ignore the significance feature. Why? Well, presumably because of the analogy I mentioned above with the misleading appearances which the progress of science has had to ignore. Cognitive psychologists are frequently dismissive of arguments of phenomenologists on the grounds that phenomenology can be very misleading as to underlying structure. The implication is that phenomenology gives us surface appearance, not anything about the nature of the explanandum.

The assumption underlying this dismissive attitude must be that the significance feature is a misleading surface appearance, like the movement of the sun, or perhaps a purely phenomenal one, like phenomenal colour or felt heat, to be set aside in any rigorous characterization of the events to be explained. This gets to seem a plausible view the more we repeat to ourselves that computing machines compute. The difference between computing and 'computing', between real and observer-relative performances, comes to seem a rather secondary matter. The significance feature comes to seem like a pure matter of the inner feel, something to do with the way the whole process is experienced from the inside, or perhaps, at best, a tag of honour we accord to agents, that they bear their predicates non-relatively; but in no case an important defining feature of the explanandum.

But this is, of course, mad. There is all the difference in the world between a creature with and one without the significance feature. It is not just a detachable feature that action has in some medium of internal representation, but is essential to action itself. The supposedly secondary, dispensable character of the significance feature disappears once we cease to dwell on that small range of actions which have plausible machine analogues. Once we look to feelings, emotions, or actions which are defined in terms of them, or of moral categories, aesthetic categories, and so on, like 'saving one's honour' or 'telling the truth', we run out of machine analogues to be bemused by.

Or if we are still bemused, it is because we are in the grip of an old metaphysical view, one embedded in our epistemological tradition, which makes us see our awareness as an inner medium of representation, which monitors (partly and sometimes misleadingly) what goes on in our bodies and the world. This is the ghost of the old dualism, still stalking the battlements of its materialist successor-states.

Consciousness is primarily understood as representation (Foucault has shown – if that is the term – how central this notion of representation is to the modern epistemological tradition). As such it is separable from the processes which it monitors, or of which it is a symptom. If it plays any role in explaining these processes, it must be in interacting with them. Since interaction is ruled out on materialist assumptions, it cannot be allowed any explanatory role at all. It can only serve as a (possibly misleading) way of access to the processes which are the stuff of behavioural science.

On this view, the primary difference between us and machines is that we are clearly conscious and they do not seem to be. Even this latter is not entirely sure, and cognitive theories begin to hedge bets when they are dragged on to this terrain: perhaps after all one day machines will get sufficiently complex to have consciousness? And will we ever know?

The discussion here gets ragged and rather silly; a sign that we are on the wrong track. And so we are. For the crucial difference between men and machines is not consciousness, but rather the significance feature. We also enjoy consciousness, because we are capable of focussing on the significance things have for us, and above all of transforming them through formulation in language. That is not unimportant; but the crucial thing that divides us from machines is what also separates our lesser cousins the dumb brutes from them, that things have significance for us non-relatively. This is the context in which alone something like consciousness is possible for us, since we achieve it by focussing on the significance of things, principally in language, and this is something we *do*.

The crucial distinction to understand the contrast between us and machines is not mental/physical, or inner/outer, but possessing/not possessing the significance feature. Once we understand this, we can see that this feature cannot be marginalized as though it concerned merely the way things *appear* to us, as though it were a feature merely of an inner medium of representation. On the contrary, it plays an absolutely crucial role in explaining what we do, and hence defines the kind of creatures we are.

We can see this best if we look again at our emotions, such as the example of shame above. As beings capable of shame, we experience certain emotions, and we react in certain ways to our situation and to each other. This is not just a fact of how things appear to us inside; this is a crucial fact about how we are and what we do. This is evident in the fact that in order to explain our behaviour, we have to use emotion terms like 'shame' and corresponding situation descriptions like 'humiliating'. In accounting for what we do, there is no substitute for a language which only makes sense applied to beings with the significance feature, the language of shame, humiliation, pride, honour, dignity, love, admiration, and so on. It is as fundamental as that.

