Offprinted from *MIND: A quarterly review of philosophy*
Vol LXXXVI No 342 April 1977

The Language of Thought. By JERRY FODOR. New York: Thomas Crowell, 1975; Hassocks, Sussex: Harvester Press, 1976. Pp. x + 214. £8.95.

D. C. DENNETT

We and other creatures exhibit intelligent behaviour, and since the regular production of such behaviour requires thought, and since thought requires representation, and since nothing can represent except within a system, we must be endowed with and utilize a system of internal representation having its own 'grammar' and 'vocabulary', which we might call the language of thought.

This argument has seldom been brought into the open and examined, but behind the scenes it has motivated and flavoured large bodies of philosophical doctrine, and strongly influenced research strategies and theories in psychology, linguistics, computer sciences and neurophysiology. It is worth asking why such an influential move has been so comfortably ignored until recently. It is not plausibly one of those drifts of thought that seem too obvious to need spelling out; perhaps it has been avoided because once one attempts to put the argument in proper and explicit shape, incoherencies, paradoxes, infinite regresses and other alarming implications seem to arise at every turning. Now Jerry Fodor has done us the fine service of propounding and defending a vigorous, unblinking, and ingenious version of the argument. Many of his conclusions seem outrageous, and the threats of incoherency are now close to the surface, but Fodor argues with great persuasiveness that these are in fact parts of the foundation of deservedly esteemed schools of thought in philosophy, cognitive psychology and linguistics. If he has produced an unintended *reductio ad absurdum* (a possibility he cheerfully admits), some of our favourite edifices will topple with him. He may be wrong, of course, but the challenge is well presented, and since recently thinkers in all the jeopardized fields have been converging on just the perplexities Fodor discusses, the challenge will not be ignored. The main issue treated in Fodor's book is fast becoming a major topic of interdisciplinary interest, and philosophers of mind who have squeezed the last drops of enlightenment out of the debate over the identity theory or the individuation of actions should be pleased to find here some important and fascinating problems to engage their talents. What is needed is nothing less than a completely general theory of representation, with which we can explain how words, thoughts, thinkers, pictures, computers, animals, sentences, mechanisms, states, functions, nerve impulses, and formal models (*inter alia*) can be said to represent one thing or another. It will not do to divide and conquer here—by saying that these various things do not represent in the same sense. Of course that is true, but

what is important is that there is something that binds them all together, and we need a theory that can unify the variety. Producing such a theory is surely a philosophical endeavour, but philosophers must recognize that some of the most useful and suggestive work currently being done on the problems is being done by psychologists, linguists, and workers in artificial intelligence.

For what it is worth, Fodor probably holds uniquely strong professional credentials for the task of consolidating the insights from these fields, for he holds a joint appointment in psychology and philosophy (at M.I.T., a major centre for work in linguistics and artificial intelligence) and has made important contributions to experimental psycholinguistics and linguistics in addition to his work in philosophy. Fodor is not beset by the philosophical naiveté of many of his colleagues in psychology and he has as powerful a grasp of current thinking in linguistics and psychology as anyone in philosophy. Indeed, the overall savvy of his book is one of its most striking characteristics, mainly for good but also for ill. There cannot be many readers well equipped or disposed to appreciate all his knowing nudges; the uninitiated will perhaps be the unpersuaded (and unamused) as well. I fear that Fodor's unfailing high spirits and jocosity may hurt his cause by irritating as many readers as they amuse. I find the book genuinely witty, however, and can only urge those who resent being tickled while engaged in such serious business to make an extra effort to distinguish the medium from the message.

Fodor's message has three parts. First he describes and promotes a brand of theorizing he calls cognitive psychology, but clearly he means to cast his net wider, and in places narrower, than that term would suggest. The distinguishing mark of this theorizing is the unapologetic utilization of Intentional characterizations of processes and 'intellectualist' analyses of perception and other 'cognitive processes' in terms of information-flow, hypothesis-testing, inference and decision-making. Within its boundaries fall much current psychology, linguistics, artificial intelligence, and some strains of thought in current philosophy. Let us call it neo-cognitivism, for it is not markedly continuous with earlier schools of cognitive psychology, nor is it all clearly psychology. It has developed largely in recognition of the impotence of (psychological and logical) behaviourism, and its inspiration is drawn largely from linguistics, computer science and (come to think of it) the last three hundred years of epistemology. Fodor attempts to establish the credentials of neo-cognitivism by showing how it avoids the doldrums of Rylean logical behaviourism, steers between the Scylla of dualism and the Charybdis of reductionism to emerge as the only straw floating—as Jerome Lettvin once put it. In this first part Fodor has a strong and persuasive case on almost all counts.

Fodor's second task is to show that this best hope for a confirmed, powerful psychological theory inescapably requires the postulation of internal representational systems. These systems, though designed for computation rather than communication, have structures—and other features—so like those of natural languages that we may—and should—

speak of the language of thought: the medium in which the computational transactions are performed that ultimately govern our behaviour and the behaviour of other intelligent creatures as well. This is the philosophic heart of Fodor's book, and will receive detailed attention below.

