REPRESENTATION AND EXPLANATION*

DAVID PAPINEAU

Department of History and Philosophy of Science
University of Cambridge

Functionalism faces a problem in accounting for the semantic powers of beliefs and other mental states. Simple causal considerations will not solve this problem, nor will any appeal to the social utility of semantic interpretations. The correct analysis of semantic representation is a teleological one, in terms of the biological purposes of mental states: whereas functionalism focuses, so to speak, only on the *structure* of the cognitive mechanism, the semantic perspective requires in addition that we consider the *purposes* of the cognitive mechanism's parts.

It seems undeniable that beliefs and other propositional attitudes *represent*. But the most plausible theories of such mental states, namely the various forms of functionalism, face a difficulty in accounting for the representational powers of propositional attitudes. In this essay I shall explore how far functionalism and related approaches in the philosophy of mind can deal with this difficulty. In the course of this I hope to cast light both on the nature of representation and on the viability of the functionalist program.

I

I shall take functionalism to be the view that being in a given mental state is a matter of being in some physical state which is causally related in a certain way to certain other physical states, and in particular to perceptual inputs and behavioral outputs, the way in question being specified by the appropriate psychological theory.

This is an overly crude characterization in (at least) three dimensions. There are questions to be asked: (a) about what qualifies as "the appropriate psychological theory," (b) about the scope of the "some" in "some physical state," and (c) about how far the causal net has to extend before we reach "perceptual inputs" and "behavioral outputs." However, nothing in my argument will depend on how these questions are answered. (In particular, as will become clear in Section III below, problems about representation are *not* removed simply by extending the causal net to in-

*Received October 1983; revised March 1984.

Philosophy of Science, 51 (1984) pp. 550–572.

clude the "external" causes of perceptions and the "external" results of actions.)

The central point about functionalism, for our purposes, is that it takes mental states to be explanatory entities which mediate between inputs to the mind and outputs from it, and that it takes such explanation to involve a theory of the internal causal workings of the mind. True, functionalism abstracts from the physical substance of such causal workings, in the style of a high-level description of a computer program, or, again, of a Ramsified version of a scientific theory; but, for all that, at bottom it pictures mental states as elements in a system of causal pushes and pulls inside the head.

The problem I want to attend to is that on this account it is not at all clear what call there is to think of mental states as *representing* the world. Given the idea of propositional attitudes as components in a system of "horizontal" relationships mediating causally between inputs and outputs, why go on to postulate "vertical" relationships between those attitudes and the things we intuitively think of them as standing for (Loar 1981, p. 57)? If the functionalist characterization suffices, as it is designed to, to serve any explanatory purposes we may have with respect to the formation of mental states and the execution of subsequent behavior, then what is added by an *interpretation* which relates those attitudes to "external" states of affairs? Why, as it is sometimes rather confusingly put, should we think of propositional attitudes as having a semantics in addition to their syntax? Or, to put it more bluntly, why should we think of them as being *about* anything?

I am inclined to argue that this difficulty arises not just for functionalism but for any physicalist (or, for that matter, dualist) position that takes mental states seriously as internal states with causal powers. But functionalism has the merit of bringing the problem out into the open. For it forces us to recognize that there are *two aspects* to our notion of belief: beliefs as fillers of causal roles, and beliefs as things which represent other things (McGinn 1982). And by concentrating exclusively on the first aspect, it makes us realize that there is no immediate reason why something with a causal role should have representational powers.

II

Some commentators would argue that the emergence of this puzzle is simply an index of the wrong-headedness of functionalism: "Of course," they will say, "if you start by assuming that psychological states explain in virtue of their 'causal roles', then you will have a difficulty finding a place for representation. But what reason do we have in the first place for supposing that the explanatory import of propositional attitudes can be detached from their representational powers? Functionalism is simply wishful thinking, in that it credits us with a sophisticated general theory we do not have. And it is misguided wishful thinking at that, in that the desire for such a theory stems from a misunderstanding of the structure of psychological explanation."

I shall not deal with this kind of complaint directly. Instead I shall proceed by taking functionalism at face value, and showing how it can solve in its own terms the problem of representation it lands itself with. Once this has been done the right response to general suspicions of functionalism will be clearer. I shall return to the general issues in the last two sections of the paper.

But before I do proceed there is a more immediate worry about functionalism to be dealt with. This worry relates to the way in which we identify beliefs (I'll stick to beliefs for the time being and bring desires back in later), as, the belief *that Paris is the capital of France,* the belief *that p,* the belief *that q,* etc. It certainly seems at first sight that such content clauses work by mentioning what we normally think of as the objects of the belief in question. And if this is our way of identifying beliefs, does it not then follow that our primary notion of belief is a notion of something-which-represents-something-else, rather than of something-with-a-certain-causal-role?

But the functionalist has plenty of room here to resist the thought that we only, so to speak, get at beliefs *by way of* a grasp of what they're about. For he can view the content clauses simply as *labels,* or indices, for the functional roles of the beliefs in question (Dennett 1978, pp. 26–27).

The picture the functionalist can offer is this. On the one hand we have a set of possible functional roles. On the other we have a system for labelling them ("Functional role$_{743}$," "functional role$_{82}$"). The identification of beliefs can then proceed directly by identifying their functional roles. We need not think of ourselves as *first* identifying a possible state of affairs, and *then* identifying a belief as the belief about that state of affairs. From this functionalist point of view it is then almost incidental that we use the English sentences we do to label beliefs—any isomorphic labelling system (the negations of those sentences, or, indeed, their Gödel numerals) could in principle be used to the same effect.