In other words, when we say that the significance feature is essential to our self-understanding as agents, we are not saying that it is inseparable from our representations in an inner medium, whose deliverances are as dispensable to an explanation of behaviour as our perceptions of the sun in the sky are to our account of the solar system. We are rather saying that once we understand ourselves as *agents*, rather than, say, as physical objects on all fours with others, including inanimate ones, we understand ourselves as beings of whom the significance feature is an essential character, as beings such that it is essential to what has to be explained, if we want to explain their behaviour.

Once we see this, we have to stop treating it as a matter of surface appearance, and the plausibility begins to dissipate that surrounds the notion that we can explain computing, and much else, by the 'computing' of machines.

But perhaps one more desperate measure is possible. Supposing we challenged our entire self-understanding as agents. Perhaps it is all systematically misleading. Perhaps the only way to explain what we do is to look at ourselves as machines, and explain what we do in the same terms.

This is a radical suggestion, and one which undercuts cognitive psychology from another direction. Its ambition is just to give an account in psychological terms, terms that apply peculiarly to human beings, and perhaps some animals, and that can be seen as developments or more rigorous variants of the terms we understand ourselves with in ordinary life. Cognitive psychology is looking for a relatively autonomous science of human behaviour. It would not be satisfied just with a science that entirely abandoned the psychological, and dealt with us simply in the language of physics, say.

But it is also a suggestion that does nothing to solve our problem. For we cannot abandon our understanding of ourselves as agents. This is bound up with our practice as agents. Self-understanding is constitutive, as we saw, of what we are, what we do, what we feel. Understanding ourselves as agents is not in the first place a theory, it is an essential part of our practice. It is inescapably involved in our functioning as human beings.

The significance feature is at the centre of human life, most palpably in that we come to understandings with people about the significance of things. There is no relationship, from the most external and frosty to the most intimate and defining, which is not based on some understanding about the meanings things have for us. In the most important cases, of course, one of the things whose significance is understood between us is our relationship itself.

That is why the significance perspective is not an arbitrary one among human beings, one way of explaining how these organisms work among other possible ones. It is not even primarily a theoretical perspective on our behaviour. We could not function as human beings, that is as humans among other humans, for five minutes outside this perspective.[5]

In other words, we could have no relations at all if we did not treat ourselves and others as agents. (But by this, I do not mean that we necessarily treat them ethically, or as ends in themselves. Even our exploitive behaviour in the vast majority of cases takes our victims as agents. It can be argued, however, that there is a profound connection between our status as agents and the validity of such moral precepts as those of Kant.)

We can put this another way, and say that this self-understanding as agents is part of the reality it purports to understand. That is why a science of behaviour in terms of physics alone, even should such a thing prove possible, would still not answer the legitimate questions which psychology sets for itself: what is it that underlies and makes possible our functioning as agents, and the self-understanding that goes with it?

But, to sum up the objection announced at the start of this section, it is not at all clear how the machine paradigm is going to help us with these questions either.

IV

But hold on. I do not think one can say flat out that the machine paradigm will not help us. Maybe it can produce some startling goods further down the road. What can be said is that it is not much more plausible than a number of other approaches; and that it only looks strongly plausible as long as you overlook the significance feature. And you only do *that*, I think, if you are still in the grip of the dualist metaphysic (even though

[5] I think this is what emerges from the very interesting analysis in P. F. Strawson, 'Freedom and resentment', *Proceedings of the British Academy*, 1962, reprinted in Strawson (ed.), *Studies in the Philosophy of Thought and Action* (Oxford, 1968).

transposed in a materialist key) which comes to us from the epistemological tradition.

