RECENT WORK IN THE PHILOSOPHY OF MIND

Philosophy of mind, a hybrid of philosophy of psychology, philosophy of language, metaphysics and epistemology, has come of age. Even with lingering unclarity in the notion of mental representation, many take the field to be on the threshold of genuine progress. Whether the hopes are frustrated or fulfilled, there is little doubt that the questions on centre stage are fundamental ones: As responsible thinkers informed by developments of science, how are we to understand mental phenomena? What kinds of states will be invoked by a scientific psychology, which aims at causal explanation of behaviour?

Despite the widespread optimism about the prospects for a comprehensive science of the mind, there is less than universal agreement about what form it will assume. As a result, philosophy of mind is tendentious and the battles sometimes bitter. Although I have joined the fray elsewhere (in Saving Belief: A Critique of Physicalism, Princeton, 1987), I think that readers of this brief survey of somewhat technical work will be served better by a more descriptive essay that sets out issues and suggests problems, but aims primarily to be a bibliographic guide. With the bevy of noteworthy books published in the past couple of years, however, I am forced to be selective.

Folk Psychology

A major question, around which much recent work has coalesced, concerns the fate of propositional attitudes like belief, desire and intention. Attitudes identified by 'that'-clauses are constitutive of our 'folk psychology' — the commonsense framework in terms of which we explain and predict each other's behaviour. Philosophers divide over whether or not (portions of) folk psychology can be made scientifically respectable — over whether or not, as they put it, the attitudes can be 'naturalised.' Those who think so try to provide a naturalistic semantics for the attitudes — that is, an account of the conditions under which a state is one of believing that snow is white, for example. For the account to be naturalistic in the intended sense, such conditions must be specified without using 'that'-clauses and without using explicitly semantic terms like 'refers to'.

Vol. XXX No. 1 January 1989
ISSN 0031—8051 $1.00
What makes the propositional attitudes so intractable for philosophers trying to naturalise them is their intentionality, the fact that they purport to be about or to represent states of affairs, whether those states of affairs obtain or not. Philosophers of mind have attempted to show, in naturally respectable terms, how an internal state can represent (or misrepresent) a given state of affairs, such as water’s being wet.


Second are those who conclude that there is something amiss in the naturalisation project itself and thus that failure to specify nonintentional and nonsemantic conditions for attitudes like belief is no warrant to draw radical conclusions about our commonsense framework. See Hilary Putnam, *Representation and Reality* (MIT/Bradford, 1988); Tyler Burge, ‘Individualism and Psychology’, *The Philosophical Review* Volume 95 (1986); and Stephen Schiffer, *Remnants of Meaning* (MIT/Bradford, 1987).

Another philosopher of mind and language who opts out of the naturalisation project is Donald Davidson. According to Davidson’s ‘anomalous monism’, all phenomena are physical; mental phenomena are mental by virtue of their having mentalistic descriptions. Although there will be no scientific laws that employ such descriptions, mental phenomena will be explained under their physical descriptions. Davidson’s development of these views is older, but it remains influential. See the two volumes of his collected papers, *Essays on Actions and Events* (Clarendon, 1980) and *Inquiries into Truth and Interpretation* (Clarendon, 1984), as well as the more recent *Essays on Davidson: Action and Events*, edited by Bruce Vermazen and Merril B. Hintikka (Clarendon, 1985) and *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*, edited by Ernest Le Pore (Basil Blackwell, 1986).

### Functionalism

For the past decade or two, the dominant approach to naturalising the mind has been functionalism. Functionalism is the view that mental states are functional states, where functional states are defined by their causal connections to inputs, outputs and other mental states. According to functionalism, human mental states are computational states of a brain.


Most functionalists, unlike Stich, also endorse the hypothesis of a full-fledged language of thought. Perhaps we have syntactically-structured, semantically interpreted internal representations realised in our brains. For example: a belief that snow is white may be understood as a relation between a believer and the proposition (or sentence) ‘Snow is white’. What makes something a belief, as opposed to a desire or an intention, is that it has a certain computational/causal role in the mental economy of the believer; what makes it a belief that such-and-such, as opposed to a belief that so-and-so, is not a matter of internal role at all, but rather it is a matter of relations between the believer and the environment.

An important work that takes this line is Jerry A. Fodor’s *Psychosemantics*. On Fodor’s representational theory, psychological states not only are causally efficacious for behaviour, but they also have semantic or representational properties expressed by ‘that’-clauses. By postulating states with both causal and semantic properties, Fodor aims to vindicate our folk psychology.