Still, one wants to say, *we* do use English sentences to identify beliefs—how exactly is this supposed to work as a labelling system? This is, to say the least, a matter of some delicacy for the functionalist, and I don't want to get bogged down in it. But one brief remark is worth making. Even if the functionalist admits (as I think in the end he ought to) that content clauses somehow mention those things that we intuitively think of as the objects of the relevant beliefs, this does not remove the need to explain "aboutness," to explain why we are *right* so intuitively to think of those beliefs as about those objects. For all the functionalist has admitted is that certain objects are involved in our system for labelling the causal roles of beliefs. And that in itself does nothing at all to show why we should think of those beliefs as *about* those objects, as *representing* states of affairs involving them. (Compare the way in which the Gödel *numbers* are involved in a system for labelling the wffs of PM. This doesn't immediately mean that all those wffs are *about* those numbers. It's a good trick to find a particular wff that *is* about its own Gödel number.)

III

The puzzle which I am laboring has not been as widely recognized as one might expect. No doubt one reason is that it is fairly easy to suppose that representation is simply a causal/functional matter: what a belief stands for is the kind of circumstance that causally produces it.

The trouble with this is that it lets far too much in. Given any belief, there are circumstances which can reasonably be counted as causes of that belief, but which we would not think of as being represented by that belief, as forming part of the truth condition for that belief.

This objection arises in two independent dimensions. Firstly, consider the chain of causes leading up to a given belief: the internal structure of some physical object, its manifest features, electromagnetic radiation, retinal stimulation, optic nerve activity, and so forth. At which point in this chain are we to locate the truth condition for this belief? At the beginning? At the end? Throughout?

And then, secondly, even if this problem could be solved, there is the point that beliefs can be produced by inappropriate causal chains, which don't contain their truth conditions anywhere. People can be misinformed, led astray by misleading association, etc. And this goes for observational beliefs as well as indirectly inferred ones. People will believe there is a tree in front of them not only when there is a real tree there, but also when faced with a good tree replica. Yet we still want to count the belief as about *trees* and not tree replicas.

One might be inclined to appeal here to some such notion as the "typical" or "characteristic" causes of the belief. But the problem is that we have as yet given no independent substance to the distinction between "typical" and "untypical." In particular it is not part of our psychological thinking (and therefore, presumably, not part of the theory the functionalist "Ramsifies") that the unwanted cases are some kind of random aberration in our cognitive workings: the man who sees a tree replica is in perfectly good working order, and what happens to him is perfectly well-understood by our psychological theory. We can certainly explain his belief just as well when it is produced by a tree replica as when it is produced by a tree.

IV

Some functionalist writers have attended explicitly to the problem of representation. And one suggestion is that the reason we should see sentences (and beliefs) as correlated with truth conditions is that this enables us to use people's utterances as reliable indicators of further facts (Field 1978; also Loar 1981; McGinn 1982). We know, given the way English speakers operate, that it is in general a sensible practice when somebody utters the words "It is raining" to infer that it is raining—and similarly with all other English sentences and the states of the world we take them to stand for.

On this conception viewing words and beliefs as about things is a matter of *calibrating* people as instruments for detecting states of the world in general, as one might calibrate a thermometer as an instrument specifically for detecting temperatures. The points made in the last section show that if all we were after from our instruments was *reliability,* then the standard calibration would be less than optimal—people who say "There's a tree" are less reliable when taken as indicators of trees than they would be if taken as indicators of trees-or-tree-replicas. But one can see how if one wanted to maximize not just reliability, but some mixture of reliability with *informativeness,* then one might indeed be led towards something like the standard pairing of beliefs with truth conditions.

The difficulty with the calibrational suggestion seems to me to be the following. The idea that beliefs are about things is supposed to be justified instrumentally, in terms of its enabling us to infer conclusions about the world. But inferring conclusions is itself a matter of forming beliefs. Why is it to our instrumental advantage to form beliefs in the way allowed by the standard calibration of others' utterances? Well, since those other people are generally reliable in the required way, the beliefs we infer from their utterances will generally be "true," in that the states of affairs the calibrational interpretation pairs with those beliefs will generally obtain when we hold them. But why should we want our beliefs so to be "true"? Well, any more people who now infer beliefs from our expressions of our inferred beliefs won't be misled. But this is just to say, as above, that *their* beliefs will be "true," and simply takes us once more round a circle which fails to explain why truth in the sense so far explained is a good thing.

Clearly we want to be able to say at least this much: true beliefs are a good thing because actions based on true beliefs will succeed. (At its simplest, if you want result *r* and believe action *a* will bring about *r*, and therefore do *a*, you will get *r* if your belief is true.) But without further elaboration the notion of truth yielded by the calibrational account gives us no hold on why this should be so.

V

Suppose then we simply start afresh with the idea that the truth conditions for a given belief are circumstances the obtaining of which will guarantee that actions based on that belief will succeed. This then deals directly with the connection between the truth of belief and the success of action. Unfortunately this suggestion, as it stands, radically underdetermines the fixing of truth conditions.

The trouble lies with the notion of success. The success of an action does not consist merely in its having a certain physically specifiable effect. In addition that action must fulfill the desire to which the action was oriented. But what is it for an effect to fulfill a desire? Clearly the notion of fulfillment conditions for desires is closely analogous to that of truth conditions for beliefs, and raises just the same puzzles. Why ascribe fulfillment conditions to desires? Does not the specification of a desire's functional role not give us all we need? These questions are no easier to answer than they were for truth conditions and beliefs. (One surprisingly popular response, analogous to the "typical cause" suggestion for belief discussed in Section III above, is that fulfillment conditions are those which generally extinguish the relevant desire. This seems most unlikely to pick out the right states of affairs: unfulfilled desires can fade away, while others can be fuelled by their own satisfaction.)

It is worth spelling out how the constraint that truth should guarantee success underdetermines the fixing of truth conditions. Suppose for instance that we ascribed the fulfillment condition *light* to the desire we currently label as (that indeed is) the desire for warmth. (What is there stated in the constraint to stop us doing this?) Then we will get "truth conditions" which guarantee the "success" of the relevant actions (actions which we currently think of as aimed at warmth) by similarly substituting light in the truth conditions of those beliefs we normally specify by reference to warmth.