Once you do see the importance of the significance feature, it is evident that computing machines can at best go some of the way to explaining human computation, let alone intelligent, adaptive performance generally. To be told that underlying my ball catching are patterns of firing in the cortex analogous to those in electronic computers gives me as yet no idea of how these can help to account for (non-observer-relative) *action*, producing as they do a quite different kind of operation in the machine. What we have to discover is how processes analogous to machine computations could combine with others to produce real action, if this paradigm is to have a future. And this is no mean task. Indeed, no one has the slightest idea even how to go about pursuing it. In this context, the glaring disanalogies between machine and human performance, for instance the features discussed in the first section, can no longer be dismissed as mere appearances. They are rather major challenges to the very legitimacy of the paradigm.

Machine-modelled explanations of human performance, of the kind cognitive psychology offers, would have to relate to this performance understood as action in the role of an underlying explanation. We have this when phenomena on one level are explained by a theory invoking factors at another level, where this second level offers us the more basic explanation. An explanation in theory T is more basic than one in T′, where the explanatory factors ultimate for T′ are in turn explained in T.

We can clarify the predicament of cognitive psychology if we lay out three types of cases of such underlying explanation.

Case 1. The descriptions made and factors cited at the higher level turn out to be confused or mistaken when we achieve the deeper level explanation. In this case, we have not so much an explanation as an explaining away. An historical example of this is the distinction in Aristotelian cosmology between the supra-lunar and incorruptible, and the infra-lunar and corruptible. This was important to explain a whole host of things, including why the stars above go in perfect circles. The whole thing was just a mistake, and what survives is just *appearances* which can be explained in terms of the new cosmology; but the crucial distinctions of the old one turn out to be unfounded. We can now explain why things *looked* that way, but we know they are not.

The higher level explanation is discredited, because the distinctions it draws do not in fact correspond to any genuine explanatory factors. The higher level operates with concepts and descriptions among which no explanatory factors are to be found. There never was a science here – just as if I tried to explain the movements of the planets in terms of their colours in the telescope. I might note all sorts of patterns, but I should never in a million years be able to explain why they move as they do. For the relevant factors are mass and distance.

Case 2. Here we have a genuine explanation on the higher level, which is the object of a more basic explanation on the lower. As an example: we explain the wood disintegrating into ash by its being put in the fire. But we can give a deeper explanation in terms of the kinetic energy of the molecules. This is more basic, in that it accounts for the regularity by which we explained things at the higher level. With the kinetic theory, we understand the why of heat-transmission in general, and can see now why the same effect could be produced by a laser, for instance; why similar effects do not flow from heating metal, and so on.

The higher level explanation is genuine; in this, the case differs from 1. But it is dispensable. The higher explanation can always be eliminated in favour of the lower without loss. The latter not only gives us more, but covers all the same terrain as the former. There are no factors explanatory of heating/burning phenomena which are available on the higher level in such terms as 'fire' or 'charring', for which there are no correlates on the lower level, which can do the same explanatory job in the context of a more comprehensive theory. So for explaining heat, there is nothing we do with our phenomenal language which we could not do better in the kinetic-theory language. The phenomenal language is indispensable for describing how things are for us in our everyday identification of things; we need them to identify things as they figure in our perceptions, but otherwise, for the purpose of scientific characterization of the domain, not at all.

Case 3. Here there is also a valid higher level explanation. And there is also a theory of underlying structures which helps us explain how things happen as they do, and gives us some of the conditions of the higher level events occurring as they do. But unlike case 2, we cannot dispense with the higher level descriptions for the purpose of explaining the phenomena of the domain concerned. Some of the crucial explanatory factors are only available at this level; or to put it negatively, they cannot all be identified at the lower level. To seek them all there would be as fruitless as seeking the factors explaining planetary orbits in their colours.

I do not have an incontestable example. Let me just offer one which is relative to our explanatory resources at the present time, without prejudging whether we will take things beyond this in the future or not. A fleet assembles for war. This is a pattern of ship movements. At what corresponds to the more basic level, these can be explained in terms of the operations of engines, pistons, screws, etc. This level is essential if we are to get an explanatory handle on some of the features of this pattern. For instance it is indispensable to explain why, in some cases, ships stopped and began to be tossed by the sea (cases of engine failure), why some ships went faster than others, why some took a circuitous route (e.g., to get more fuel), and so on.

