Folk-Psychological Explanations

Jonathan Bennett

Before we can reasonably decide anything about the future of folk psychology we need a better grasp of what it is and how it works. Since folk psychology more or less defines our chief psychological concepts, exploring it is doing conceptual analysis. Many philosophers these days contemn conceptual analysis or condescend to it; I don't join them, but nor shall I argue with them here.

In this paper I hope to contribute a little to the understanding of folk psychology by setting out the reasons why the generalizations on which folk psychology rests are *explanatory*, reasons that do not require us to get mired in the question of whether those generalizations are causal.

1. Intentionality in simple systems?

We must start with the belief-desire-behavior triangle. The founding triangular idea is that a thinking system does what it thinks will bring about what it wants. Two of these three concepts are said to involve 'intentionality'; a better, because more explanatory, label 'cognitive teleology'—what a system has if it has thoughts that guide it to its goals.

The conceptual structure that this involves is illustrated by the behavior of a thermostat: the thermostat 'wants' the room to be warmer, 'thinks' that closing the switch will bring this about, and accordingly closes the switch. It is not illustrated by vending machines that have been used for that purpose by Ned Block. He describes a machine that will give you a Coke for a dime when it is in state S_1 and will give you a Coke for a nickel when it is in state S_2 (you get it from S_1 to S_2 by putting a nickel in), and he describes state S2 as a low-level analog of *desire for a nickel*.[1] This has nothing to be said in its favor. There is no truth of the form: 'When the machine is in state S_2 it does what it "thinks" will bring it a nickel', and so the most elementary, non-negotiable aspect of intentionality or cognitive teleology is absent. The same applies to the use of a vending machine in the one unsuccessful chapter of Dennett's latest book.[2]

Though thermostats are to be favored over vending machines, they should be approached gingerly. I don't side with those who get furious when Dennett writes indulgently of taking 'the intentional stance' towards a thermostat; on the contrary, there is something to be learned from doing just that. But there is also something wrong about doing it, as I shall now explain.

All the behavior of the thermostat that might be handled

[1] Ned Block, 'Troubles with Functionalism', in Ned Block (ed.), *Readings in Philosophical Psychology*, vol. 1 (Harvard University Press: Cambridge, Mass., 1980), pp. 268–305, at p. 271.

[2] Daniel C. Dennett, *The Intentional Stance* (M.I.T. Press: Cambridge, Mass., 1987), chapter 8, 'Evolution, Error and Intentionality'.

teleologically, or in intentional terms, is explained by a single mechanism, a single kind of causal chain that can be fully described without any use of intentional concepts. We can replace 'The thermostat does what it can to keep the temperature of the room close to 68 degrees' by 'The thermostat's switch closes whenever its temperature falls to 66 degrees and opens whenever its temperature rises to 70 degrees', and we can explain the latter generalization without any mention of 68 degrees as a goal and without mentioning beliefs and desires or anything like them.

In short, the one intentional account of the thermostat's behavior is matched by a single physicalistic account; and I submit that when that is the case, the latter account should prevail and the former, though perhaps stimulating and interesting for philosophical purposes, is false and should be rejected. For genuine teleology or intentionality, I contend, *the unity condition* must be satisfied. That is, a system x's intentionality is genuine only if

> Some class of x's inputs/outputs falls under a single intentional account—involving a single goal-kind G such that x behaved on those occasions because on each of them it thought that what it was doing was the way to get G—and does not fall under any one mechanistic generalization.

Where that is satisfied, applying intentional concepts to the system brings a conceptual *unity* to some set of facts about it—a set that is not unifiable under a mechanistic description.

The unity condition marks off the systems some of whose behavior falls into intentional patterns that are not coextensive with mechanistic patterns. Only if a system's behavior satisfies that condition, I contend, is it legitimate for us to exploit its intentional patterns in our thought and speech. The marking-off is of course a matter of degree. It rejects intentionality when the intentional pattern coincides with a single mechanistic one; it welcomes it when such a pattern utilizes thousands of different mechanisms; and it gives an intervening judgment—'intentionality in this case is so-so: permissible but not very good'—for many intermediate cases.

The fuzzy line drawn by the unity condition seems to correspond roughly with much of our intuitive sense of which systems do and which ones don't have thoughts and wants. Consider a chameleon flicking out its tongue and catching a fly with it. One can plausibly think of this as goal-pursuing behavior: it wants to eat the fly and thinks that this is the way to bring that about. But suppose we find that one uniform physical mechanism controls this pattern of behavior—a relatively simple causal tie between proximity of fly and movement of tongue, and between location of fly and direction of tongue movement, with, in each case, a few parameters in the one governing a few parameters in the other. Thoughtful people will regard this as evidence that the cognitive-teleological account of the behavior was wrong because really only a single mechanism was involved. The plausibility of the response 'Oh, so *that's* all it was!' is evidence for the truth of the unity thesis.

The thesis also corresponds to the best *defence* there is for using intentional concepts.