However, Fodor takes individualism — the view that states are individuated with respect to their local causal powers — to be a general presupposition of science, and his commitment to individualism forces him to distinguish between ‘wide’ and ‘narrow’ content. The wide content of a state, which is expressed by a ‘that’-clause, is determined by the subject’s relation to the environment; narrow content, which is inexpressible in public language, depends only on the subject’s internal states considered in isolation from the environment. The notion of narrow content as determined wholly by the subject’s internal states thus is taken...
o be suitable for incorporation into scientific psychology.

But what, one may wonder, is narrow content? Fodor’s answer is that narrow content is a function from the subject's environment or ‘context’ to truth condition, where context must be construed nonsemantically and unintentionally as much as the account is to be naturalistic one. In the absence of such specification of what is to count as a context, however, one may suspect a vicious circle: Narrow content is a function from context to truth condition; context is whatever is relevant to determining truth condition, given narrow content.

Perhaps the most persistent critic of individualism in psychology has been Tyler Burge, whose ‘Individualism and the Mental’, in Studies in Metaphysics, Midwest Studies in Philosophy, Volume IV, edited by Peter A. French et al., (Minneapolis, 1979) has become a classic. In ‘Individualism and Psychology’, Burge argues that psychology as practised does not conform to the constraints of individualism; nor is there good reason to revise the language of psychology so that it would be individualistic. Also see Robert van Gulick’s ‘Metaphysical Arguments for Internalism and Why They Don’t Work’ in Re-Representations: Readings in the philosophy of Mental Representation, edited by Stuart Silvers (Kluwer Academic Publishers, forthcoming).

William G. Lycan develops the idea of a language of thought somewhat differently from Fodor. In Judgement and Justification (Cambridge UP, 1988) and in Logical Form in Natural Language (MIT/Bradford, 1984), he combines the hypothesis of sentences-in-the-head with a Davidsonian truth-conditional semantics. Since Lycan is not committed to individualism, he is not saddled with the idea of narrow content. He looks toward a psychology that recognizes wide semantic properties like reference, which, he argues, are needed in certain psychological explanations (for example, of learning from authority). This raises interesting questions about the character of psychological explanation: Are explanations by semantic properties causal explanations? If so, the behaviour of physical duplicates could have different causal explanations. If not, we would have to countenance noncausal explanations of behaviour in scientific psychology.

Criticism of functionalism comes from several quarters. Hilary Putnam, one of the originators of functionalism as a model of mind, has been a critic of functionalism for the past thirteen years. In his most sustained attack, Representation and Reality, he argues that meaning and reference cannot be understood in functionalist terms.

Others, notably John Searle, object to the supposition that mental states may be understood in wholly computational terms. In Minds, Brains and Science (Harvard, 1984), Searle offers a lively critique of what he calls 'strong AI' in the form of a thought experiment. Strong AI endorses the analogy that mind is to brain as software is to hardware. Searle's own view (in Intentionality, Cambridge, 1983) is that the brain has causal powers that no mere computational state can have. Both the critique and the positive view remain controversial.

Hubert Dreyfus, a longstanding critic of strong AI, has recently co-authored (with Stuart Dreyfus) a book for a popular audience, Mind Over Machine: The Power of Human Intuition and Expertise in the Era of the Computer (The Free Press, 1986). In the tradition of Heidegger, Merleau-Ponty, and Wittgenstein, Dreyfus emphasises skills and 'know-how' over propositional knowledge and argues that symbol-manipulating models of the mind are inadequate. Although philosophers in the 'analytic' mode, which dominates philosophy of mind, should take Dreyfus more seriously, this breezy book may not be the best vehicle for a serious critique. Dreyfus's earlier What Computers Can't Do: A Critique of Artificial Reason, Revised Edition (Harper & Row, 1979) remains the most comprehensive critical investigation of the claims of strong AI.

The New Connectionism and Artificial Intelligence

So far, the issues I have been discussing arise within the traditional computational approach to the mind, dubbed Good Old-Fashioned Artificial Intelligence (GOFAI) by John Haugeland in his useful Artificial Intelligence: The Very Idea (MIT/Bradford Press, 1985). Recently, that standard approach has been challenged by the New Connectionism, so-called because it is a revival and development of an earlier idea in AI, the idea that intelligence could be modelled by neural networks rather than by explicit symbol systems. Each node of a net is connected to numerous other nodes from which it receives signals, which activate it to a certain level. When a node is activated above a certain threshold, it then propagates activation along the links to other nodes.