But if you want to understand why they are steaming towards this pattern, you have to be told that war has been declared, and that they are forming the fleet for such and such an offensive action. You need to have recourse here to the 'highest level' language of policy and politico-military goals and intentions. If you remain on the lower, engine-room level, you will never identify the crucial factors, in the same way as the factors behind planetary motion could not be found in colour discourse.

I repeat that this example is relative to our present explanatory resources. It is not meant to *prove* that we could not discover one day some explanation on a neurophysiological level which would render our policy- and intention-descriptions dispensable. I am just trying to give a picture of a third possible case, which *may* turn out to have instances at the end of the day. Because, though no one can say that such a neurophysiological language of explanation is impossible, there is even less ground for assuring us that it must be there to be found. Case 3 may yet turn out to be the model for deeper level explanations of human behaviour. My hunch is that it will.

But forget my hunches. The point of this was to provide a typology in which to understand the possible relations of underlying explanations to our action account.

It is clear that case 1 has no application here. To say that it is analogous to the infra-/supra-lunar distinction amounts to saying that our classifications of events as actions are wholly illusory. But since the self-understanding of ourselves as agents is essential to our acting, this is a claim which must remain meaningless and preposterous to us. Really to see the distinction between action and non-action as like the infra-/supra-lunar one would be to be incapable of acting. This is not an alternative we need consider.

There remain 2 and 3. The assumption of cognitive psychologists seems to me that case 2 offers the appropriate model. The underlying explanation, in a language appropriate to computing machines, gives us all the explanatory factors; the action account presents things as they look to us.

The model here would be the kinetic theory in relation to a phenomenal account of heating and heat transmission.

But I have argued above that this claim is prematurely made. Certainly the machine paradigm at present does not offer any hint of how we could hope to discover all the explanatory factors in its terms. In particular, we do not have the foggiest idea how it might help us to account for the significance feature of agents. If we ever do manage to account for the significance feature in mechanistic terms, then we will indeed have instantiated case 2. But until that day – should it ever come – case 3 has got to figure as a very plausible contestant.

For in fact, that is where we are now. Underlying explanations, especially neurophysiological, can offer us more basic explanations of some important phenomena: of certain features of development, of differential capacities, of breakdowns, and a host of other matters. But to explain fully motivated behaviour, we need an account of what is happening in terms of action, desires, goals, aspirations. We have no metaphysical guarantee that after an immense series of discoveries, refinements, and breakthroughs, the basic structure of our explanations of ourselves will not still be the same: a variant of case 3. What purport to be assurances to the contrary are based on the illusions of traditional dualism.

On one reading of the term, case 2 can be called a case of reduction of the higher to the lower level. (In a more denigrating sense, we sometimes reserve 'reduction' for cases of 1.) On this reading, it looks as though I am classing cognitive psychologists as proponents of reduction, more particularly, reduction of psychology to some underlying explanation. But this they (or many of them) claim not to be.

We have only to look at Fodor's book.[6] In his first chapter, he defends the independence and viability of the psychological enterprise against both behaviourism and physicalistic reductivism. A reductivist relationship holds, Fodor argues, between a special science (like psychology) and a more general one (like physics), when the laws of the former can be linked to laws of the latter via correlation statements which are themselves law-like. The crucial feature of this relationship would be that the natural-kind terms of the special science, those in which its laws could be formulated, would be type-indentical with the natural-kind terms of the general science, that is, physics.

Fodor's characterization of reduction resembles case 2 above, in that the special science is dispensable – although perhaps he makes the

[6] Fodor, *The Language of Thought.*

requirement a bit too stringent in demanding that the correlations be all law-like.