The question of the legitimacy of intentional explanations of behavior ought to be faced squarely. Since chemical explanations involve principles that go wider and deeper, and theoretically admit of greater precision, why should they not always be preferred to explanations in terms of thoughts and wants?

Some of the more libertine and 'instrumental' ways of talking about intentionality have given the impression that no justification is needed—that it is simply up to us to decide whether we want to talk and think in a certain way about

people and thermostats and vending machines and lecterns. I hope that nobody really believes that.

If justification is to be given, there are three prima facie possible ways of doing this. **(1)** The most completely justifying (were it true), but also the least credible, is the Cartesian thesis that some animal movements cannot be explained chemically but can be explained in terms of thoughts and wants. **(2)** The next strongest justification is the one yielded by my unity thesis, as I shall explain in a moment. **(3)** Finally, there is the fact that we often don't know the chemical explanation, which entitles us to use intentional explanations *faute de mieux*.

Evidently **(1)** is not available at the actual world, and it would be a sad day for belief and desire if **(3)** was the best we could do. So let us focus on **(2)**, which says that an intentional explanation of the given behavior brings out patterns, provides groupings and comparisons, that a chemical explanation would miss. What the animal did belongs to a class of behaviors in which it wants food and does what it thinks will provide food, and there is no unitary chemical explanation that covers just this range of data. This animal seeks food in many different ways, triggered by different sensory inputs, and it is not credible that a mechanistic, physiological view of the facts will reveal any unity in them that they don't share with behaviors that were not food-seeking at all. If this unifying view of the facts answers to our interests, gives us one kind of understanding of the animal, and facilitates predictions of a kind that are otherwise impossible (predictions like 'It will go after that rabbit somehow'), we have reason for adopting it. These reasons leave us free still to acknowledge that each of the explained facts, taken separately, admits of an explanation that is deeper and more wide-ranging and—other things being equal—preferable.[1]

2. Some objections answered

When I first said this, Davidson thought I had implied that a thing could lose its entitlement to intentional treatment because we discovered a single mechanism underlying all the input-output relations that we had hitherto grouped under some generalization about thoughts and wants.[2] That was a misunderstanding. The line around fully legitimate intentional explanations depends upon whether there is a single mechanism, not on whether we know it.

Peacocke has rejected my unity thesis because it implies 'that if we discover a creature that has only one way of catching flies, an intentional explanation of the creature's behavior is spurious'.[3] This does not address itself to the question of how the 'intentional stance' is to be justified with respect to a given animal, and presumably it is meant as a naked appeal to conceptual intuition. That is all right: if the appeal were a resounding enough success, that would be evidence that I have been talking about something which, however worthy and interesting, is not the conceptual underlay of our ordinary uses of 'think' and 'want' and 'intend' and 'in order to' and so on. Peacocke's appeal to intuition, however, has no such success. In fact, there are two different things it might be: one of them is not true, and the other does not conflict with my account. (i) If Peacocke's creature catches flies by a technique that involves one motor kind movement

[1] For more along this line, see Jonathan Bennett, *Linguistic Behaviour* (Cambridge University Press, 1976; Hackett Publishing Company: Indianapolis, 1989), sections 21–22; Daniel C. Dennett, *The Intentional Stance*, op. cit. chapter 2.

[2] Bennett, loc. cit.; Donald Davidson, 'Rational Animals', *Dialectica* 36 (1982), at p. 232.

[3] Christopher Peacocke, 'Demonstrative Thought and Psychological Explanation', *Synthese* (1981), pp. 187–217, at p. 212.

upon receipt of one sensory kind of stimulus, there is no strong intuitive support for the claim that this is cognitively guided goal-seeking behavior. I would think worse of my theory if it implied that such a creature brought thoughts and wants, or any analogue of them, to bear on its getting of food. (ii) If the behavior in question involves one kind of movement upon receipt of a wide variety of different sensory clues, that does look like cognitive teleology, but then it also conforms to the unity condition. I have tended to illustrate the condition by contrasting simple-input/simple-output with complex-input/complex-output, but I didn't have to. So long as the input side is complex in the right way, the behaviors in question can't be brought under a single non-intentional explanation; and that is all I demand.[1]

I have met the objection that if the legitimacy of the intentional stance depends on the unity condition then we ought never to have much confidence, of any organism, that it really does have thoughts and wants. 'Given how little we know about what in detail goes on in the central nervous systems of animals,' the challenge goes, 'how could we be entitled to think that a given range of behavior was probably not under the control of a single mechanism?' I think we could easily be entitled to think this. Our generalization implying that the animal does what it thinks will bring it food brings together a certain class of behaviors and a certain class of sensory inputs. Among the behaviors are cases of running, dodging, climbing, digging, swimming, leaping, biting, keeping still, keeping quiet, etc.—involving lots of different muscles and different uses of some of the same muscles. The sensory inputs include a variety of different kinds of sight, smell and sound. In the light of all this, we are soberly entitled to suppose that no one mechanism explains all this behavior.