VI

One is inclined at this point to return to the calibrational suggestion. For, while the above example succeeds in showing that success isn't enough, it is obvious that the relevant "beliefs" will not be very *reliable* under the suggested "interpretation." Even though it is sometimes light in the relevant place when people have the belief that we currently think of as about warmth, that belief is a much *better* indicator of warmth than of light. (People are indeed sometimes led to the belief in question by the presence of light—but this is only an occasional and secondary source for the belief.)

What we seem to want, then, is some analysis which will combine the virtues of the calibrational and success-guaranteeing suggestions. I think this can be achieved if we adopt an explicitly biological perspective on our problem. For there *is* something that stops it from being an arbitrary matter whether we deem the belief and desire in the previous section to be focused on warmth rather than light. Namely, that it is in some sense the biological *purpose* of the relevant desire to lead us towards heat, and *not* towards light; and correspondingly, it is the purpose of the relevant belief to get us to act in ways appropriate to the presence of heat, and not light.

It might be thought that the notion of biological purpose is already built into the functionalist attitude to mental states. If we view mental states "functionally," aren't we automatically seeing them in terms of their biological purposes, or "functions"? However, nothing I have said up to this point justifies this, nor is there anything in the classic statements of functionalism, such as Putnam (1967), or Lewis (1972), to warrant it. Identifying mental states in a second-order way, as states with a certain structure of causes and effects, is not yet to view them as things with purposes, as things which can be *explained* in terms of certain of their effects. If there is a connection between "functionalism" about mental states and biological "functions," it needs to be argued for. It is in this spirit that I shall be arguing that functionalists need to attend to the biological purposes of mental states in order to account for the representational powers of those states. (For a different connection between the two meanings of "function," relating to the variable realizability of mental states, see Papineau 1984). To avoid confusion in what follows I shall use "teleological" instead of "functional" whenever I am talking about biological purposes.

One further clarificatory point before proceeding. The analysis of the notion of biological purpose is itself a matter of some controversy. I favor the view that talk of biological purposes needs underpinning by reference to explanations in terms of natural selection (see Wright 1973). Accordingly in the next three sections I shall assume that the biological purposes of mental states depend on how those states are to be explained in natural selection terms. Concessions for those who would prefer to detach teleology from natural selection are offered in Section X.

VII

Let us once more focus on belief. As before we shall return to desires in due course. Consider how we explain beliefs. In the first instance we explain particular beliefs (token beliefs) by reference to the particular circumstances that cause them. But, as we saw in III above, this casts the net too wide to capture the notion of representation—more circumstances cause tokens of a given belief than we want to think of as forming part of the belief's truth condition. But suppose we step back a level and ask for an explanation of our general disposition to form tokens of that type—an explanation of our having that belief-type in our repertoire in the first place. At this level we *do* get a distinction among the different possible causes of the belief. The biological explanation of our disposition to form the belief that it is warm somewhere is that (a) that belief has typically arisen on occasions when it is warm in the relevant place, and (b) on just those occasions the actions which issue from that belief have had selectively advantageous results. This explanation does not work if we replace "warmth" by "light." Although the belief has often arisen on occasions when it was light in the relevant place, the presence of light does *not* ensure that the consequent actions will be advantageous. Or, to take another example, consider beliefs about trees: although such beliefs actually arise more often in the presence of trees-or-tree-replicas than in the presence of trees, the actions they direct are advantageous in the presence of *trees,* and not, as a rule, in the presence of tree-replicas.

I want to consider, then, the following general analysis of truth conditions. The disposition to form a given type of belief is explained by the fact that that belief has typically arisen in certain circumstances, and in those circumstances the actions that it has directed have been selectively advantageous. The typical circumstances in question are the belief's truth conditions. And talk of "typicality" here is no longer empty, for it is precisely because of what happened on those "typical" occasions in the past (warmth present), and not on others (light), that we have our current belief-forming dispositions.

It is worth noting immediately that this suggestion does not require that all belief-forming abilities are all innate, or "hard-wired." It works just as well with acquired dispositions to form beliefs ("new concepts") as with hard-wired ones. Natural selection takes place in learning as well as in intergenerational evolution (though then it is natural selection of cognitive components rather than genes). New concepts are "fixed" by learning precisely when the beliefs they give rise to are advantageous; or when, to be precise, the circumstances in which those beliefs typically occur are ones in which the actions directed by those beliefs have advantageous effects. And here again we can say that those "typical" circumstances are the truth conditions of the beliefs: the ability to form those beliefs is explained by the fact that they "work" in those circumstances. (An interesting special case is the acquiring of concepts as the result of specifically linguistic training. Here the relevant actions will be utterances; the relevant advantages will be parental or other social rewards; and the concepts that get fixed will be those that will enable the child to recognize the circumstances that the utterances in question represent.)

Another point worth mentioning at this stage is that a natural selection account of representation does not mean that beliefs are always, or even usually, true. Admittedly, for the simple observational beliefs I have so far discussed, there will always be the "typical" circumstances relating to the explanation of that belief-type, and when a token of that type occurs in those "typical" circumstances, it is *ipso facto* true. But, as the tree-replica type of example shows, beliefs can arise in the absence of their typical causes, and in such cases they are then false.

Nor does it even follow from the natural selection story that a given belief will be true more often than not. Perhaps this is less obvious. According to the natural selection story it is the fact that a belief-type "typically" obtains in certain circumstances that will explain our having it in our repertoire. But if such occasions of true belief are to account for the belief-type's being selected for, won't they need to be far more frequent than cases of false belief where the belief arises in "untypical" circumstances and so leads to disadvantageous action? However note that it is only the *past* predominance of true belief over false that is required: as with any explanation by reference to a selection process, what matters is the effects which have been produced in the past by the item to be explained, not those which might or might not be produced in the future. And so the natural selection approach leaves it open that the statistical norm from now on might be falsity rather than truth.

