Three kinds of intentional psychology[1]

D. C. DENNETT

1

Suppose you and I both believe that cats eat fish. Exactly what feature must we share for this to be true of us? More generally, recalling Socrates' favourite style of question, what must be in common between things truly ascribed an *intentional* predicate – such as 'wants to visit China' or 'expects noodles for supper'?[2] As Socrates points out, in the *Meno* and elsewhere, such questions are ambiguous or vague in their intent. One can be asking on the one hand for something rather like a definition, or on the other hand for something rather like a theory. (Socrates of course preferred the former sort of answer.) What do all magnets have in common? First answer: they all attract iron. Second answer: they all have such-and-such a microphysical property (a property that explains their capacity to attract iron). In one sense people knew what magnets were – they were things that attracted iron – long before science told them what magnets were. A child learns what the word 'magnet' means not, typically, by learning an explicit definition, but by learning the 'folk physics' of magnets, in which the ordinary term 'magnet' is embedded or implicitly defined as a theoretical term.[3]

Sometimes terms are embedded in more powerful theories, and sometimes they are embedded by explicit definition. What do all chemical elements with the same valence have in common? First answer: they are disposed to combine with other elements in the same integral ratios.

[1] I am grateful to the Thyssen Philosophy Group, the Bristol Fulbright Workshop, Elliot Sober and Bo Dahlbom for extensive comments and suggestions on an earlier draft of this paper.

[2] Other 'mental' predicates, especially those invoking episodic and allegedly *qualia*-laden entities – pains, sensations, images – raise complications of their own which I will not consider here, for I have dealt with them at length elsewhere, especially in *Brainstorms* (1978). I will concentrate here on the foundational concepts of belief and desire, and will often speak just of belief, implying, except where I note it, that parallel considerations apply to desire.

[3] The child need learn only a portion of this folk physics, as Putnam argues in his discussion of the 'division of linguistic labour' (1975).

Second answer: they all have such-and-such a microphysical property (a property which explains their capacity so to combine). The theory of valences in chemistry was well in hand before its microphysical explanation was known. In one sense chemists knew what valences were before physicists told them.

So what appears in Plato to be a contrast between giving a definition and giving a theory can be viewed as just a special case of the contrast between giving one theoretical answer and giving another, more 'reductive' theoretical answer. Fodor (1975) draws the same contrast between 'conceptual' and 'causal' answers to such questions, and argues that Ryle (1949) champions conceptual answers at the expense of causal answers, wrongly supposing them to be in conflict. There is justice in Fodor's charge against Ryle, for there are certainly many passages in which Ryle seems to propose his conceptual answers as a bulwark against the possibility of *any* causal, scientific, psychological answers, but there is a better view of Ryle's (or perhaps at best a view he ought to have held) that deserves rehabilitation. Ryle's 'logical behaviourism' is composed of his steadfastly conceptual answers to the Socratic questions about matters mental. If Ryle thought these answers ruled out psychology, ruled out causal (or reductive) answers to the Socratic questions, he was wrong, but if he thought only that the conceptual answers to the questions were not to be given by a microreductive psychology, he was on firmer ground. It is one thing to give a causal explanation of some phenomenon and quite another to cite the cause of a phenomenon in the analysis of the concept of it.

Some concepts have what might be called an essential causal element.[4] For instance, the concept of a genuine Winston Chruchill *autograph* has it that how the trail of ink was in fact caused is essential to its status as an autograph. Photocopies, forgeries, inadvertently indistinguishable signatures – but perhaps not carbon copies – are ruled out. These considerations are part of the *conceptual* answer to the Socratic question about autographs.

Now some, including Fodor, have held that such concepts as the concept of intelligent action also have an essential causal element; behaviour that appeared to be intelligent might be shown not to be by being shown to have the wrong sort of cause. Against such positions Ryle can argue that even if it is true that every instance of intelligent behaviour is caused (and hence has a causal explanation), exactly *how* it is caused is inessential to its being intelligent – something that could be true even if all intelligent behaviour exhibited in fact some common pattern of causation.