3. Developing the unity thesis

We have a mechanistic generalization if all the relevant inputs are of some one *sensory* kind and all the relevant outputs are of one *motor* kind.[2] The emphasized adjectives are important. If in the relevant class of situations, x is confronted by evidence that something it could do would lead it to food, its inputs all belong to a single kind, namely the kind 'constituting evidence that something x can do would lead it to food'; but this is an *evidential* and not a *sensory* kind. What unites the inputs is something that involves the notion of seeming, or of evidence, or the like, and not something that could be stated just in the language of the intrinsic nature of inputs. Similarly, if in the relevant class of situations, x always moves in some way that is likely to get food, those movements belong to the kind 'being likely to lead to getting food'; but this is an *instrumental* and not a *motor* kind; that is, it is a kind defined in terms of probable upshot, not in terms of the intrinsic nature of movements.

(Whether a class of situations falls within single sensory kind depends not on how its members strike us but on how they strike the animal x whose behavior we are trying to explain. Even if it seems to us that the relevant class of inputs have in common only that in each of them there is evidence that some other animal is frightened, it might be that in all of those situations x detects a single characteristic kind of smell, in which case x's inputs in those situations belong

[1] As for simple-input/complex-output: that would involve an animal whose pursuits of a certain kind of goal were triggered by some relatively simple kind of stimulus, with no significant differences amongst the occasions on the input side, but were executed by a variety of different kinds of movements that have in common only their being apt to produce the goal. That would be magic, and is therefore negligible.

[2] Or a class of sensory kinds whose members differ only in different settings of some small number of parameters, and similarly with motor kinds.

to a single sensory kind. On the other hand, it presumably couldn't happen that we find only an instrumental kind of unity among the outputs although there is a motor kind from the standpoint of x.)

A class of situations covered by something of the form 'x receives input of sensory kind K_S and makes a movement of motor kind K_M' might also be covered by something of the form 'x receives evidence that it can do something that will lead to G and it does that something'. But the sensory-motor generalization prevails over the evidential-instrumental one. We are not fully entitled to employ the latter unless that is our only way of bringing the phenomena under a single generalization. So what is needed for a justified intentional explanation, abstractly stated, is a class of behavioral episodes whose inputs all answer to this description and to no 'lower' one:

> There is a kind K of movement such that: (1) x gets sensory evidence that if it performs a K movement it will get G, and (2) x performs a K movement.

From now on, to keep things simple, I shall focus on the (1) component of the analysis, leaving (2) to tag along unaided.

The unity thesis helps with a problem that is aired at some length in Dennett's first paper on cognitive ethology.[1] There is a tendency to think that any behavioral regularity is probably due to hard wiring ('tropism' or 'instinct' are Dennett's terms), or to a low-level acquired stimulus-response pattern. In Dennett's words:

> The oft-repeated, oft-observed, stereotypic behavior of a species. . . is just the sort of behavior that reveals no particular intelligence at all—all this behavior can be explained as the effects of some humdrum combination of 'instinct' or tropism and conditioned response. It is the novel bits of behavior, the acts that couldn't plausibly be accounted for in terms of prior conditioning or training or habit, that speak eloquently of intelligence. (Dennett, p. 348a)

But the alternative to oft-repeated kinds of behavior are the behavioral episodes that get reported in anecdotes, and we are assured that real science can't be based on those. This threatens to close down any gap through which we might conduct a scientific—or at least a respectably disciplined—study of cognition, especially high-level cognition.

Dennett's solution is to say that anecdotes may be all right if we have lots of them, as we do to support our opinions about one anothers' mental level:

> 'As we pile anecdote upon anecdote, apparent novelty upon apparent novelty, we build up for each acquaintance such a biography of *apparent* cleverness that the claim that it is all just lucky coincidence—or the result of hitherto undetected "training"—becomes the more extravagant hypothesis.' (Dennett, p. 348b-c).

But he does not discuss how piling up anecdotes differs from discovering a behavioral regularity, nor does he spell out what makes an anecdote evidence of 'apparent cleverness'. I shall make a suggestion about that shortly.

What Dennett calls (apparent) 'novelty' is what used to be called 'insight'. It is a real phenomenon, but in the initial 'insight' literature it was often implied to involve intellectual feats that owed nothing to the animal's past experience. If that were really the case, the feats would have to be (if not miraculous) hard-wired, mere tropisms having their first outing, and therefore not evidence of high-level intellect. A better way of viewing such 'novelties' is this: the animal solves a 'new' problem, or finds a 'new' solution for an old problem, by

[1] Daniel C. Dennett, 'Intentional Systems in Cognitive Ethology: the "Panglossian Paradigm" Defended', in *The Intentional Stance*, op. cit., pp. 237–268.

extrapolating or generalizing from its past experience *across an impressively large qualitative gap*. (It probably got across the gap with help from imaginary trial-and-error approaches to the problem, and that is impressive too.)[1] What impresses Dennett about it is the evidence it gives that the animal's successes in achieving its goals are not all products of habit, dumb training, low-level conditioning. That seems right, but I don't think it is quite central to the issue that Dennett and I are wrestling with. It is an approach to 'Is this dumb tropism or something higher?' that seems to offer no help at all with the question 'Is this a little higher than dumb tropism or a lot higher?' The account I shall give will give help with the second question as well as the first. The 'novelty' or 'insight' idea is not something I shall discuss, but I think it can be simply added to what I shall say.