One obvious way in which this might come about is through a change in the environment. Suppose that tokens of a given belief-type are caused by certain cues (the visual appearance of a tree) which in the past have almost invariably coincided with the relevant truth-condition (an actual tree). But suppose the environment changes so that this coincidence is disrupted (most of the trees on earth come to be imitations put there by considerate Martians). Then, in so far as we are constituted to arrive at tree beliefs on the basis of visual observation, most of our beliefs as to the presence of trees will henceforth be false. But even so the right explanation of our having the belief-type in our repertoire would still be that (in the past) tokens of that type had "typically" arisen in the presence of trees; and to that extent it would still be true that the purpose of that belief was to register the presence of trees. (Perhaps the situation wouldn't be stable: in so far as the false beliefs led to disadvantageous action the disposition to form those beliefs would be selected against. But that's another story, and one which on my approach would lead to a different belief-type and/or a different truth-condition.)

VIII

Thus far, and in particular in these last few remarks about false beliefs, I have been assuming that all beliefs are unstructured "feature-placers" arising directly from observation ("tree," "warmth," "light," etc.). But of course this is a gross oversimplification. To start with, any sensible version of the functionalist approach will need to allow that beliefs are made up of various components ("concepts") combined in various ways, and that the overall causal role of any belief depends on the components it contains and the way they are combined (Field 1978). And, correspondingly, when we turn to representational questions, we would expect the truth-condition of any given belief to be made up of elements and according to a mode of composition which were again functions of the conceptual components and structure of the belief in question.

How does this come out in teleological terms? We can say that the biological purpose of a given concept is to allow us to have certain beliefs, and that the purpose of such beliefs is, as before, to be present when certain states of affairs obtain. But now we should allow that which states of affairs these are will depend in turn on the concepts and modes of composition making up the belief in question. There is clearly an element of circularity here. But it is not vicious. Indeed this kind of circularity is a familiar theme in the theory of meaning: it is only in the context of a sentence (or belief) that a word (concept) has a meaning, but at the same time the meaning of a sentence (belief) depends on the words (concepts) it contains. And it is clear enough how the requisite natural selection story would go. Concepts get selected because the way they combine to give rise to beliefs ensures that those beliefs are typically found in circumstances where the actions they direct will be advantageous. At bottom it is the *concepts* that get selected for, because what beliefs we form and what circumstances they arise in depends on our concepts and how they operate. But it is only when concepts *do* combine into beliefs that they have an effect on action, and so it is only by way of ensuring that certain belief states typically arise in certain circumstances that concepts get selected for at all.

There is another dimension in which the analysis so far has been oversimple. Not only are our belief-types not unstructured, they are not in general purely observational either. Certain concepts are essentially such that their application is a matter of inference. This phenomenon is extremely widespread. Apart from the traditional examples of concepts for scientific "unobservables," there are good reasons for holding that no dispositional concepts ("soluble," "fragile") or multi-criterial concepts ("mass," "length") can be defined in terms of observational concepts or truth-functional combinations thereof (Papineau 1979, Ch. 1).

Such nonobservational concepts have, so to speak, a perfectly determinate place in the Quinean web—it is clear enough what inferences from observation are supposed to warrant their assertion or their denial. But the fact that they can't be equated with any of their observational manifestations makes it natural to think of them as having truth-conditions which in some sense stand behind and transcend the observable world. From a biological perspective it is clear why we should have developed the general cognitive ability to form concepts of this kind. Since many significant features of the world aren't directly (or invariably, or conclusively) observable, it is clearly advantageous that we be able to construct a Quinean web, a cognitive network in which the identity of the conceptual knots depends not on direct links with observational processes, but also on their inferential connections with other such knots. And given that we do construct mental models of this kind, we need to think of the referents of such concepts, not in terms of the properties it is the biological purpose of certain observational processes to register, but rather, in a holistic way, as those properties which have a structure of causal relations in the actual world which mirrors the inferential structure constituting our cognitive network. This is still to construe representation teleologically. But truth-conditions do not now depend on the biological purpose of a single belief-forming process specific to a given belief-type, for nonobservational beliefs do not derive from such "dedicated" processes. Rather their truth-conditions depend on the biological purpose of our "model-building" ability as such—namely, to give us beliefs which will have advantageous results just in case they occur in the presence of properties whose place in the causal structure of the world corresponds to the place of the relevant concepts in the inferential structure of our mental model.

The existence of nonobservational beliefs is relevant to the possibility of falsity. At the end of the last section I argued that it is possible for many (indeed for most) of our beliefs to be false. But at that stage I was presuming that all beliefs were observational. If we abandon this assumption the issue becomes more complicated.

To start with, once we recognize that the concept "tree," say, is not a simple visual concept (since there are any number of independent and indirect ways of telling if something is a tree or not), we can't continue to assume that visual replicas of trees will automatically give rise to false tree beliefs. For the alternative ways of deciding on treehood can come into play and override the misleading visual information. However, this does not rule out the possibility of false belief altogether. Even if nonobservationality multiplies the possible ways of arriving at a given belief, and thereby provides some check on error, it won't eliminate error entirely, for even a plurality of checks might still end up validating a belief on an occasion when it is false. Thus the Martian tree artifacts might be good sensory (olfactory, chemical, etc.) replicas, as well as good visual imitations. (Wouldn't the Martian "trees" have to be somehow detectably different from real trees, to give substance to the thought that it is *false* to believe that Martian replicas are trees? But the question is not whether there exists an identifiable battery of processes which will succeed in distinguishing the imitation trees, but rather whether that battery exhausts the ways in which we *do* arrive at tree beliefs. Given that it doesn't, given that our psychological constitutions will sometimes lead us to tree beliefs in situations where we are not biologically supposed to have those beliefs, the possibility of false beliefs remains with us.)