Third, Fodor completes his book with two lengthy chapters purporting to show how evidence from linguistics and psychology establishes answers to an impressive variety of questions about 'the structure of the internal code'. Having proved the existence of Planet X, he proceeds to detail its climate and geography for us, using data that had been available but hitherto mute. These chapters are undeniably compelling, for every now and then one gets glimmers of the sort of fruitful falling-into-place so seldom encountered in psychology or philosophy of mind. Whereas, for instance, behaviourism has always worn the guise of a properly endorsed method (a 'methodology') in dogged search of results, here we seem to see an abundance of results and tempting hypotheses to test for which we must somehow concoct methodological permission.

For example, linguists have devised a variety of competing formal systems for more or less algorithmically generating or analysing sentences, and a question the psycholinguist asks is which if any proposed formalism has 'psychological reality', or in other words describes or mirrors real psychological processes occurring in the production or comprehension of sentences. This empirical question is to be settled independently of the elegance or power of the formal systems. (One way of dividing 67 by 12 is to subtract 12 from 67, then subtract 12 from 55, and so forth, while counting the subtractions; another is long division; which if either has 'psychological reality' for an individual human calculator is surely an empirical question, and *asking the calculator* is not the only, or always the best, way of answering the question.) Subtle studies of reaction times, relative difficulty of comprehension, patterns of errors, and so forth often provide satisfyingly clearcut verdicts on these questions, but often only if we make just the sorts of assumptions about representational machinery Fodor is attempting to vindicate. Whether these investigations will continue to ramify nicely is far from assured, however, and there is an abundance of danger signals for sceptics to make of what they can.

The conclusion Fodor wishes to draw from this examination is bracingly unqualified: '. . . having a propositional attitude is being in some *computational* relation to an internal representation'. 'Attitudes to propositions are . . . "reduced" to attitudes to formulae, though the formulae are couched in a proprietary inner code' (p. 198). The inner code is innate, and one's innate vocabulary of predicates must be sufficient to represent, by logical construction, any predicate of any natural language one can learn. Once one learns such a predicate one may augment one's inner code with a synonym, as it were, of the natural language predicate and henceforward use this non-native inner word as an abbreviation for the cumbersome truth-functional molecule of native mentalese (p. 152). We aren't born with an inner code word for 'airplane' but if we couldn't form at the outset a predicate of inner mentalese at least coextensive with 'airplane' we could never learn what 'airplane' meant, could never add an 'airplane'-synonym to our basic stock. So there is a sense in which

one cannot 'acquire new concepts' by learning a language, even one's mother tongue.

All this (and there is more) is hard to swallow, but what are the alternatives? Thinkers as diverse as B. F. Skinner, Norman Malcolm and Hubert Dreyfus have insisted that the very concept of neural systems of representation is a monstrous error. Let us call that the extreme right wing view. On the extreme left, then, would be researchers such as McConnell and Ungar, who take brain-writing so literally that they suppose one might physically extract token sentences of the inner code from one creature and teach another by injection or ingestion. (Ungar reports he has trained cats to fear the dark and then isolated a substance in them, 'scotophobin', which injected into untrained cats causes them to fear the dark!) Middle-of-the-road positions have yet to be formulated in satisfactory detail, but it is safe to say that Fodor has laid claim to a position far to the left of centre and is insisting that no less extreme position can provide the foundations for the promising theories of neo-cognitivism.

Let us return to the beginning and examine Fodor's case in some detail. Fodor takes his first task to be protecting neo-cognitivism from two philosophic threats: Ryle's attack on intellectualist theorizing, and the physicalist demand that all theoretical terms be reducible somehow to the terms of physics. Fodor sees these as in different ways suggesting the charge that neo-cognitivism is dualistic (a verdict Fodor would view as at least discouraging and probably fatal). The charge is familiar: the characteristic predicates of cognitive psychology are Intentional or 'mentalistic' idioms, and since mentalism is dualism, cognitive psychology is dualistic. Certainly in the past this has been an influential train of thought; Brentano did after all reintroduce the concept of Intentionality precisely as the distinguishing mark of the non-physical, and (though probably not influenced by Brentano) Skinner has for years seen the spectre of dualism in every variety of 'mentalistic' theorizing. The claim has not however figured influentially in recent philosophic work in the area. On the contrary, the coexistence of physicalistic doctrine with Intentional or mentalistic vocabulary, while perhaps not having received the justification it ought to have, is a typically undefended and unattacked feature of current discussions.

It is a bit curious then that a rebuttal of the dualism charge should find pride of place in Fodor. Perhaps he is addressing the many psychologists who haven't heard and are still swayed by Skinner's suspicions. More curious still is Fodor's choice of Ryle as the initial target of his rebuttal. In *Psychological Explanation* (1968) Fodor went to great lengths to refute his version of Ryle's 'logical behaviourism' and in 1975 he has still not been able to remove his hands from this tarbaby. Now it is clearer why Ryle should exercise him so, for he has clarified his interpretation with a cute example. Why are Wheaties (as the ads say) the breakfast of champions? Because, says the dietician, they contain vitamins, etc. Because, says the Rylean, they are eaten for breakfast by a non-negligible number of champions. The former is a 'causal' explanation, the latter is

a 'conceptual' explanation and, according to Fodor, Ryle's view is that the latter explanation is in competition with the former. When a question should have a conceptual answer it can't have a causal answer as well. Questions like 'What makes the clown's clowning clever?' have conceptual answers and, according to Fodor's Ryle, therefore can't have causal answers—'Alas for the psychology of clever clowning'.