Anecdotes are also made more admissible, Dennett says, if they report episodes that were controlled by the anecdotalist:

> 'Similar stratagems can be designed to test the various hypotheses about the beliefs and desires of vervet monkeys and other creatures. These stratagems have the virtue of provoking novel but interpretable behavior, of generating anecdotes under controlled (and hence scientifically admissible) conditions.' (Dennett, p. 348d)

I submit that control has nothing to do with it. When you know what you are looking for, control gives you a better chance of finding it; but that is a practical convenience, and can't help with the basic problem of how to get scientifically valid results from data that are not about regularities. A solution to that problem has to depend on what the results are, not on how they were arrived at, e.g. whether through a controlled experiment or just through passively observing an animal with which one was not interfering at all.

The right solution, I suggest is as follows. If we can't bring a given behavioral episode under a generalization about that animal's behavior, we can't confidently make *anything* of it—that it manifests thoughts and wants, or for that matter that it comes from instinct or low-level stimulus-response. So we need generalizations about the animal's behavior, which is to say that we need behavioral regularities; and the problem is to say *what marks off the regularities that are evidence of high-level cognition from those that are not*.

Here is what does it: If the generalization that we establish about the animal's inputs and outputs colligates the data under sensory kinds of input and motor kinds of output, it provides no evidence of cognitive mentality; but if it pulls the inputs together in evidential rather than sensory kinds, and if there is no 'lower' unity to the inputs, then the behavior in question is evidence that the animal behaves as it does because of beliefs and desires.

What Dennett calls 'piling up anecdotes' might be the accumulation of plenty of evidence for a generalisation about a class of sensorily diverse inputs and perhaps outputs that are diverse in their motor respects. Reports on such episodes might be called 'anecdotes' just because of their sensory and perhaps motor diversity. If they are Dennett's topic, then what he says is right, but his presentation is misleading. Once the content of the relevant generalisations is understood, we can see that there is really no tension or difficulty here at all, and we need not be pushed into giving weight to an unexamined notion of 'novelty' or a fundamentally irrelevant notion of 'control'.

[1] For an expanded version of these compressed remarks, see Jonathan Bennett, *Rationality* (Routledge and Kegan Paul: London, 1964; Hackett Publishing Company: Indianapolis, 1989), the final section ('Insight').

4. Further use for the unity thesis

It is often held by philosophers of mind that there are senses of 'higher' and 'lower' that make true something that Dennett has called *Lloyd Morgan's Canon*: 'If two hypotheses about an animal equally fit its behavior, and one attributes to it mental capacities that are higher than those attributed by the other, the latter hypothesis should be preferred.' If something like this is right, the unity thesis might be seen as the special case of it where the higher attribution involves some cognitive mentality and the lower involves none. In other special cases both competitors would attribute cognitive mentality, but one would attribute more of it, or more complexity or sophistication in it, or the like. This sloppy formulation is meant as a reminder that I haven't offered to define the higher/lower distinction, and so a fortiori I haven't put myself in a position to defend Morgan's Canon. Those are two nontrivial tasks which I cannot embark on here. In this paper I shall help myself to the assumption that Morgan's Canon is correct when interpreted in conformity with our intuitive sense of what counts as 'higher' than what.

If that is right, and if the unity thesis is a legitimate special case, my way of handling the unity thesis could help us to deal with other higher/lower issues. Consider the question: When the monkey gave its warning cry, did it want its companions to *believe there was a leopard nearby* or merely to *climb a tree*? I assume on intuitive grounds that the former is 'higher' than the latter: it credits the monkey with a thought about beliefs, whereas the other credits it merely with a thought about movements. According to my present hypothesis, we should adjudicate between the two by finding the 'lowest' evidential property that is possessed by all and only the environments in which the monkey utters that sort of cry. (If that class of environments is marked out by a sensory kind, that undercuts any evidential kind, and the explanation of the cries ought to be something right off the bottom of the intentionality scale.) The rival kinds of evidential property are these:

Low: The environment offers evidence to the calling monkey that that sort of cry will cause the other monkeys to climb trees;

High: The environment offers evidence to the calling monkey that that sort of cry will cause the other monkeys to believe there is a leopard nearby.

If we are to be entitled to think High is true of an environment, we must have grounds for attributing to a monkey a belief about the beliefs of other monkeys. What basis could we possibly have for this? Well, the functionalism that explicates our opinions about what monkeys believe must be supposed also to explicate *their* opinions (if they have any) about what other monkeys believe. That is, if they have a concept of belief, it like ours must be supported by the belief-desire-behavior triangle. Fortunately, for my present purposes I can take a somewhat simplified version of this idea. I shall say that an environment satisfies High if:

High*: The environment offers evidence to the calling monkey that that sort of cry will cause the other monkeys to act in a manner appropriate to the information that there is a leopard nearby.