In addition, the nonobservationality of belief also allows the possibility of a quite different kind of error. I suggested above that a nonobservational concept would refer to that property whose role in the causal structure of the world corresponded to the role played by the concept in question in the cognitive network. But what if there is a mismatch between mental model and the world's causal structure? What, for instance, if the cognitive network embodies the assumptions of some theory which quite misrepresents the way the world actually works? In such a case it is natural to conclude that there is no determinate answer to the question of which property in the world corresponds to a given concept in the cognitive structure. What quantity did the medieval notion of "impetus" stand for? Force? Energy? Momentum? (Feyerabend 1962, Section 6) What is referred to by a concept like the Zande notion of *ira mangu*? Witch-hood? Having an inflated gallbladder? (Papineau 1978, Ch. 6)

How mistaken a conceptual structure needs to be before we decide that its elements fail of determinate reference is a debatable question. And there is room to dispute the extent to which the assumptions of corrigible theories do get built into our concepts. But my present concern is not to resolve these questions, but merely to point out that the teleological approach to representation leaves open the possibility of such nonreferring concepts. We have already seen how, relative to given concepts, the processes by which we arrive at beliefs don't always produce the results they are biologically supposed to, and so leave us with false beliefs—we now see that in addition the "model-building" process by which we construct new concepts needn't always work as it is supposed to either, with the result that we can be left with nonreferring concepts. (Given that falsity and reference failure are possible, one would like some assurance that scientific theorizing in general, and our current theories in particular, are not so deficient. This is not the place to address this task. But I do take it to be a virtue of the teleological approach to representation that it implies there is a substantial question to be answered as to whether our beliefs get reality right or not.)

IX

Let us consider desires once more. I shall simplify by ignoring the complexities of the last section and returning to the fiction of unstructured observational contents. It might seem that we can simply say that the satisfaction condition for a given desire is that characteristic result of the actions it directs which has been selectively advantageous, and the production of which therefore explains the disposition to form that desire. But which result? Take the desire for sweet things. Is the relevant result the taste sensation? The ingesting of sugar? The increased metabolic activity? The survival? The maximizing of inclusive fitness?

The actions which result from a given desire can be "concertinaed" out into a characteristic succession of results, each a means to the subsequent one and so eventually to the maximization of inclusive fitness. Moreover, this chain sometimes breaks down, and one gets the earlier stages (the taste sensation, the sugar, even the survival) without the final evolutionary payoff. And so if we want a characteristic result of a desire which is *always* selectively advantageous, we will have to accept that all desires are the desires for the same end, namely the maximization of inclusive fitness. And this in turn will force us to recognize a lot of beliefs that we didn't know we had, such as the belief that eating sweet things will enhance maximal inclusive fitness. (If the only desire were the "master" desire to maximize inclusive fitness, we would need this belief to get us actually eating.)

One can see some sense in this line of thought. In effect it takes the state that we normally think of as the desire for sweet things, and then construes it instead as the belief that eating sweet things will enhance inclusive fitness. And if one *is* going to construe that state as a belief, then that is indeed the right truth condition to give it (for it is precisely the cases where eating sweet things has enhanced inclusive fitness that have led to the state of wanting sweet things being selected for.) The trouble with all this, of course, is that it misrepresents our psychological structure; it makes us out to be more rational than we actually are. Desire's aren't beliefs, and they don't behave in the same way. Crucially, they don't respond to evidence in the same way. If the state in question really were a belief that eating something sweet will enhance inclusive fitness, then it would simply disappear given information, say, that such an action would simply enhance obesity. But of course (unfortunately for many people) it doesn't, which is why we should count it as a desire and not a belief.

One can think of the situation as follows. From natural selection's point of view the only end *is* maximizing inclusive fitness. Thus natural selection selects actions according as they are an effective means for achieving this end. But which actions will be so effective will depend on environmental circumstances. So natural selection selects organisms with cognitive mechanisms which take as input variable environmental circumstances and which have as output actions which will be effective in those circumstances. But rather than "designing" these cognitive mechanisms so that they take into account all environmental circumstances that might affect whether or not an action was an effective means to eventual inclusive fitness (apart from anything else, the computations involved would no doubt be unmanagable), natural selection takes a "short cut" by having brains take into account only those circumstances that will affect whether the action is an effective means to some "proxy" end (such as sweetness; or, again, sex; or security, etc.).

It is important to keep our distance from the metaphors here. Talk of "proxy" ends needs to be thrown out: what it means is simply that the cognitive mechanism is *not* sensitive to evidence which bears on whether or not the result in question will enhance inclusive fitness.

Thus the full biological explanation of somebody eating something sweet *will* allude to the fact that eating sweet things is by and large good for inclusive fitness. But it is, so to speak, the process of natural selection, and not ourselves, that "believes" in the connection between sweetness and fitness. From the point of view of our cognitive mechanism, this connection is simply "taken for granted"—again, in the specific sense that information about this connection does not affect the operation of our cognitive mechanisms.

It is true that there is another sense in which our cognitive mechanisms do seem to be sensitive to information about the effectiveness of the results we think of our desires as focused on. This relates to the acquisition of new desires: it seems likely that we acquire new desires precisely insofar as the actions they direct (drinking alcohol, say) have in our experience conduced to some further, innately desired, result (feelings of well-being). However, I would prefer to view this as part of the "initial designing" of our "cognitive mechanisms," and not as a feature of their internal operation. This is in line with the suggestion made earlier, that natural selection operates during learning as well as in intergenerational evolution. And the point of looking at the matter this way is, once more, that desires for alcohol are not responsive to further information about the effects of alcohol in the way that a belief that alcohol will make one feel better would be. (After all, if such desires were so responsive, what need would we have for a notion of acquired desires?)