The revised edition of Paul M. Churchland's Matter and Consciousness (MIT/Bradford, 1988) has a good introduction to Connectionism, which is also known as Parallel Distributed Processing (PDP). Also Daedalus: Journal of the American Academy of Arts and Sciences, whose Winter 1988 issue is devoted exclusively to artificial intelligence, contains a good introductory survey of Connectionism for a general audience.

Some PDP models are more congenial to behaviourism than to mentalism. Instead of being programmed with a set of rules, Connectionist models of learning are 'trained up' by a learning process that consists of a number of repetitions of a training cycle, with a feedback mechanism (back-propagation) to signal whether the response was correct and, if not, to adjust the strengths of activation among the units. PDP models are also more like holistic approaches to information than like the traditional models employing units at specific memory addresses; information is stored in patterns of weighted connections produced by learning. (This accounts for the 'distributed' in PDP, 'parallel', contrasted with 'serial', refers to the fact that numerous processes are going on at the same time.)

There are many sources of the appeal of the New Connectionism. First, it seems especially useful in describing the information-processing that utilises many sources of information interacting with one another — for instance, when we recognise a word from an ambiguous display in which the letters, considered separately, are not identifiable, but in which the visual information of each letter constrains the identities of the others. In
addition, PDP models seem to be physiologically plausible since the brain is itself a large number of interconnected units, which are connected by the excitatory and inhibitory signals that they send to each other. Finally, whereas the rival sequential models of human thought would require a much faster brain than we have, the massive parallel structure of connectionist architecture may be able to overcome the time limitations of serial models.

For descriptions of PDP models and much more, see Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volumes 1 and 2, edited by David E. Rumelhart, James L. McClelland, and the PDP Research Group, (MIT/Bradford Books, 1986). These volumes have become the standard work, which anyone interested in Connectionism should consult. In addition, Neurophilosophy by Patricia Churchland, as well as articles in Modularity in Knowledge Representation and Natural Language Understanding, edited by Jay L. Garfield, (MIT/Bradford, 1987), look to PDP for philosophical significance.

Optimism about PDF models is not universal, however. One early hurdle concerns the relation between the unstructured units postulated by New Connectionist models and the representational forms in which we express our knowledge. In particular, supposing that (at least sometimes) we express our thoughts by uttering sentences, how is the generativity of natural languages accommodated by connectionist models? This kind of criticism is developed by Jerry A. Fodor and Zenon W. Pylyshyn in ‘Connectionism and Cognitive Architecture: A Critical Analysis’ in Cognition, volume 28 (1988), 2—71. For a connectionist’s response, see Paul Smolensky, ‘The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn’ in Connectionism and the Philosophy of Mind, edited by Terence Horgan and John Tienson (Volume 26, Supplement to The Southern Journal of Philosophy, published by Memphis State University, 1988). This volume also contains useful introductory surveys aimed specifically at philosophers.

It is too soon to predict the fate of Connectionism, or even, if it is successful, to predict its relevance to issues that drive people into the philosophy of mind in the first place. The results of science do not wear their philosophical implications on their sleeves. Assuming, as most philosophers of mind do, that philosophy should be continuous with science in some sense, there will still be philosophical work to be done no matter what the fate of Connectionism. In particular, we will still have the problem of understanding intentionality, or of showing how we can do without it.

**Reliable Indication**

An approach to naturalising intentionality that is neutral with regard to the existence of a language of thought treats belief as a kind of indication. Just as the pointer of the fuel gauge is at a certain position, in normal conditions, only if the tank has a certain amount of gas in it, so too, one believes that such-and-such, in normal conditions, only if one is in a state that is a reliable indicator that such-and-such is the case. For example, suppose that Jones believes that the toast is burning. Then she is in a state that, in normal conditions, she is in only if the toast is burning, and, in normal conditions, the toast’s burning causally contributes to her being in that state.

This line has been pursued by Fred I. Dretske in Explaining Behavior: Reasons in a World of Causes (MIT/Bradford, 1988), Robert C. Stalnaker in Inquiry (MIT/Bradford, 1984), and Fodor in Psychosemantics. Dan Lloyd’s Simple Minds (MIT/Bradford, forthcoming) will develop this approach still further.