Of course any environment that satisfies Low also satisfies High*. But we can't be entitled to associate the warning cries with a desire to produce the belief that there is a leopard nearby unless they occur in a class of environments that is united under High* but not under Low. Nor under anything else that is lower than High*. For example, if the cry is sometimes given when all the monkeys within earshot are visibly in trees already, the entire class of relevant environments may be united by this property:

> The environment presents evidence to the calling monkey that that sort of cry will cause the other monkeys to be in a tree, i.e. to go into a tree or if already in a tree to stay there.

That is different from Low, but it is lower than High* and therefore disqualifies the latter.

So, what is needed for us to be fully entitled to read the calls as intended to produce beliefs rather than to produce behavior is that they occur in a class of environments which is a vastly complex jumble unless we bring it under the unifying concept of 'environment in which it seems to the monkey that a warning call will lead the others to act in a manner appropriate to there being a leopard nearby' or else under some concept that is even higher than that—for example, '. . . environment in which it seems to the monkey that a warning call will lead the others to act in a manner appropriate to the caller's believing that there is a leopard nearby', or '. . . appropriate to the caller's wanting the hearers to believe that there is a leopard nearby', and so forth.

I offer that as an example of how the structure of my applications of the unity thesis might be used also higher up the ladder, to help bring discipline into questions about which of two competing intentional explanations should be adopted. By these standards it is unlikely that we shall ever be entitled to think that any nonhuman animal has tried to get another to believe something; but I don't say that in criticism of the standards.

5. Descartes on complexity

A consequence of the unity thesis is that an animal can have a goal and the intellectual ability to recognize means to it only by virtue of having packed into it large number of mechanisms. Descartes said that a physical replica of a man would not behave in every way like a man, and he gave two reasons for this. Here is one of them:

> 'Even though [such physical replicas] might do some things as well as we do them, or perhaps even better, they would inevitably fail in others, which would reveal that they were acting not through understanding but only from the disposition of their organs. For whereas reason is a universal instrument which can be used in all kinds of situations, these organs need some particular disposition for each particular action; hence it is morally [*moralement*] impossible for a machine to have enough different organs to make it act in all the contingencies of life in the way that our reason makes us act.'[1]

We must agree with Descartes that a purely physically controlled system would need a distinct physical mechanism to ensure obedience to each distinct conditional of the form 'In an E environment perform an A action', and that our reason puts us in command of countless such conditionals.[2] But if we assume (as I do and Descartes didn't) that the doings of reason are supervenient on physical happenings, we must conclude that reason generates all those conditionals because its activities are supervenient on those of a vast stock of distinct mechanisms taking us causally from initial states to resultant states, including taking us from sensory inputs

[1] René Descartes, *Discourse on the Method* 5, AT 6.56f.

[2] If two conditionals differ only in having different settings of two or more parameters, they could be kept true by a single mechanism that had re-settable parameters in it. So when I speak of how many distinct conditionals our reason makes true, I mean how many conditionals that differ from one another in more ways than that.

to behavioral outputs. This is not in any way impossible, and thus is not 'morally impossible', whatever Descartes meant by that. He was probably helped to think otherwise by having no idea of how small the working elements of a brain are. He may have been affected also by the assumption that each distinct conditional requires a distinct 'organ', i.e. a physical arrangement that has no physical overlap with any arrangement governing some other conditional. That assumption is false, of course; there is no reason why two mechanisms should not share most of their matter.

Anyway, we don't have to be materialists to think that a universal instrument must be a compendium of particular instruments. Descartes' thinking otherwise is a sign of his tendency to assume—in Wittgenstein's great phrase—that the mind is 'a queer kind of medium' in which things happen that could not possibly happen anywhere else.

So we have to view a thinking, wanting, planning, goal-pursuing being as a tight cluster of a large number of mechanisms whose over-all effect is to make it register evidence about things it can do that will produce some state of affairs and then do those things.

If I have seemed to imply that for an animal to house a mechanism is for some input-output conditional to be durably true of it, I retract that. Most of the relevant conditionals about actual animals are switched on or off according to the animal's state of alertness, sexual satiety, blood-sugar level, and so on. I leave these toggles out of my account for simplicity's sake; in my main line of argument the omission is harmless.

6. Some further aspects of intentionality

I have been contending that we are not entitled to apply intentional concepts to a system unless (i) its input/output relations fall into a certain kind of pattern and (ii) they satisfy the unity condition. In my next section I shall introduce a further necessary condition for intentionality or cognitive teleology—one that will occupy the rest of this paper. That, however, will not be an attempt to strengthen my account of what is needed for intentionality so as to turn it into an account of what suffices for it. Other required elements will certainly be missing.