Fodor's demolition of this notion should be, and is, obvious, and as an interpretation of Ryle it is almost right; the Wheaties example does most effectively illuminate a central Rylean distinction, and there are many passages in *The Concept of Mind* that could be cited to support the claim that Ryle deserves to be so interpreted. But Ryle does not, as Fodor thinks, offer *The Concept of Mind* as a psychological theory or as a substitute for psychology or as a proof that psychology can't be done. Fodor seems to be pointing out that questions like 'What makes the clowning clever?' are ambiguous, but he does not see, or accept, the implication that in such cases there are two questions one can be asking. If there are two questions, it can be true that one cannot answer a question requiring a conceptual answer with a causal answer, which is Ryle's point, without it being true that psychology and philosophy of mind are in competition. Fodor has construed Ryle's attack on intellectualist theorizing (involving the postulation of inner cognitive processes) as an attack on intellectualist solutions to problems in psychology, while Ryle intended it primarily as an attack on intellectualist solutions to the conceptual problems of philosophy. In fairness to Fodor's interpretation, Ryle does strongly suggest that cognitivistic or 'paramechanical' hypotheses and the like are bankrupt as psychology as well (see especially the last chapter of *The Concept of Mind*) and against that excessive strain in Ryle's thought Fodor's arguments—and indeed the whole book—are a welcome antidote. But in the process of magnifying and rebutting the worst in Ryle, Fodor misconstrues Ryle in another fashion that leads him to overlook a more penetrating Rylean objection to his enterprise 'Ryle assumes', Fodor tells us, '. . . that a mentalist must be a dualist; in particular, that mentalism and materialism are mutually exclusive'. Hence the 'tendency to see the options of dualism and behaviourism as exhaustive in the philosophy of mind' (p. 4). Were we to replace 'Ryle' with 'Skinner' and 'philosophy of mind' with 'psychology' in this passage there would be no quarrel, but in the sense of the term in which behaviourism is the chief rival empirical theory to Fodor's mentalism, Ryle is no behaviourist but a sort of mentalist himself. Ryle does not attempt, as Skinner does, to explicate mentalistic predicates '(just) in terms of stimulus and response variables' (p. 8). On the contrary, his explications are typically replete with Intentional idioms. Ryle's familiar account of vanity, for instance (whatever its problems) is not that vanity is a disposition to perform certain locomotions, utter certain sounds, respond to certain stimuli, but that it is a disposition to try to make oneself prominent, to ignore criticism, talk about oneself, avoid recalling past failures, 'indulge in roseate daydreams about his own successes' (*The Concept of Mind*, p. 86). What kind of behaviourism is that? Not any kind to be found in psychology. Ryle's disagreements with

Fodor are fundamental, but they are not to be discovered by allying Ryle with Skinner.

Perhaps Ryle's view can again be illuminated by a fanciful example. Suppose someone were benighted enough to think the monthly bank statement he received was a historical description of actual transfers of currency among thousands of labelled boxes in bank vaults. He is informed of an overdraft and puts forward a theory of anti-dollars, vacuums and vortices to explain it. The Rylean explains that *nothing like that* is what makes it the case that the account is overdrawn and gives a 'conceptual' account of the situation. The 'logical behaviourist' account of overdrafts is the one we are usually interested in. Of course there is a mechanical story about what happens at the bank that can be told as well, and perhaps knowing it will help us understand the conceptual account, but the two are distinct.

Fodor does not seem to see this point in application to psychology, for he wishes to maintain with regard to the clever clown 'that it is the fact that the behaviour was caused by such [inner cognitive] events that makes it the kind of behaviour it is; that intelligent behaviour *is* intelligent because it has the kind of etiology it has' (p. 3—but see also n. 2 of p. 29, where Fodor qualifies this). This claim burkes Ryle's distinction and leads Fodor, I hope to show later, to a mistaken account of what makes it the case that something represents something.

Setting aside this difficulty, Fodor has shown, contra Ryle, that there is some real work that the mentalistic terms of cognitive psychology might do, but could they do this work while being faithful to the spirit of materialism? Fodor argues that the reasonable belief in the generality of physics, and the reasonable desire that the various sciences be somehow unified, have engendered unreasonably strong demands that the theoretical predicates of the 'special' sciences, and psychology in particular, be 'reducible' to the predicates of physics. Fodor's critique of reductionism and concomitant defence of functionalism is consonant with other recent accounts, especially Putnam's, but makes important additions of detail to this emerging orthodoxy. The unreasonableness of reductionism is nicely illustrated by a discussion of its application to Gresham's Law, a very clear account is given of type and token physicalism and natural kinds, and there is an especially useful development of the claim that it is a mistake to try to make the laws of the unreduced sciences exceptionless. We should look for the laws of physics to be exceptionless, but these laws should not, as the reductionist requires, guarantee that the laws of the reduced sciences have no exceptions, but rather provide an explanation of the exceptions encountered. To reconstrue the laws of the special sciences so that their predicates were locked with the predicates of physics would be to abandon the very utility of the predicates that gave birth to the special sciences in the first place.