A further complication. Desires fluctuate, and indeed do so in response to environmental circumstances. Thus, for instance, hunger depends on the blood sugar level: one wants to eat if blood sugar is low and eating will restore it to the appropriate metabolic level. But does not my argument then imply that what we naturally think of as the desire for food should instead count as the belief that eating will restore the blood sugar to the appropriate metabolic level? This belief would direct pretty much the same actions as hunger (namely, eating), and it does seem evidentially responsive to the relevant facts. But the responsivity here is of the wrong kind. Evidence that eating will not in fact increase the blood sugar level (because one has some digestive abnormality, say) won't stop one feeling hungry. Again, in the terms I have been using, it is best to think here of different cognitive mechanisms being "switched on" at different times, rather than of one mechanism which "takes account" of what will affect blood sugar level. While in a sense the switching is a matter of some system (the switching system) "believing" that food will restore the blood sugar level, we don't want to count *ourselves* as having that belief, precisely because of the lack of appropriate responsivity of the "belief" to the relevant evidence.

Let me sum up. An action stemming from a desire will have a concertina of effects which are relevant to its enhancing inclusive fitness. As we proceed outwards, so to speak, we will go past effects which are taken to be relevant only in virtue of current beliefs, to effects the relevance of which is assumed by natural selection but not by the agent, and ending up (if everything works well) with enhanced fitness. The satisfaction condition of the desire is the *first* effect which is taken to be relevant by evolution but not by the agent. That's what the desire is *for*—to give rise to actions which have *that* effect. It's not for earlier effects, because whether actions have those effects at all depends on what beliefs the desire is interacting with. And it's not, in the first instance, for later effects, because the actions it directs are designed to produce those later effects only *through* producing that first effect.

Now that we have these conclusions about satisfaction conditions for desires, we can see that the biological explication originally offered for the truth conditions of beliefs was too crude. I originally presented the natural selection of belief-forming abilities as hinging simply on the relevant beliefs having "selectively advantageous" effects. But one implication of the above points about our cognitive mechanisms is that in general beliefs have selectively advantageous effects only insofar as they have effects which satisfy desires. So we should count as the truth conditions of beliefs not simply circumstances in which the resulting actions have advantageous effects, but rather circumstances in which those actions lead to the satisfaction of desires.

In effect what we now have is the earlier idea that truth conditions are circumstances which ensure the success of action, but with the notion of success limited by evolutionary considerations. It is no longer an arbitrary matter what result we count as making an action successful, as satisfying the desire behind it—the result in question is specifically one the past production of which explains the desire being there in the first place.

X

It might seem odd that truth, of all things, should depend on natural selection. Surely, one feels, whether certain states have representational powers depends on how they work *now,* not on where they came from. Suppose, for instance, that you didn't exist, but that a being just like you had spontaneously assembled itself a moment ago as a result of some cosmic accident, some random coagulation of just the requisite molecules, and now found itself in just your situation. Wouldn't that being have just the same beliefs, and about just the same objects, as the beliefs you actually have?

I do indeed want to deny this. And I recognize that denial is, to say the least, counterintuitive. But there are a number of things that can be said in defence of my position here.

The first thing to note is that given the two-aspect approach to belief, there is a sense in which the "accidental replica" does have beliefs, indeed all the beliefs that you have. For on the two-aspect view, to ascribe the belief *that p* to someone is to do two things: firstly, to indicate the causal role of an internal state (by means of the "that *p*" label); and, secondly, to indicate that state's representational powers. And on the first count there is nothing wrong with ascribing beliefs to the replica. By hypothesis, it does have internal states which play exactly the same causal roles in it as your beliefs do in you. (And no doubt what it would phenomenologically be like to be that being is just what it is like to be you.)

But this only removes part of the difficulty. One's intuitions are not only that the replica would have internal states functionally identical to yours. One would also like to say that those beliefs, like yours, are *about* trees, *about* heat, *about* the state of the nation, etc. And my account of representation does preclude me from saying this.

It will be helpful here to separate out the two assumptions that lead to my ending up in this position. In the first place there is the thesis that representation is essentially a teleogical matter. Then there is the further thesis that teleological claims need to be reduced to explanations in terms of natural selection.

The first of these theses can perfectly well be detached from the second. Many commentators (indeed perhaps the majority) would hold that teleological explanation requires only that the item to be explained be shown to have some kind of beneficial effect for some wider system. How the item came to be there (and in particular whether this resulted from selection in virtue of the said beneficial effect) is for them a further question.

Someone who took this line could well accept my teleological account of representation and yet avoid my difficulty with the accidental replica. For it would be open to them to say that the replica had the kind of structure (just like yours) which meant its desires had the "purpose" of getting it to act in ways that produced certain results, and its beliefs had the "purpose" of getting it to act in ways appropriate to specific circumstances. From this point of view the replica's deviant origins would be quite irrelevant to the ascription of biological purposes to its internal states.

Perhaps then I should restrict myself to the claim that representation is teleological. This in itself is certainly something worth showing, and by distancing myself from the commitment to natural selection I could then avoid the replica difficulty. Unfortunately I am thoroughly pessimistic about the chances for alternative accounts of teleology. The natural selection approach gives a simple and clear explication of our intuitions as to when it is appropriate to view some result as the "purpose" of some item: namely, when the presence of the item in question is causally explained by the (past) production of the result. The alternative accounts, in terms of certain kinds of (often cybernetic) structures giving rise to certain kinds of beneficial effects, are in general either too vague to account for our intuitions, or, when spelt out in any detail, easy game for counterexamples. (Imagine, for instance, trying to develop the arguments of the last section, which explicated satisfaction conditions in terms of the biological purposes of desires, purely in terms of some cybernetic analysis of the roles of desires and without recourse to their putative evolutionary history.)