[4] Cf. Fodor 1975: 7n.

That is, Ryle can plausibly claim that no account in causal terms could capture the class of intelligent actions except *per accidens*. In aid of such a position – for which there is much to be said in spite of the current infatuation with causal theories – Ryle can make claims of the sort Fodor disparages ('it's not the mental activity that makes the clowning clever because what makes the clowning clever is such facts as that it took place out where the children can see it') without committing the error of supposing causal and conceptual answers are incompatible.[5]

Ryle's logical behaviourism was in fact tainted by a groundless anti-scientific bias, but it need not have been. Note that the introduction of the concept of valence in chemistry was a bit of *logical chemical behaviourism:* to have valence *n* was 'by definition' to be disposed to behave in such-and-such ways under such-and-such conditions, *however* that disposition to behave might someday be explained by physics. In this particular instance the relation between the chemical theory and the physical theory is now well charted and understood – even if in the throes of ideology people sometimes misdescribe it – and the explanation of those dispositional combinatorial properties by physics is a prime example of the sort of success in science that inspires reductionist doctrines. Chemistry has been shown to reduce, in some sense, to physics, and this is clearly a Good Thing, the sort of thing we should try for more of.

Such progress invites the prospect of a parallel development in psychology. First we will answer the question 'What do all believers-that-*p* have in common?' the first way, the 'conceptual' way, and then see if we can go on to 'reduce' the theory that emerges in our first answer to something else – neurophysiology most likely. Many theorists seem to take it for granted that *some* such reduction is both possible and desirable, and perhaps even inevitable, even while recent critics of reductionism, such as Putnam and Fodor, have warned us of the excesses of 'classical' reductionist creeds. No one today hopes to conduct the psychology of the future in the vocabulary of the neurophysiologist, let alone that of the physicist, and principled ways of relaxing the classical 'rules' of reduction have been proposed. The issue, then, is *what kind* of theoretical bonds can we expect – or ought we to hope – to find uniting psychological claims about beliefs, desires, and so forth with the claims of neurophysiologists, biologists and other physical scientists?

Since the terms 'belief' and 'desire' and their kin are parts of ordinary language, like 'magnet', rather than technical terms like 'valence', we must first look to 'folk psychology' to see what kind of things we are being asked to explain. *What do we learn beliefs are when we learn how to use the words*

[5] This paragraph corrects a misrepresentation of both Fodor's and Ryle's positions in my critical notice of Fodor's book in *Mind*, 1977, reprinted in *Brainstorms*, pp. 90–108.

'believe' and 'belief'? The first point to make is that we do not really learn what beliefs are when we learn how to use these words.[6] Certainly no one *tells us* what beliefs are, or if someone does, or if we happen to speculate on the topic on our own, the answer we come to, wise or foolish, will figure only weakly in our habits of thought about what people believe. We learn to *use* folk psychology – as a vernacular social technology, a craft – but we don't learn it self-consciously as a theory – we learn no meta-theory with the theory – and in this regard our knowledge of folk psychology is like our knowledge of the grammar of our native tongue. This fact does not make our knowledge of folk psychology entirely unlike human knowledge of explicit academic theories, however; one could probably be a good practising chemist and yet find it embarrassingly difficult to produce a satisfactory textbook definition of a metal or an ion.

There are no introductory textbooks of folk psychology (although Ryle's *The Concept of Mind* might be pressed into service), but many explorations of the field have been undertaken by ordinary language philosophers (under slightly different intentions), and more recently by more theoretically minded philosophers of mind, and from all this work an account of folk psychology – part truism and the rest controversy – can be gleaned. What are beliefs? *Roughly*, folk psychology has it that *beliefs* are information-bearing states of people that arise from perceptions, and which, together with appropriately related *desires*, lead to intelligent *action*. That much is relatively uncontroversial, but does folk psychology also have it that non-human animals have beliefs? If so, what is the role of language in belief? Are beliefs constructed of parts? If so, what are the parts? Ideas? Concepts? Words? Pictures? Are beliefs like speech acts or maps or instruction manuals or sentences? Is it implicit in folk psychology that beliefs enter into causal relations, or that they don't? How do decisions and intentions intervene between belief–desire complexes and actions? Are beliefs introspectible, and if so, what authority do the believer's pronouncements have?