Let me mention just two obstacles that must be overcome by anyone using the idea of reliable indication to naturalise intentionality. First, one must give some reason to believe that normal conditions can be specified nonintentionally and nonsemantically. For example, Stalnaker, who takes the intentionality of mental states as independent of the intentionality of linguistic expressions, accounts for Burge’s famous ‘arthritis’ case by saying that “the fidelity conditions which are relevant to beliefs formed by a process that involves language may depend on the linguistic capacities that are normal in the believer’s linguistic community” (Inquiry, p. 67). Is there any reason to suppose that the relevant linguistic capacities can be specified nonsemantically and nonintentionally? If not, then the account is not suitably naturalistic.

Second, there is the difficulty of accounting for misrepresentation. Dretske has made several efforts (as has Fodor) to account for misrepresentation in naturalistic terms. Dretske’s approach depends upon there being natural functions, “functions a thing has which are independent of our interpretative intentions and purposes” (Misrepresentation in Belief: Form, Content and Function, edited by Radu J. Bogdan, Clarendon Press, 1986). If its natural function determines what a sign is supposed to do, then error or misrepresentation can be understood as failure to do what the sign is designed to do. A difficulty with this approach, one recognised by Dretske, is the indeterminacy of function: what principled way do we have for characterising functions? Moreover, naturalisation of intentionality along these lines would seem to require naturalisation of teleology as a further project.

**Evolutionary Instrumentalism**

One way simply to by-pass the difficulties of the naturalisation project is to be an instrumentalist about intentional states. Daniel C. Dennett has espoused just such a selective instrumentalism — instrumentalism concerning beliefs and other phenomena characterised from the ‘intentional stance’ on the one hand, realism concerning underlying physical mechanisms on the other. (For some new arguments against instrumentalism in psychology, see Jay L. Garfield, Belief in Psychology (MIT/Bradford, 1988).)

Dennett now considers it a tactical error to have called himself an instrumentalist: beliefs have the same sort of reality as centres of gravity.
Dennett insists (no doubt, correctly) that predictively successful characterisations from the intentional stance may well shed no light on the underlying mechanisms that actually produce the behaviour. But Dennett goes much farther. For one thing, he suggests that to think that there “are facts of the matter” about intentional states is thereby to resist the “idea that we are artifacts designed by natural selection” (p. 300). Such a startling suggestion needs more careful formulation and argument than Dennett gives it.

Dennett's view may be summarised by saying that all ascriptions of content are interpretive (in some unexplored sense in which there are limits to their precision, and in which interpretations contrast with descriptions of unadorned facts discovered by physical sciences), and that there is no original or underived intentionality: all intentionality is derived from the intentions and purposes either of an ascriber or of Mother Nature. Dennett credits Mother Nature with “free-floating rationales” — rationales that are never expressed or represented, ‘as if’ rationales which are “predictive and, hence, explanatory” (p. 259). Also see Ruth Garrett Millikan's Language, Thought and Other Biological Categories.

The question arises: what does it mean to suppose that Mother Nature has intentions and purposes, or that “it is only relative to [Mother Nature's] design ‘choices’ or evolution-‘endorsed’ purposes — raisons d'être — that we can identify behaviors, actions, perceptions, beliefs, or any of the other categories of folk psychology”? (p. 300).

First, it is difficult to see how most cases of intentional ascription could plausibly or nontrivially be related to evolutionary considerations at all. (For instance, ‘She bid two-no-trumps because she thought that her partner had two aces.’) Second, even in cases in which evolutionary considerations may be relevant (for instance, ‘He shot the intruder because he thought his life was in danger’), there seems to be a dilemma lurking. Appeal to the intentions of Mother Nature is either itself relative to our intentions or not. If it is not, then Mother Nature has intrinsic, intentionality and the view loses its naturalistic credentials. If Mother Nature’s intentions are relative to our intentions, then it is difficult to see how appeal to Mother Nature’s intentions can explain our intentions in any sense at all. To say in this context that our intentionality “is derived from the intentionality of natural selection” is to walk a small circle.

Dennett has most recently developed his position in The Intentional Stance (MIT/Bradford, 1987) and tangentially in Elbow Room: The Varieties of Free Will Worth Wanting (MIT/Bradford, 1984). The Intentional Stance reprints six previously published essays with newly appended ‘Reflections’ and adds four new papers. Earlier critics who thought that Dennett’s instrumentalism was unstable may not be altogether reassured by his latest refinements; still, The Intentional Stance is an important (and readable) addition to the literature.