What then of the replica case, where intuition seems (at least *modulo* the teleological theory of representation) to count *against* the natural selection theory of teleology? Well, perhaps there is room to shake this particular intuition. Consider a simpler, accidental creature, a randomly coagulated little green entity that just happened to have a protuberance which enabled it to reach berries down from the trees it happened to find itself nearby. (And make sure you're not thinking of it as constructed by some *other* naturally selected being, for then its characteristics would have derivative purposes.) Is it clear that we should say that this "limb" was there *in order to* pick berries with? And if this isn't clear—if we want to say the limb's not there *for* anything, the creature is just *lucky* to have it—then perhaps we should think twice about saying that your full-blown replica's internal states are desires *for* certain results, or beliefs *about* certain objects. (Perhaps our inclination to say this is merely due to our inability to imagine seriously a complete human replica really arriving by accident.)

In any case, should intuition be decisive here? In general a natural selection understanding of teleology is certainly revisionary of our *concept* of teleology (or, I would say, *was* so revisionary a hundred years ago), and as such could be expected to alter (to have altered) our intuitions about particular cases. If, as I am arguing, representation should come to be seen as a special case of teleology, and hence of natural selection, one could expect this to alter certain intuitions about particular cases of representation too. (And if someone now asks "*Why* we should revise our concepts of teleology to make them answerable to natural selection?" I would reply "How else can we have explanation by reference to effects in a physical world governed by causes?")

XI

I return now to general objections to functionalism. It might seem that I have now vindicated the original suspicions of the view that psychological explanation is a purely "syntactic" matter of causal pushes and pulls. For haven't I now shown how explanations involving propositional attitudes do after all involve the representational powers of those attitudes?

But this would be a confusion. The explanations the functionalist is centrally concerned with, and about which his critic wants to disagree, are explanations of particular actions in terms of particular beliefs and desires, and of particular beliefs and desires in terms of prior circumstances, while the explanations I have now shown to bring in representational considerations are rather explanations of our having dispositions to form beliefs and desires in the first place.

Christopher Peacocke has suggested that "the 'because' in 'He did it because he had such and such reasons' will have something in common with the 'because' in 'the leaf moved because it obtains more light in its new position'" (1981, p. 213). This seems to me to be a mistake. It is rather the "because" in "He had such and such reasons because he then did it" that shares its structure with the biological "because." (Consider, for instance, "We believe the fire will hurt because that leads us to avoid it.")

The situation seems to be this. There is the causal structure postulated by the functionalist, of dispositions to form beliefs and desires given certain circumstances, and of dispositions to act given certain beliefs and desires. By reference to this structure we can explain particular psychological attitudes in terms of the circumstances that produce them. (Thus we can explain somebody's belief that there is a tree in front of them in terms of the presence of a tree, or, for that matter, in terms of the presence of a tree replica.) And we can explain particular actions in terms of the particular beliefs and desires that produce them.

In all this no essential role is played by representational considerations; we simply take the functionalist causal structure as given and proceed from there. Indeed to this extent somebody would be justified who argued, with Stephen Leeds (1978), that there is no explanatory significance to the standard pairing of beliefs with their truth conditions. From the point of view of explaining current beliefs, desires and actions there is indeed no need to bring in "aboutness," and representation can as well be regarded as a more or less useful fiction. But once we start explaining the elements themselves of the functionalist causal structure, the notion of "aboutness" comes into serious play. Once we ask where our dispositions to form beliefs and desires come from in the first place, we need to focus specifically on truth and satisfaction conditions amongst the other systematic causes and effects of our beliefs and desires.

By way of analogy, consider some simpler biological subsystem, such as, say, the liver. We have some more or less structural understanding of how the thing works. We can, if we like, reflect on the evolutionary purposes of the various parts, and this will demarcate a picture of the "normal" working of the organ. But we know that the organ doesn't always work as it is supposed to, that it has various "pathological" states and activities. The point I am stressing is that relative to our grasp of the organ's structure the pathological processes are as well understood, and in the same way, as the normal processes. The distinction between "normal" and "pathological" has to do with the original *point* of the system's parts, and as such plays no substantial part in our grasp of how the thing currently works.

Thus also with our understanding of human psychology. We know how it is supposed to work—which states of affairs beliefs are supposed to register, what results our desires are supposed to eventuate in. But we also know it doesn't always work like that. We can believe that there is a tree out there when there isn't. Even when our beliefs are true, there is the further possibility (which I haven't previously brought out) that our choice of action can be "irrational," and for that reason fail to satisfy our desires. Yet we have no difficulty in such cases in giving perfectly ordinary psychological explanations of what someone does—we might, say, explain somebody's attempt to cut something down in terms of their believing it is a tree, or, again, in terms of their wanting to sell it at a profit.

We are fully aware that our psychological processes do not always work as they should, and thus they can offer such explanations even when, "pathologically," there isn't a real tree there or in a case where more careful reflection would show the agent that there was no profit to be made. (Of course we still identify the belief in such a case as the belief *that there's a tree there,* or the desire as the desire *for a profit*. But this does not bring the object of the attitude into the explanation. It is simply a matter of what we *call* the attitude. See Section II above.)

XII

Let us turn now to one final set of worries. It might be argued that in the case of the liver we have some grasp of the molecular makeup of the organ, and that it is because of this that we understand its pathological workings as well as its normal ones. But it is precisely such direct physiological knowledge that we lack in the case of the mental. Are we really entitled then to credit ourselves with a knowledge of mental structure which goes beyond that yielded by teleological considerations?

Daniel Dennett has argued in this connection that any organism that has evolved by a process of natural selection (or, for that matter, any nontrivial machine designed by such organisms) will contain *some* kind of cognitive mechanism which enables it to respond appropriately to its environment. But no doubt the structure (and not just the physics, but the second-order, functional structure) of such cognitive mechanisms will be different in different such organisms, including humans. Still, argues Dennett, this does not mean that we cannot apply our everyday belief-desire psychology to the general range of such intelligent beings. For according to him our everyday psychology is not designed to answer to the specifics of cognitive structure, even in humans, but it is rather an all-purpose explanatory approach the applicability of which is ensured solely by the fact that the behavior of the organisms in question is somehow appropriately directed (Dennett 1978; 1981; 1982).