Fodor offers a specific positive account of the logical relations that may hold between the terms of a special science (say psychology) and a reducing science (either physiology or physics) which goes far toward establishing the proper independence of the former. I think it could go farther. Fodor shows how a special science can be neutral with regard

to variation in physical realization, and can tolerate variety in the physical tokenings of its types even within the individual, but I think he unnecessarily rules out the possibility that there could be a law of a special science, even an exceptionless law, where there were no *laws* of the reducing sciences relating all the tokenings (because the regularities in token sequences could only be described by conditionals with highly disjunctive antecedents and consequents). It seems essential that he allow for and explain this possibility, for it is fundamental to the capacity of well designed systems to 'absorb' random or merely fortuitous noise, malfunction, interference. The account Fodor gives does not permit the brain to tolerate typographical errors in the inner code, so far as I can see.

Fodor supposes his arguments obtain methodological permission to use mentalistic predicates in theory construction. Why should we want them? Because it is 'self-evident that organisms often believe the behaviour they produce to be of a certain kind and that it is often part of the explanation of the way that an organism behaves to advert to the beliefs it has about the kind of behaviour it produces' (p. 28). In other words, Fodor does not believe another reasoned obituary of behaviourism would be worth space in his book. Very well, but what, exactly, is 'self-evident'? Fodor believes that the everyday, lay explanations of behaviour (of both people and beasts) in terms of beliefs and desires are of a piece with the sophisticated information-flow explanations of the neo-cognitivists, so that the self-evident acceptability of 'the dog bit me because he thought I was someone else' ensures the inevitable theoretical soundness of something like 'the dog's executive routine initiated the attack subroutine because in the course of perceptual analysis it generated and misconfirmed a false hypothesis about the identity of an object in its environment'. Fodor recognizes that it is a fairly large step from everyday, personal-level Intentional explanations to theory-bound sub-personal level Intentional explanations but impatiently dismisses the worry that anything important to his enterprise might hinge on how he took the step: 'There is, obviously, a horribly difficult problem about what determines what a person (as distinct from his body, or parts of his body) did. Many philosophers care terrifically about drawing this distinction . . . but . . . there is no particular reason to suppose that it is relevant to the purposes of cognitive psychology' (p. 52). We shall see.

Fodor's next task is to show how neo-cognitivist theory is unavoidably committed to a language of thought. He begins by offering three different but related demonstrations, and similar problems attend each. First, Fodor presents a schema for neo-cognitivist theories of 'considered action'. Any such theory will suppose the

> 'agent finds himself in a certain situation (S) . . . believes that a certain set of behavioural options . . . are available to him . . ., computes a set of hypotheticals roughly of the form if B_1 is performed in S, then, with a certain probability, C_1. . . . A preference ordering is assigned to the consequences. . . . The organism's choice of

behaviour is determined as a function of the preferences and the probabilities assigned' (p. 28–29).

In other words, a normative decision theory is to be adapted as a natural history of cognitive processes in the organism, and for such a history to be true, agents must 'have means for representing their behaviour to themselves'. 'For, according to the model, deciding is a computational process; the act the agent performs is the consequence of computations defined over representations of possible actions. No representations, no computations. No computations, no model' (p. 31). Moreover, 'an infinity of distinct representations must belong to the system' for 'there is no upper bound to the complexity of the representation that may be required to specify the behavioural options available to the agent' (p. 31).

Note that this argument assumes there is a clear line between computational processes and other processes, and another between considered action and mere reactivity. Fodor does not intend his argument to apply only to the psychology of human beings, but how plausible is it that a mole or a chicken or a fish is capable of representing behavioural options of unbounded complexity? The famous four F's (fighting, fleeing, feeding and sexual intercourse) would seem to be a plausible initial tally of options, and even if we allow, say, a dozen variations on each theme we hardly need a productive representation system to provide internal vehicles for them all, and the process that led to the appropriate 'choice' in such a case would not often appear to be computational, unless all processes are. Presumably a diving bell does not compute its equilibrium depth in the water, though it arrives at it by a process of diminishing 'corrections'. Does a fish compute its proper depth of operations? Is there an important qualitative difference between the processes in the fish and the diving bell?