All these questions deserve answers, but one must bear in mind that there are different reasons for being interested in the details of folk psychology. One reason is that it exists as a phenomenon, like a religion or a language or a dress code, to be studied with the techniques and attitudes of anthropology. It may be a myth, but it is a myth we live in, so it is an 'important' phenomenon in nature. A different reason is that it seems to be a *true* theory, by and large, and hence is a candidate – like the folk physics of magnets and unlike the folk science of astrology – for

[6] I think it is just worth noting that philosophers' use of 'believe' as the standard and general ordinary language term is a considerable distortion. We *seldom* talk about what people *believe*; we talk about what they *think* and what they *know*.

incorporation into science. These different reasons generate different but overlapping investigations. The anthropological question should include in its account of folk psychology whatever folk actually include in their theory, however misguided, incoherent, gratuitous some of it may be.[7] The proto-scientific quest, on the other hand, as an attempt to prepare folk theory for subsequent incorporation into or reduction to the rest of science, should be critical, and should *eliminate* all that is false or ill-founded, however well-entrenched in popular doctrine. (Thales thought that lodestones had souls, we are told. Even if most people agreed, this would be something to eliminate from the folk physics of magnets prior to 'reduction'.) One way of distinguishing the good from the bad, the essential from the gratuitous, in folk theory is to see what must be included in the theory to account for whatever predictive or explanatory success it seems to have in ordinary use. In this way we can criticize as we analyse, and it is even open to us in the end to discard folk psychology if it turns out to be a bad theory, and with it the presumed theoretical entities named therein. If we discard folk psychology as a theory, we would have to replace it with another theory, which while it did violence to many ordinary intuitions would explain the predictive power of the residual folk craft.

We use folk psychology all the time, to explain and predict each other's behaviour; we attribute beliefs and desires to each other with confidence – and quite unself-consciously – and spend a substantial portion of our waking lives formulating the world – not excluding ourselves – in these terms. Folk psychology is about as pervasive a part of our second nature as is our folk physics of middle-sized objects. How good is folk psychology? If we concentrate on its weaknesses we will notice that we often are unable to make sense of particular bits of human behaviour (our own included) in terms of belief and desire, even in retrospect; we often cannot predict accurately or reliably what a person will do or when; we often can find no resources within the theory for settling disagreements about particular attributions of belief or desire. If we concentrate on its strengths we find first that there are large areas in which it is extraordinarily reliable in its predictive power. Every time we venture out on a highway, for example, we stake our lives on the reliability of our general expectations about the perceptual beliefs, normal desires and decision proclivities of the other motorists. Second, we find that it is a theory of great generative power and efficiency. For instance, watching a film with a highly original and

[7] If the anthropologist marks part of the catalogue of folk theory as false, as an inaccurate or unsound account of the folk craft, he may speak of *false consciousness* or *ideology*; the role of such false theory in constituting a feature of the anthropological phenomenon is not diminished by its falseness.

unstereotypical plot, we see the hero smile at the villain and we all swiftly and effortlessly arrive at the same complex theoretical diagnosis: 'Aha!' we conclude (but perhaps not consciously), 'he wants her to think he doesn't know she intends to defraud his brother!' Third, we find that even small children pick up facility with the theory at a time when they have a very limited experience of human activity from which to induce a theory. Fourth, we find that we all use folk psychology knowing next to nothing about what actually happens inside people's skulls. 'Use your head' we are told, and we know some people are brainier than others, but our capacity to use folk psychology is quite unaffected by ignorance about brain processes – or even by large-scale misinformation about brain processes.