For example, our concepts of belief and desire are probably such as to require that the inner routes from input to output satisfy certain constraints. Searle's 'Chinese room' thought experiment seems to indicate that there are such constraints, though it gives us only negative information—i.e. tells us almost nothing—about what they are. One possibility is that a system counts as thinking and wanting only if the following is true:

> If two input/output pairs contribute to a single teleological pattern, that increases the probability that there is some physical overlap between the inner routes that they involve.

Other ideas also suggest themselves. If I were pursuing sufficient conditions (that is, pursuing all the necessary conditions) for intentionality, I would have to dig into this topic, but I am not, so I shan't.

Again, all actual intellect involves cognitive dynamics: often enough a given item of sensory input has no immediate effect on behavior but makes a difference to the behavioral upshots of later inputs by affecting the animal's 'cognitive maps'. Block's vending machine does model that much, because giving a penniless machine a nickel doesn't make it do anything, but changes its cognitive map so as to alter what it does when the next nickel is fed to it. Now, perhaps this is conceptually required. Perhaps it is the case that if it were clear to us that a given system was not subject to such cognitive dynamics, that would automatically satisfy us that

it wasn't a genuine thinker and wanter. If so, then that is a further necessary condition that I am ignoring.[1]

Well, so be it. I want to tell one part of the story properly, and am content in this paper to leave other parts untold.

As for the phenomena that I am setting aside—the ones that are naturally described in terms of cognitive dynamics—could they be described in terms of my apparatus of input/output conditionals? That is, if I wanted to enrich my account to take them in, could I do it by moving on from where I am, or would a fresh start be needed? I think a fresh start would be needed. To force the input/output conditionals to cover the phenomena in question, I should have to make them astronomically complicated and astronomically numerous. Indeed, the case for hypothesizing cognitive states that are affected by sensory inputs and that also combine with sensory inputs to produce behavior is just that without that hypothesis we have a horrendous clutter of input/output conditionals.

Still, the story I am telling in terms of such conditionals is a legitimate abstraction from the thicker story. I claim to have made some good use of it, and I now proceed to try to make more. This brings me to where I was at the end of Section 5.

7. Intentionality as a source of explanations

In the account I have been giving, nothing rules out its being a mere coincidence that this single system houses a lot of mechanisms whose over-all effect is to make the system a G-seeker; and if it is a coincidence, the system's intentionality cannot be used to explain its behavior. Here is an analogous case. Suppose that of the cities Joe is acquainted with he hates all and only those whose city government has a ward system; there are about forty of them, and Joe's emotions about them have forty different reasons, their common political systems being a sheer coincidence. That gives us a generalization on the strength of which we can 'unite' Joe's hatred for Detroit with his hatred for Chicago, and so on, but it doesn't give us the faintest *explanation* for any of the hatreds or, therefore, any reason to expect that he will hate the next such city that he encounters.

It does give us an explanation for his hatred for Detroit *today*—namely that he has a deep-seated and longstanding hatred for Detroit; what it doesn't do is to give us any carry-over from one city to another. Similarly, the account I have given of intentionality up to here may enable us to explain the animal's going on this occasion from a stimulus of kind S to a movement of kind M: it has done this often enough to convince us that it has some settled disposition to link this kind of input with that kind of output. But that link between a sensory kind of input and a motor kind of output corresponds to a single mechanism; an explanation that exploits it is, precisely, an explanation that does not make use of any intentional concepts.

I have argued elsewhere that the concepts of belief and desire are nothing if they are not explanatory, and in this paper I shall take that for granted.[2] I shall also work on the assumption that we have something explanatory if we have something that would have licensed a prediction, but not otherwise.

To put intentionality to work, then, we need to be able to explain or predict one link between sensory input and motor output on the basis of links between other pairs—ones in

[1] The importance of this omission is one of many things that were made clear to me by Sydney Shoemaker's acute, searching, constructive comments on an earlier version of this paper.

[2] See Bennett, *Linguistic Behaviour*, op. cit., pp. 42–44.

which the sensory kinds are different (and perhaps the motor kinds as well). If an animal goes after rabbits in a variety of different ways, on the basis of a variety of different sensory kinds of clue, that gives us *some* reason to predict that it will go after rabbits on the basis of kinds of clues that we haven't so far observed it to use; but the account I have given so far doesn't lay any basis for this. That is because it doesn't rule out its being a coincidence that the relevant cluster of mechanisms all exists under a single skin. How, then, can we repair that hole in the account?

(I am not insisting that attributions of beliefs and desires be *causally* explanatory. I don't care whether the kind of explanatoriness that I shall find for folk psychological statements is causal in nature, and indeed I doubt if the question is determinate enough to be worth addressing. Even further off my path is the question of whether beliefs and desires are causes. This question requires us to reify or eventify beliefs and desires, i.e. to find not only truth conditions for 'x thinks that P' and 'x wants it to be the case that Q' but also application conditions for the noun phrases 'belief that P' and 'desire that Q'. It is better to ask whether attributions of beliefs and desires are causally explanatory than to ask whether beliefs and desires are causes;[1] but it is better still to keep causation right out of the picture.)