Fodor's way of dealing with these problems is best understood by contrast with the paths not taken. Fodor could have claimed that only the behaviour of human beings (and other smart creatures of his choosing) is governed by truly computational processes, or he could have gone to the other extreme and granted the diving bell its computational processes. Or he could have defended an intermediate position along these lines: all creatures of noticeable intelligence make decisions (I who 'care terrifically' would insist that at best something decision-like occurs within them) and as we ascend the phylogenetic scale the decision-processes are more and more aptly characterized as computational; all creatures of noticeable intelligence have at least rudimentary representational systems, but only in higher creatures are these systems language-like in being productive or generative. Instead, he adopts the line that there is a radical discontinuity between computational and non-computational processes: 'What distinguishes what organisms do . . . is that a *representation of the rules they follow constitutes one of the causal determinants of their behaviour*' (p. 74, n. 15). If Fodor is to distinguish this claim from the other options he must mean that these rules are *explicitly* represented (not implicitly represented in virtue of functional organization, that is), and this is the radical heart of Fodor's position. I will discuss some problems with it later.

Fodor's second demonstration concerns what he calls concept learning, roughly, coming to distinguish and attach importance to some particular class of things or stimuli in one's environment (e.g., learning about green apples, or learning not to press the bar until the buzzer sounds, or learning to put the red circles in one pile and everything else in another). Fodor claims that 'there is only one kind of theory that has ever been proposed for concept learning—indeed, there would seem to be only one kind of theory that is conceivable—and this theory [that concept learning proceeds by hypothesis formation and confirmation] is incoherent unless there is a language of thought' (p. 36). Why? Because the hypotheses formed must be 'couched' in representations. A striking point Fodor makes about this is that experiments can distinguish between *logically equivalent* but 'notationally' different formulations of hypotheses in concept learning. The idea is that taking the spades out of a deck of cards is easier, oddly enough, than leaving all the cards that are not spades in the deck. When presented with the latter task if one does not think 'in other words . . .' one's performance will suffer. Does this consideration, and similar ones, not establish beyond a shadow of a doubt the 'psychological reality' of the representations? As in the first case, it all depends on how far Fodor is prepared to descend with his talk of representations. What is somewhat plausible in the case of human beings is not at all plausible in the case of lower animals, and it seems that even insects can achieve *some* concept learning. Either some very primitive concept learning does not require hypothesis formation and confirmation, or if it all does, some hypotheses are formed but not 'couched', or just about any feature on the inside of a creature can be considered a representation. Certainly the psychological reality of something functioning in some ways rather like a representation is established by the results Fodor cites, but Fodor has prepared a buttered slide for us and anyone who does not want to get on it might well dig in the heels at this point and ask for more details.

Fodor's third demonstration concerns perception. Here his point is that, as empiricists have insisted (for all the wrong reasons) 'the sensory data which confirm a given perceptual hypothesis are typically internally represented in a vocabulary that is impoverished compared to the vocabulary in which the hypotheses themselves are couched' (p. 44). For instance, the empiricists would say that my hypothesis is that there is an apple out there, and my data are that I seem to see this red round patch. Fodor likes the idea of perception proceeding by a series of computational processes taking descriptions in one vocabulary and using them to confirm hypothesized descriptions in another vocabulary and the main thing he sees wrong with empiricist versions of this is their penchant for couching the given in the 'theory-free language of qualia' rather than the 'theory-laden language of values of physical parameters' (p. 48). What is given is the excitation of a sensory mechanism sensitive to a physical property. 'Hence, there is no reason to believe that the organism cannot be mistaken about what sensory descriptions apply in any given case' (p. 48). But here Fodor seems to me to have lost track of the important distinction between the content of a signal to the system it informs and the content

we on the outside can assign it when we describe the signal and the system of which it is a part. For instance, one badly misconceives the problem of perception if one views the retinal receptors as 'telling' the first level of hypotheses testers 'red wavelength light at location L again', for that level does not utilize or understand (in *any* impoverished sense) information of that sort. What it gets in the way of data are at best reports with uninterpreted dummy predicates ('it is intensely F at location L again') and out of these it must confirm its own dummy hypotheses.[1] In fact is either way of talking appropriate? The *vocabulary* of the signals is not something that is to be settled by an examination of tokens, at least at this level, and when we turn to indirect evidence of 'psychological reality' any evidence we turn up will perforce be neutral between interpreted and uninterpreted predicates.

What content is to be assigned to events in the nervous system subserving perception? That, I take it, is a rather important question for cognitive psychology to answer. It cannot be answered, I submit, until one gets quite careful about who (or what, if anything) has access to the candidate representation—for whom or for what the thing in question is a representation. As Michael Arbib has suggested, what the frog's eye tells the frog's brain is not what the frog's eye tells the frog.