This line of thought, then, might make one doubtful of a purely "syntactic" psychology. I have argued that the semantic properties of mental states come into play once we consider the biological purposes of those states. But on Dennett's account the explanatory workings of beliefs and desires depend on nothing other than a general presumption of biological purposiveness. So it might seem that belief-desire explanations cannot proceed independently of semantic relations after all.

However, I see no reason to accept Dennett's original premise that belief-desire psychology is guaranteed by general evolutionary considerations. On the contrary, it seems clear that belief-desire psychology does lay claim to facts of cognitive structure not so guaranteed (and, correlatively, that its explanations do not hinge on representational relations).

As far as I can see, the general evolutionary considerations in question guarantee no psychology at all. Without some substantial idea of an organism's cognitive mechanism, all we would be able to do by way of explanation, given some bit of its behavior, would be to point out that behavior's general biological appropriateness—that it had some effect, if it did, that it enhanced inclusive fitness. And correspondingly we could make a vague, general prediction that as a rule (though not necessarily always) what the organism does will somehow be so appropriate.

But of course we can actually do much better than that. This is because by viewing behavior as stemming from beliefs and desires, we are already postulating some specific cognitive structure behind that behavior. In a sense the whole point of bringing in belief-desire psychology is to get a definite hold on the specific ways in which an organism deviates from evolutionary optimality, on the ways in which it tends to be inaccurate and irrational. As long as an organism *is* accurate and rational, simple biological appropriateness will account for its behavior. But given that organisms aren't always biologically optimal, we want to be able to distinguish and anticipate the *different* ways in which it can go wrong.

There is, it is true, an abstract sense in which any intelligent organism's cognitive mechanism has *somehow* to register the facts on the one hand and *somehow* to get moved to certain ends on the other. But, to repeat the point, if that were all we knew about an organism, we wouldn't have any explanatory or predictive hold on its behavior other than that it would be somehow sensible. In particular we wouldn't be able to distinguish, for explanatory purposes, (and indeed on occasion to anticipate, for predictive purposes) cases where its behavior is inappropriate because the beliefs behind it are false, as opposed to cases where the behavior is inappropriate because of misplaced desires (not to mention the yet further cases where the behavior is inappropriate because the "choice" of behavior is unsuitable).

As Dennett himself has stressed, there needn't be anything corresponding to a neat structure of actual beliefs and desires inside an intelligent organism's head. It's various cognitive states might have information and affect all muddled up together in a way that nevertheless worked out all right. Or it might work by very quickly switching from one single-purpose "frame" to another. No doubt artificial intelligence researchers could suggest further alternatives. The point I am making is that if we were dealing with such organisms then, *pace* Dennett, we would be ill-advised to continue attributing beliefs and desires to them. For the explanatory and predictive distinctions that give such attributions point will not apply to such organisms, evolved and intelligent as they may be.

Of course there is no reason to take it for granted that *we* work according to belief-desire psychology. Certainly the detailed structure I uncovered in our psychological theories about ourselves in Sections 7 to 9 raises questions about the availability of evidence for those theories. On the other hand, one could argue that the practical success of everyday psychology in itself shows that it must at least be approximately true. But, in any case, if we indeed don't work according to belief-desire psychology, the moral is not that belief-desire psychology is some kind of guaranteed transcendental perspective not answerable to the structure of our brains, but simply that we ought to change our theory of our psychological workings.

More generally, it should now be clear what is wrong with the suggestion that entirely general evolutionary considerations make it possible for us to understand propositional attitudes directly in terms of their representational powers, and independently of any assumptions about cognitive mechanisms; namely, that (whatever the correct theory) it is only in the context of *some* theory of cognitive structure that there is any point in introducing propositional attitudes at all.

Of course, once we have got such a theory, and with it the attribution of propositional attitudes and other mental states, we can *then* ask about the evolutionary explanation of dispositions to form those states, and thereby bring in representational powers. But these evolutionary considerations (the ones which I have been elaborating in this paper) will arise after we have some theory of cognitive structure; they will not lead us to it.

REFERENCES

Dennett, D. (1978), *Brainstorms*. Montgomery, Vermont: Bradford Books.

———. (1981), "Three Kinds of Intentional Psychology", in *Reduction, Time and Reality,* R. Healey (ed.). Cambridge: Cambridge University Press.

———. (1982), "Beyond Belief", in *Thought and Object,* A. Woodfield (ed.). Oxford: Clarendon Press.

Feyerabend, P. (1962), "Explanation, Reduction and Empiricism", in *Minnesota Studies in the Philosophy of Science. vol. III.* H. Feigl and G. Maxwell (eds.). Minneapolis: University of Minnesota Press.

Field, H. (1978), "Mental Representation", *Erkenntnis 13*: 9–61.

Leeds, S. (1978), "Theories of Reference and Truth", *Erkenntnis 13*: 111–29.

Lewis, D. (1972), "Psychophysical and Theoretical Identifications", *Australian Journal of Philosophy 50*: 249–58.

Loar, B. (1981), *Mind and Meaning*. Cambridge: Cambridge University Press.

McGinn, C. (1982), "The Structure of Content", in *Thought and Object,* A. Woodfield (ed.). Oxford: Clarendon Press.

Peacocke, C. (1981), "Demonstrative Thought and Psychological Explanation", *Synthese 49*: 187–217.

Papineau, D. (1978), *For Science in the Social Sciences*. London: Macmillan.

———. (1979), *Theory and Meaning*. Oxford: Clarendon Press.

———. (1984), "Social Facts and Psychological Facts", in *Popper and the Human Sciences,* G. Currie and A. Musgrave (eds.). The Hague: Martinus Nijhoff.

Putnam, H. (1967), "Psychological Predicates", in *Art, Mind and Religion,* W. Capitan and D. Merrill, (eds.). Pittsburgh: University of Pittsburgh Press.

Wright, L. (1973), "Functions", *Philosophical Review 82*: 139–68.