Suppose there is a single common cause for all the input-output connections that add up to the animal's having a teleological pattern of behavior. Would that provide us with teleological explanations of the behavior? It would do so only if it entitled us, having seen some parts of the pattern, to predict others; and clearly it would not do the latter. If in some astronomically improbable way a single large genetic mutation led to offspring that had a lot of G-getting mechanisms, where the parents had had none, this common cause would not make it legitimate to explain anything the offspring did in terms having G as a goal. The observation of behavioral upshots of some of the mechanisms wouldn't provide valid evidence for the existence of any others of them. Or, to revert to a parallel that I used earlier: we aren't helped to explain or predict Joe's hatred for Detroit through his hatred for Chicago just because both hatreds were caused by a single bad dream.

What we need for explanatoriness is that there should be a unitary causal explanation not merely for

> the system's having mechanisms $M_1,\ldots,M_k$,

where in fact its possession of those mechanisms make it a G-seeker, but for

> the system's having a lot of mechanisms that make it a G-seeker.

This is a weaker explanandum in one way, because it doesn't list the mechanisms. But I am more interested in the respect in which it is stronger, namely its including the fact that the mechanisms make the system a G-seeker.

8. One source of explanatoriness: evolution

One way of filling the gap in the account is through an appeal to evolution, and for my purposes a simplified pop evolutionary story is good enough. Of all the potential mechanisms that got a genetic fingerhold on the animal's ancestors through random mutations, relatively few survived; among the survivors were the bunch of mechanisms that make their owner a G-getter, and *that is why they survived*. Why does this animal contain a lot of mechanisms that make it a G-getter? It inherited those mechanisms from a gene pool that contained them *because they are mechanisms that*

[1] I here rely on how thing- and event-causation differ from what I call fact-causation on the other. See Jonathan Bennett, *Events and their Names* (Hackett Publishing Co.: Indianapolis, 1988), Section 8.

make their owner a G-getter.

That answers to my specifications for something that makes it more than a coincidence that the animal has many mechanisms that are united in their G-getting tendency. And it lays a clear basis for explanations that bring in intentionality. That a species has evolved a G-getting tendency that is manifested in this, that and the other links between sensory kinds of input and motor kinds of output creates some presumption that it has evolved other links that also have a G-getting tendency. So there is something predictive in this, and thus something explanatory as well.

If there had been no evolution but animals had been produced by a designing designer, the foregoing account would still hold, *mutatis mutandis*, just so long as the designer had included all the G-getting mechanisms in order that the animal should be a G-getter. As has often been pointed out, there is a strong analogy between the workings of evolution and the workings of a person executing a design, and the analogy goes far enough to spread across my present topic.

I haven't yet said that without an evolutionary explanation or something sufficiently like it (e.g. a designing designer) we couldn't use attributions of intentionality to explain or predict. But even if I did, that claim should be sharply distinguished from Dennett's thesis that it is only because we can appeal to what he metaphorically calls 'the intentions of Mother Nature' that we are in a position to make fairly determinate statements about the thoughts and wants of animals.[1] My account does not imply that we need help from evolution in order to answer the question: 'What, if anything, does this animal think and want?' The force of 'if anything' is that it might be a coincidence that this part of the physical world has packed into a bunch of mechanisms that give it intentional patterns of behavior; so that even when we have established the whole intentional story, we should hesitate to *tell* it, to *explain* anything in terms of it, unless we are sure that it is no coincidence and that the mechanisms are interconnected in the right way. This is not Dennett's claim that without an appeal to evolution we can't establish the story in the first place.

9. A second source of explanatoriness: educability

Now, consider an animal whose behavior falls under intentional concepts in a very nontrivial way—the generalization about the circumstances under which it seeks G as a goal covers a vast number of different mechanisms—but it doesn't contain the means for modifying any of this apparatus in the light of its experience. It picks up from its environments all kinds of information about ways to get G, and acts accordingly, but if one of these input-output pairs starts to let it down, leading not to G but to something unpleasant, that does not lead the animal to delete that input-output pair from its repertoire. Nor does it ever add anything to its repertoire in the light of chance discoveries about what works.

I'll bet that there are no such animals. It is vastly improbable that the required kind and degree of complexity should evolve without being helped along by the evolution of a degree of individual adaptability to discovered changes in circumstances. Still, it could happen. The idea is not incoherent or absolutely impossible; we know what it would be like for there to be such behaviorally frozen animals. They would cope successfully and (it would seem) intelligently with their environments, but as soon as these altered a bit in some relevant way, the animals would be incurably in difficulties, and after a modest number of such alterations

[1] Dennett, *The Intentional Stance*, op./ cit., chapter 8.

the animals would be dead.

We can imagine a world at which great behavioral complexity did have great survival value whereas individual adaptability didn't. At such a world, frozen complexity might well evolve, and my demands for intentional explicability would be met. Animals at that world would have richly intentional patterns of behavior—hard-wired instincts generating a multiplicity of fine-grained minutely appropriate ways of behaving whose over-all effect was to make the animal a G-seeker for this or that value of G.