Fodor rather nonchalantly dismisses such distinctions. Why are they important? Suppose we make the following extension of his main argument. The only psychology that could possibly succeed is neo-cognitivist, which requires the postulation of an internal system of representations. However, nothing is intrinsically a representation of anything; something is a representation only *for* or *to* someone; any representation or system of representations requires at least one *user* of the system who is external to the system. Call such a user an *exempt agent*. Hence, in addition to a system of internal representations, neo-cognitivism requires the postulation of an inner exempt agent or agents—in short, undischarged homunculi. Any psychology with undischarged homunculi is doomed to circularity or infinite regress, hence psychology is impossible.[2]

The problem is an old one. Hume wisely shunned the notion of an inner self that would intelligently manipulate the ideas and impressions, but this left him with the necessity of getting the ideas to 'think for themselves'. His associationistic couplings of ideas and impressions, his pseudo-chemical bonding of each idea to its predecessor and successor, is a notorious non-solution to the problem. Fodor's analogous problem is to get the internal representations to 'understand themselves', and one is initially inclined to view Hume's failure as the harbinger of doom for all remotely analogous enterprises. But perhaps the *prima facie* absurd notion of self-understanding representations is an idea whose time

1 Cf. J. J. C. Smart, *Philosophy and Scientific Realism* (1963), on 'topic-neutral reports'. The epistemic status of reports with uninterpreted predicates and reports with qualia-predicates would seem to be the same.

2 Cf. my 'Why the Law of Effect Will Not Go Away', *J. Theory of Social Behavior*, October 1975, and Fodor's 'The Appeal to Tacit Knowledge in Psychological Explanation', *J. Philosophy*, lxv (1968), 627–640.

has come, for what are the 'data structures' of computer science if not just that: representations that understand themselves? In a computer, a command to dig goes straight to the shovel, as it were, eliminating the comprehending and obeying middleman. Not *straight* to the shovel, of course, for a lot of sophisticated switching is required to get the right command going to the right tools, and for some purposes it is illuminating to treat parts of this switching machinery as analogous to the displaced shovellers, subcontractors and contractors. The beauty of it all, and its importance for psychology, is precisely that it promises to solve Hume's problem by giving us a model of vehicles of representation that function without exempt agents for whom they are ploys. Alternatively, one could insist that the very lack of exempt agents in computers to be the users of the putative representations shows that computers do not contain representations—real representations—at all, but unless one views this as a rather modest bit of lexicographical purism one is in danger of discarding one of the most promising conceptual advances ever to fall into philosophers' hands.

Fodor almost parenthetically makes these points (in a footnote on p. 74 where he roundly rebuts an ill-considered version of the homunculus argument of mine). He is justly unafraid of homunculi, for they are at most just picturesquely described parts of the switching machinery that ensures the functional roles of the inner messages, but he fails to recognize that they still play the theoretical role of fixing the 'topic' and 'vocabulary' of the messages they communicate. If viewing messages of the inner code as self-understanding representations in this fashion can save Fodor's enterprise from incoherence—and in principle I think it can—it does so by adding constraints to the notion of an internal representation system that emphasize rather than eliminate the distinction between personal level attributions of beliefs and desires and sub-personal level attributions of content to intra-systemic transactions. If there is any future for internal systems of representation it will not be for languages of thought that 'represent our beliefs to us', except in the most strained sense. Fodor notices the strain (p. 52) but decides to tolerate it. The result, for all its vividness, is at least misleading in a way that has an analogy in the history of science. The problem of genetic inheritance used to look all but insoluble. Did the sperm cell contain a tiny man, and if so did the tiny man have sperm cells containing tiny men and so forth *ad infinitum*? Or did the sperm cell contain a picture or description of a human being, and if so, what looked at the picture or read the description? The truth turns out to be scarcely less marvellous than the 'absurd' speculations, what with self-reading, self-duplicating codes and their supporting machinery, but anyone who had insisted all along that somehow the mother finds out from the sperm what sort of baby the father wants would not have been pointing in just the right direction. (This sidelong glance at DNA serves the additional purpose of reminding the sceptics who view the contraptions of artificial intelligence as hopelessly inefficient and 'inorganic' that nature has proved not to be stingy when it comes to micro-engineering solutions to hard problems.)

Earlier I claimed that Fodor's view of computational processes commits him to a radical view of representation. The problems with this hard line on the psychological reality of explicit representations are apparent—indeed, are deliberately made apparent, to Fodor's credit—in his discussion of langage learning. His argument is that the process of learning the meaning of a word, even the initial words of one's native tongue, is and must be a process of hypothesis formation and confirmation, and in particular,

> among the generalizations about a language that the learner must hypothesize and confirm are some which determine the extensions of the predicates of that language. A generalization that effects such a determination is, by stipulation, a *truth rule* (p. 59).

For instance, the truth rule for 'is a chair' is '⌜y is a chair⌝ is true iff Gx' where 'G' is a predicate of one's internal code. Fodor seems to think that the only hypotheses which could *determine* the extension of a natural language predicate would have to be confirmed hypotheses explicitly about that predicate and having the explicit form of a truth rule. But to play Fodor's own game for a moment, couldn't a child learn *something that determined* the extension of 'is a chair' by *disconfirming* the following hypotheses (and others):

⌜x is a chair⌝ is true iff x is red
⌜x is a chair⌝ is true iff x is in the living room
⌜x is a chair⌝ is true iff x has a cushion

and, perhaps, confirming—even *mis*confirming—others, e.g., '⌜x is a chair⌝ is true if x is this object here or that object there', and '⌜x is a chair⌝ is true only if x would support Daddy's weight', without ever *explicitly* representing a confirmed truth rule for 'is a chair'? I suspect that Fodor's reply would be that to learn something determining the extension of 'is a chair' the child must explicitly conjoin all the confirmed hypotheses and the negations of the disconfirmed hypotheses and that somehow this amounts to the confirmation of the explicit truth rule for 'is a chair', but aside from the implausibility of this as a story of real computational processes (but remember DNA) one wants to know what could conceivably count against the presumably empirical claim that whatever could determine the extension of a predicate has the explicit form of a truth rule, if this example did not.