Now Fodor thinks that this kind of reductive relation is very unlikely to hold between the sciences of man and physics. He takes an example from economics: Gresham's Law. It is surely extremely unlikely that all cases of monetary exchange of which Gresham's Law holds, that is, where moneys of different quality are in circulation, should all fall under the same physical description; or otherwise put, that the physical description of all such cases should exhibit a natural kind of physics. 'Banal considerations suggest that a physical description which covers all such events must be wildly disjunctive.'[7] Even if one should manage, at the moment when human society was about to go under, to survey all previous cases of monetary exchange, and find some vast baggy disjunction under which all these cases fit in physical terms, this would still fail to be a law; because it would not necessarily help at all in counterfactual cases. We would not be able to conclude that, if the universe had gone on for another year, the physical conformation of the monetary exchanges in it would have been such and so.[8]

But, Fodor argues, we do not need to espouse type-type identities in order to save materialism, science, and so on. It is sufficient to embrace what he calls 'token physicalism': 'the claim that all the events that the sciences talk about are physical events'.[9] This is compatible with the type of event that a special science picks out (like monetary exchange) being realized physically in an indefinite number of ways – so long as it is always realized physically.

Espousing token physicalism, and rejecting type-tvpe identities, allows for the special sciences deploying concepts which are unsubstitutable. The special sciences need these if they are to 'state such true, counterfactual supporting generalizations as there are to state'.[10] For if the natural-kind terms of a special science only correlate with loose, open disjunctions in another science, then we cannot state the laws explanatory of the events that the special science deals with in the other science. For to explain, to give an account of what happens, is to license counterfactuals; and open disjunctions by definition license no counterfactuals. The natural-kind terms of our special science are in this case unsubstitutable.

Another science may cast a great deal of light on the underpinnings of these natural kinds. In particular, it may lay bare important conditions of

[7] *Ibid.*, p. 15. [8] *Ibid.*, p. 16. [9] *Ibid.*, p. 12. [10] *Ibid.*, p. 25.

their functioning as they do; so that the other science may give us explanations of exceptions and breakdowns. But it cannot substitute for the special science; and in this sense, we can see the natural kinds this latter science designates as part of the furniture of things.

Fodor's description of the status of a special science like psychology fits my case 3. The special science is indispensable, because the crucial explanatory factors (read, natural kinds) are only discoverable on its level; on the lower level they are not identifiable. Just as, for example, the class of planets of a given mass form an indefinitely open disjunction described in colour terms, so do the cases of monetary exchange in physical terms.

But then surely I am wrong to tax cognitivists with reductionism, with taking case 2 as their model?

No, I am not; first, because we are not talking about the same things. When Fodor talks of the relation of psychology to physics, he is not talking about our account of ourselves as agents. His 'psychology' is an account of what we do in computational terms, and the reductive issue for him arises between an account at this level and one at the physical or neurological level. He is quite oblivious of the difference between an account in computational terms and one which characterizes us as agents with the significance feature.

Indeed, Fodor's thesis of the irreducibility of psychology emerges originally from a reflection on computing machines. It was the recognition that two machines might be the same in the program run on them, and yet be very different in their physical structures and principles, which gave rise to the notion that an account of what they do in computational terms could not correlate with general laws on a physical level.

This was the basis of the thesis known as 'functionalism' in psychology. But this was because it was simply taken for granted that a 'psychological' account of what we do would be a computational one analogous to those we apply to machines. Fodor clearly makes this identification. Part of his argument against reductionism assumes it. Even if there are neurological kinds coextensive with psychological kinds as things stand now, he argues, the correlations cannot be lawful. 'For it seems increasingly likely that there are nomologically possible systems other than organisms (viz. automata) which satisfy the kind predicates of psychology but which satisfy no neurological predicates at all.'[11]

The 'psychology' here is obviously not what I am talking about. What

[11] *Ibid.*, pp. 17–18.

we normally understand as the predicates of psychology, those which involve the significance feature, plainly do not apply to machines. Nor have we anything but the vaguest fantasies as to how they might apply to machines we design in the future. The 'kind predicates' of psychology which we might think it 'increasingly likely' that automata will satisfy are computational performance terms applied in their weak, observer-relative sense.