The behavior of such creatures could be explained and predicted intentionally. If an animal has a lot of (for short) G-seeking input-output patterns, that is evidence that they have been selected because they let the animal get G; and that is evidence that other input-output links that have the same upshot will also have been selected. By the prediction test, therefore, we can use the premise that the animal is a G-seeker to explain a new bit of G-seeking by it; the premise is at least somewhat projectible, and is not a mere summation of observed behavioral episodes. And all this applies *mutatis mutandis* to frozen creatures that resulted not from evolution but from the activities of a designer.

So we can have explanatory intentionality even where there is no educability, just so long as the animal's origin makes it more than a coincidence that it houses a lot of mechanisms whose over-all effect is to make it a G-getter. What about the converse? That is, what about educability without evolution or any substitute for it?

Well, consider again the case of educable parents that have an educable offspring with a goal that they didn't have: the offspring is the locus of a large number of G-getting mechanisms, none of which were present in the parents, their presence in the offspring being the result of a very radical and sheerly coincidental set of genetic mutations. It is to be understood that the offspring's inherited educability extends to its pursuits of the goal G. (I suppose the educability to be inherited so as not to make the story more biologically bizarre than I need to for my purpose.)

This story, though utterly improbable, seems to be coherent and to state a real possibility. If we knew that it was true of a given animal, we could *explain* some of the animal's behavior in terms of its having G as a goal. For

(i) its having G as a goal and

(ii) its being able to learn from experience

jointly give us reason to predict that it will pursue G in ways (and on clues) that we have not previously seen it employ. Such a prediction presupposes that it is on the cards that the animal has previously employed those ways and clues or ones from which it has been able to reach those though some kind of generalization, imagined trial-and-error, 'insight', or the like (cf. Section 3 above). That presupposition distinguishes this from the evolutionary case. In the latter, we have some grounds for predicting that the animal will pursue G through a certain input-output pair without knowing anything about its past experience; but of course we have to assume that many of its forebears have experienced that pair, for otherwise the trait linking them could not have been selected. This difference between the two is, on reflection, just what one would expect. What evolutionary adaptability is to a species, educability is to an individual; so explanations in terms of the former are likely to say things about the species that will be said about the individual in explanations in terms of the latter.

11. Appeals to intuition

I have not been inviting you to consider various possible kinds of animal and to judge whether you would be willing to describe such an animal in terms of beliefs and desires. I

have not been holding up examples of an educable animal that did not evolve, an evolved animal that is not educable, one that has both features, and one that has neither, and asking you 'Does this strike you, intuitively, as an animal that thinks and wants?' Out at the margins where we are, such appeals to conceptual intuition are not worth much. I have not engaged in them, and do not need them.

My strategy has been different. I argue for the unity thesis, according to which the range of a folk psychological generalization concerning a particular animal should correspond to a lot of different generalizations relating sensory kinds of input to motor kinds of output. I add to this the premise, argued for elsewhere, that the concepts of belief and desire are legitimate only if they can help to *explain* behavior. That raises the question of how a folk psychological generalization can be genuinely explanatory, by the acid test according to which what can explain could have supported a prediction. To that I have given the best answers I can find—answers that mercifully spare us from the quicksand question of whether beliefs and desires can be causes.

It happens that those answers, developed in order to satisfy a certain theoretical demand, do also serve to bring the account closer to what intuition demands. In *Linguistic Behaviour* I left educability and evolvedness out of my account of basic teleology. I rightly said that evolution gave the best answer to questions of the form 'Why does this animal have that goal?', that is, 'Why is a set of mechanisms with that over-all tendency packed under one skin?' But I treated this merely as a question that might arise, not as something that is needed if teleological explanations are to be given for the behavior of a not very educable animal.

I brought in educability as helping to mark a certain difference of level: I thought that some genuinely cognitive teleology ought not to be described in terms of 'believes' and 'wants' or 'intends' but only in terms of more generic notions which I expressed as 'registers' and 'has as a goal', and I offered educability as part of what makes the difference. I was steering here by conceptual intuitions, and I think I steered a true course. But I did not realize that educability was also playing a stand-by structural role: in the absence of evolution (or divine design), educability would be needed for any concepts of cognitive teleology, even low-level ones, to be applicable.

In the book I offered an example of a lake whose behavior has a preserving-the-local-wildlife pattern, and I said that its apparent teleology is fake because the very same behavior also falls into a simple mechanistic input-output pattern. I implied (because I believed) that that failure to satisfy the unity condition was the sole obstacle to attributing goals to the lake; protests from readers made it clear that this was not intuitively acceptable; and I am now clear that at least part of the short-fall was due to the fact that cognitive teleology, even of an abysmally low-level kind, requires not only the unity condition but also something that makes the teleological generalizations genuinely explanatory. That is the gap I have been trying to fill in the last part of this paper.