Perhaps Fodor has gratuitously overstated his own best case, for it seems as if he is committed to the impossible view that only explicit representation is representation, and (roughly) nothing can be believed, thought about or learned without being explicitly represented.

> That is, one might think of cognitive theories as filling in explanation schema[ta] of, roughly, the form: *having the attitude R to proposition P is contingently identical to being in computational relation C to the formula (or sequence of formulae) F.* A cognitive theory, insofar as it was both true and general, would presumably explain the productivity of propositional attitudes by entailing infinitely many substitu-

> tion instances of this schema: one for each of the propositional attitudes that the organism can entertain (p. 77).

Perhaps we 'entertain' propositional attitudes either seriatim or at least in manageably small numbers at any one time, but the propositional attitudes we *have* far outstrip those we (in some sense) actively entertain. For instance, it should come as no news to any of you that zebras in the wild do not wear overcoats, but I hazard the guess that it *hadn't occurred* to any of you before just now. We all have believed it for some time but were not born believing it, so we must have come to believe it between birth and, say, age fifteen, but it is not at all plausible that this is a hypothesis any of us has explicitly formed or confirmed in our childhood, even unconsciously. It is not even plausible that having formed and confirmed other hypotheses entailing this fact about zebras, we (in our spare time?) explicitly *computed* this implication.

Fodor does seem to be committed to some such view as this, however. He backs into this corner by underestimating the viability of what he takes to be the only alternative, which he characterizes as a dispositional behavioural analysis of propositional attitudes. 'A number of philosophers who ought to know better do, apparently, accept such views' Fodor says, never doubting that he has seen clearly to the very heart of such silliness. His version of dispositional analysis is so simplistic, however, that he thinks the notion is adequately buried by a quip: 'Pay me enough and I will stand on my head iff you say chair. But I know what "is a chair" means all the same' (p. 63). It is of course true that the arduous piecemeal composition of dispositional definitions of propositional attitudes would be a bootless methodology for psychology (Ryle knew better than to attempt to say, precisely, just what his 'multi-track' dispositions were), but if, as Fodor supposes, the representation-talk of cognitive psychology ultimately gets vindicated by such ploys as computer modelling of cognitive systems and processes, he must be committed in spite of himself to a version of Rylism. For a computer programme is just a very complicated specification of a multi-track disposition (a disposition to be disposed under conditions A, B, C to be disposed under conditions X, Y, Z to be disposed . . . to give output O . . . etc.). Notationally distinct but equivalent programmes are equivalent precisely in that they determine the same multi-track disposition.

Suppose research reveals all the psychologically real computational processes in Mary, and artificial intelligencers programme a robot, Ruth, whose internal processes 'model' Mary's as perfectly as you like. Suppose that Mary believes that p. So then does Ruth. But suppose the artificial intelligencers then give another robot, Sally, a programme equivalent to Ruth's, but notationally and computationally different. Sally may not be a good psychological model of Mary, but Sally, like Ruth and Mary, believes that p.[1] That is, the ascription of all Mary's beliefs and desires (etc.) to Sally will be just as predictive as their ascription to Ruth so far as prediction of action goes. Sally's response delays,

1 Perfect equivalence of programmes is a very strong condition. I would hold it is a sufficient but not necessary condition for sharing Intentional characterizations.

errors, and the like may not match Mary's, but this is not what belief ascription is supposed to predict or explain (cf. Fodor, p. 123). If one agrees with Fodor that it is the job of cognitive psychology to map the psychologically real processes in people, then since the ascription of belief and desire is only indirectly tied to such processes, one might well say that beliefs and desires are not the proper objects of study of cognitive psychology. Put otherwise, cognitivist theories are or should be theories of the subpersonal level, where beliefs and desires disappear, to be replaced with representations of other sorts on other topics.

But unless I am misreading Fodor, he will have none of this. His position simply is that since believing that snow is white couldn't be having a disposition to behave, it must be having a token of the mentalese translation of 'snow is white' installed in some wonderful way in one's head. Perhaps I am misreading him by interpreting 'being in a computation relation to a formula of the inner code' as implying the existence of a real token of that formula in some functionally characterized relation to the rest of the machinery, but the weaker alternative, *viz.*, that one is in a computational relation to a formula if one *can or would produce or use* a token of that formula in some way under some circumstances invokes dispositionalism of just the sort Fodor has presumably forsworn.