The psychology whose irreducibility Fodor is defending is one which is just as much a science of computing machines as of humans. It has nothing to do with our account of ourselves as agents. The difference between these he just ignores, most likely for some of the reasons discussed above, owing to the baleful influence of traditional dualism. So whatever the relation between the computational and physical levels, Fodor plainly construes that between the computational account and the one in terms of agency in a reductive way, on the model of cases 1 or 2.

Secondly, it is not so clear after all that Fodor really can carry through his account of the psychology–physics relation as a case 3. If a paradigm of this relation is to be found in computing machines, whose program can be matched by machines of different design, then it is not so clear that counterfactuals cannot be found at the more basic level.

For any given (physical) type of machine, there are no counterfactuals on the computational level that cannot be matched, and explained by counterfactuals at the engineering level. Counterfactuals like 'if the program had been changed in such and such a way, then ...', or 'if the problem had been posed in such and such a way, then ...', can be given a deeper level explanation in terms of the way the machine is wired, connected up, or whatever. If this were not the case, we would not be able to build, design, or improve such machines.

Of course, other machines can be constructed on other principles, such that the deeper level explanations would invoke quite different factors. One machine, let us say, operates electronically; the other is run by fuel and has gears. The underlying accounts will be very different. And there may be an indefinite number of such machines which we might design to run the same types of program.

This certainly shows that the level of program design is in some way essential to us, that we could not go about what we do if we were to abandon this level of discourse. But we could not go so far as to say that the crucial explanatory factors are unavailable on the lower level.

Contrast what seems plausible with the Gresham's Law example. It is not just that one case of monetary exchange with media of different quality will involve gold and silver, the next gold and bronze, the next dollars and Deutschmarks, the next old and new currency, and so on. Even in a given case, you cannot match counterfactuals on the economics level with those on the physical level. 'If people come to believe that the king is no longer adulterating the silver coinage, then gold will come back into circulation' corresponds to no counterfactual on the level of bodily movement, say, even if we restrict our attention to this context. People can come to believe this in all sorts of ways; they can be told in French or English, or in sign language; they can come across silver coins newly minted, that seem heavier; they can deduce it from the behaviour of merchants; and so on indefinitely.

We might complain that this comparison is unfair, that we have to draw the boundaries of a context narrower in the Gresham's Law case. But this just makes the difference more palpable. We do not know how to draw such boundaries in the monetary exchange case so as to make for stable relations of deeper level explanation. The ever-present possibility of original speech acts which inaugurate new extensions of meaning makes this impossible.

By contrast, in the domain of computing machines, there are such stable relations of more basic explanation in each context; and the boundaries between the contexts are clearly and unambiguously demarcated by the (physical) type of machine. We are never at a loss for lower-level counterfactuals to explain our higher-level ones. True, there are an indefinite number of such possible contexts of computation. But they are each clearly demarcated, and within each one the relation between levels of explanation conforms to case 2. The absence of match between natural-kind terms at the two levels of discourse can itself be explained in terms of a difference between kinds, viz., the types of machine.

This suggests that we ought to distinguish two questions: (a) Do the laws and licensed counterfactuals have the same scope on the two levels? and (b) Are there laws and licensed counterfactuals at all on the lower level? The answer to (b) may be affirmative, even while that to (a) is negative. In this case, it is not unambiguously true that reductive relations do not hold. This is the kind of case where we want to speak of systems which are analogous but not homologous. For each homologous class of machines, however, the reduction is a perfect case 2, and if this were the only domain we had to consider, it would never occur to us to question a reductivist construal.

But in a genuine case 3, the answer to (b) is a negative, and this is a quite different predicament.

Fodor seems to have elided what I have called case 3 and what I might call a multi-contexted case 2; and this may be connected with his having elided the two issues: the reduction of computational psychology to physics, and the reduction of our action understanding to computational psychology; or rather, his having invisibly subsumed the second question in the first. Because the second does seem to call for a case 3 solution, while the first seems to conform to this special kind of multiplex case 2.

But this is all part of his ignoring the issue around the significance feature, which amounts, I have tried to show, to a reductionism of a very strong kind.