Other Topics

There are many other important contributions to philosophy of mind that I have been unable to discuss here. The following is just a sample: First, there have been developments in the study of sense experience and of subjectivity generally. In Sense and Content: Experience, Thought and Their Relations (Oxford, 1983), Christopher Peacocke endorses a Russellian Principle of Acquaintance. In The Subjective View: Secondary Qualities and Indexical Thoughts (Clarendon, 1983), Colin McGinn argues that the world is and must be presented to us in a subjective way. Perhaps the most influential book on the relations between the objective and on the eliminability of subjectivity is Thomas Nagel's The View From Nowhere (Oxford UP, 1986).

Second, there has been work in related areas that bears on topics in philosophy of mind: For example, most philosophers of mind share the view that mental states ‘supervene’ on physical states in that, crudely, there can be no difference in mental state without some difference in physical state; so, work on supervenience has immediate application to the philosophy of mind. (See, for example, Jaegwon Kim’s ‘Epiphenomenal and Supervenient Causation’ in Causation and Causal Theories (Midwest Studies in Philosophy, Volume 9), edited by Peter French et al., Minnesota, 1984.) Another example is Saul Kripke's influential article, ‘A Puzzle About Belief’ (in Meaning and Use, edited by Avishai Margalit (Reidel, 1979)), which has prompted much subsequent work on ascription of belief. A final example is Situations and Attitudes by Jon Barwise and John Perry (MIT/Bradford, 1983); in the course of an informal presentation of a theory of sentence meaning, Barwise and Perry develop a view of propositional attitudes.


NOTES

1. I put 'naturalised' in scare quotes because it is used in this context in a restricted sense, according to which some entity (or property) is naturalised if nonintentional and nonsemantic conditions for its existence (or instantiation) can be specified. An account that specifies such conditions is called 'naturalistic' in this special sense. An explicit statement of the need for naturalisation in this sense may be found in Fodor's Psycosemantics, p. 97.
transcendent statements may be held to consist in a grasp of their truth-conditions.

Why, then, should the idea that truth is objective be thought to be objectionable? According to the 'acquisition argument', the difficulty is that we have no idea how we are supposed to have acquired the notion that a statement might be true in virtue of the obtaining of a state of affairs which cannot be recognised to obtain. Wright is not inclined to make very much of this point, for the argument invites the response that it is in acquiring a language that we acquire a conception of unrecognisable states of affairs in the obtaining of which the truth of verification-transcendent statements consists. The role of the so-called 'manifestation argument' is to undermine this picture of statement-understanding and the notion of truth which underpins it. The manifestation argument challenges the realist to explain "what in our use of any particular statement would distinctively manifest that we understood what it was for that statement to be true in a manner transcending our capacities for verification" (p. 86). If nothing in S's use of language manifests his grasp of the concept of objective truth, then there will be no grounds for attributing the concept to S (barring a successful transcendental argument for such an attribution). If there are no grounds for attributing to S an understanding of the notion of objective truth, then, by the same token, there will be no grounds for supposing that S's capacity to understand given statements consists in his knowledge of what it would be for them to be true, where the statements are such that their truth-conditions could not be recognised to obtain. To be sure, there is a relatively harmless and platitudinous sense in which statement-understanding consists in knowledge of truth-conditions, but the platitudes which warrant such a description "do nothing to justify the idea that the notion of truth which the reference therein to 'truth-conditions' invokes is the realist's objective truth" (p. 19).

In order to assess this argument, it would be as well to distinguish several questions which might be asked about the concept of objective truth. The first has to do with its legitimacy or intelligibility: is it actually intelligible to suppose that a statement might be true even if it is impossible that we should ever come to know that it is true? Supposing that the concept is an intelligible one, a second question concerns what it is for a given speaker to have a grasp of it; what does S's understanding of the objectivity of truth consist in? A third question is concerned with the constraints upon attributing the concept to S: what would count as evidence that S does indeed understand the notion? Although these questions are related, they are not, on the face of it, different ways of asking one and the same question. Wright is rather inclined to slide over the difference between the second and third questions, for he insists that "more ought not to be read into someone's understanding of an expression than can be manifested in his behaviour" (p. 20). Wright also seems to hold that the point about manifestation holds the key to the first question, for the question in response to which the manifestation argument is first introduced into the discussion is, precisely, a question about the