None of this is to say that neural representations, even tokens of brain-writing, are impossible. It is not even to deny that the existence of such representations is a necessary condition for cognition. It may well turn out to be. But Fodor, by making explicit coding *criterial* for representation or contentfulness, has committed the very sin he imputes to Ryle: he has confused a conceptual answer with a causal answer. Like neo-cognitivists generally, Fodor wants to be able to assign content to events or other features of systems, to treat them as information-bearers or messages. What makes it the case ultimately that something in this sense represents something within a system is that it has a function within the system, in principle globally specifiable.[1] To say that it has the function of bearing a certain message or transmitting certain information is to talk in circles, but often in useful circles for the time being. Content is a function of function, then, but not every structure can realize every function, can reliably guarantee the normal relationships required. So function is a function of structure. There are, then, strong indirect structural constraints on things that can be endowed with content. If our brains were as homogeneous as jelly we could not think. Fodor, however, makes a direct leap from content to structure and seems moreover to make structure in the end criterial for content.

On his view a prescriptive theory (e.g., natural deduction or decision theory) can be predictive of behaviour only if it is descriptive of inner processes. When we predict and explain the behaviour of a system at the Intentional level *our* calculations have a certain syntactic structure: to oversimplify, they are formal proofs or derivations, e.g., of descriptions of best actions to take given certain beliefs and preferences. We predict that the physical states or events to which we assign the premises as

1 Ignoring for the moment the normative element in all Intentional attributions. See my 'Intentional Systems', *J. Philosophy*, lxviii (1971), 87–105.

formulae will *cause* those states or events whose formulae are the later lines of our calculations (see, e.g., p. 73). Fodor seems to suppose that the only structures that could guarantee and explain the predictive power of our Intentionalistic calculations (and permit us to assign formulae to states or events in a principled way) *must* mirror the syntax of those calculations. This is either trivially true (because the 'syntactic' structure of events or states is defined simply by their function) or an empirical claim that is very interesting, not entirely implausible, and as yet not demonstrated or even argued for, so far as I can tell. For instance, suppose hamsters are interpretable as good Bayesians when it comes to the decisions they make. Must we in principle be able to find some saliencies in the hamsters' controls that are interpretable as tokens of formulae in some Bayesian calculus? If that is Fodor's conclusion I don't see that he has given it the support it needs, and I confess to disbelieving it utterly.

In a recent conversation with the designer of a chess-playing programme I heard the following criticism of a rival programme: 'It thinks it should get its queen out early'. This ascribes a propositional attitude to the programme in a very useful and predictive way, for as the designer went on to say, one can usually count on chasing that queen around the board. But for all the many levels of explicit representation to be found in that programme, nowhere is anything roughly synonymous with 'I should get my queen out early' explicitly tokened. The level of analysis to which the designer's remark belongs describes features of the programme that are, in an entirely innocent way, emergent properties of the computational processes that have 'engineering reality'. I see no reason to believe that the relation between belief-talk and psychological-process talk will be any more direct.

Are all these doubts about Fodor's radical view swept away by the material in the second half of this book, where evidence is adduced about the structure, vocabulary and utilization of the inner code? The challenge of these chapters to the sceptic is to find a way of recasting what cannot be denied in them in terms less radical than Fodor's. I do not see that this cannot be done, but saying it is not doing it, and doing it would require a monograph. Fodor's account of the inner code in action is packed with detail and bold speculation, and is supported by a variety of elegant experiments and ingenious arguments. Fodor puts together a more or less Gricean theory of communication and a more or less Chomskyan view of the relation between surface features of utterances and deeper levels, but comes out forcefully against semantic primitives (at least in their familiar role in the production and comprehension of sentences). He defends images as inner representational vehicles in addition to his code formulae, and claims to show that the inner code can represent its own representations and has a vocabulary about as rich as that of English—to mention a few highlights. There are a few dubious links in the argumentation (e.g., Fodor's cat and mouse example on p. 142 seems obviously mis-analysed, but it may not matter), but time and again Fodor succeeds, in my estimation, in parrying the 'obvious' philosophical

objections. One exception is in his account of communication. Unless I am reading him too literally, he seems committed to the view that for A to communicate verbally with B, A and B must not only share a natural language but have the same version of mentalese as well. Once again this is a claim that might be trivial or might be almost certainly false, and we can't tell until Fodor is more explicit.

Faulting Fodor for not being sufficiently explicit in this instance is a bit ungenerous, for Fodor has offered a detailed theory in an area hitherto bereft of detailed theories, and has been more explicit than anybody else about many of the murky issues. The book is exceptionally clear, with excellent summaries of arguments and conclusions at just the right places. The view Fodor has put forward is a remarkably full view; seldom have stands on so many different issues been so staunchly taken in this area, and even where I think he is wrong, it is usually the crispness of his expression that suggests for the first time just exactly what is wrong. Fodor challenges us to find a better theory, and I fully expect that challenge to be met, but when better theories emerge they will owe a good deal to Fodor's reconnaissance.[1]

TUFTS UNIVERSITY

1 I am indebted to Georges Rey, Bosse Dahlbom, David Israel and Susan Stafford for criticism of the first draft of this review.

PRINTED IN GREAT BRITAIN AT THE UNIVERSITY PRESS ABERDEEN