As many philosophers have observed, a feature of folk psychology that sets it apart from both folk physics and the academic physical sciences is the fact that explanations of actions citing beliefs and desires normally not only describe the provenance of the actions, but at the same time defend them as reasonable under the circumstances. They are reason-giving explanations, which make an ineliminable allusion to the rationality of the agent. Primarily for this reason, but also because of the pattern of strengths and weaknesses just described, I suggest that folk psychology might best be viewed as a rationalistic calculus of interpretation and prediction – an idealizing, abstract, instrumentalistic interpretation-method that has evolved because it works, and works because we have evolved. We approach each other as *intentional systems*,[8] that is, as entities whose behaviour can be predicted by the method of attributing beliefs, desires and rational acumen according to the following rough and ready principles:[9]

(1) A system's beliefs are those it *ought to have*, given its perceptual capacities, its epistemic needs, and its biography. Thus, in general, its beliefs are both true and relevant to its life, and when false beliefs are attributed, special stories must be told to explain how the error resulted from the presence of features in the environment that are deceptive relative to the perceptual capacities of the system.

(2) A system's desires are those it *ought to have*, given its biological needs and the most practicable means of satisfying them. Thus intentional systems desire survival and procreation, and hence desire food, security, health, sex, wealth, power, influence, and so forth, and also whatever local arrangements tend (in their eyes – given their

[8] See my 'Intentional Systems' (1971).

[9] For a more elaborate version of similar principles, see Lewis 1974.

beliefs) to further these ends in appropriate measure. Again, 'abnormal' desires are attributable if special stories can be told.

(3) A system's behaviour will consist of those acts that *it would be rational* for an agent with those beliefs and desires to perform.

In (1) and (2) 'ought to have' means 'would have if it were *ideally* ensconced in its environmental niche'. Thus all dangers and vicissitudes in its environment it will *recognize as such* (i.e. *believe* to be dangers) and all the benefits – relative to its needs, of course – it will *desire*. When a fact about its surroundings is particularly relevant to its current projects (which themselves will be the projects such a being ought to have in order to get ahead in its world) it will *know* that fact, and act accordingly. And so forth and so on. This gives us the notion of an ideal epistemic and conative operator or agent, relativized to a set of needs for survival and procreation and to the environment(s) in which its ancestors have evolved and to which it is adapted. But this notion is still too crude and overstated. For instance, a being may come to have an epistemic need that its perceptual apparatus cannot provide for (suddenly all the green food is poisonous but alas it is colourblind), hence the relativity to perceptual capacities. Moreover, it may or may not have had the occasion to learn from experience about something, so its beliefs are also relative to its biography in this way: it will have learned what it ought to have learned, *viz.* what it had been given evidence for in a form compatible with its cognitive apparatus – providing the evidence was 'relevant' to its project then.

But this is still too crude, for we understand that evolution does not give us a best of all possible worlds, but only a passable jury-rig, so we should look for design shortcuts that in specifiably abnormal circumstances yield false perceptual beliefs, etc. (We are not immune to illusions – which we would be if our perceptual systems were *perfect*.) To offset the design shortcuts we should also expect design bonuses: circumstances in which the 'cheap' way for nature to design a cognitive system has the side benefit of giving good, reliable results even outside the environment in which the system evolved. Our eyes are well adapted for giving us true beliefs on Mars as well as on Earth – because the cheap solution for our Earth-evolving eyes happens to be a more general solution.[10]

I propose that we can continue the mode of thinking just illustrated *all the way in* – not just for eye-design, but for deliberation-design and belief-design and strategy-concocter-design. In using this optimistic set of assumptions (nature has built us to do things right; look for systems to believe the truth and love the good) we impute no occult

[10] Cf. Sober (unpublished) for useful pioneering exploration of these topics.

powers to epistemic needs, perceptual capacities and biography, but only the powers common sense already imputes to evolution and learning.

In short, we treat each other as if we were rational agents, and this myth – for surely we are not all that rational – works very well because we are *pretty* rational. This single assumption, in combination with home truths about our needs, capacities and typical circumstances, generates both an intentional interpretation of us as believers and desirers and actual predictions of behaviour in great profusion. I am claiming, then, that folk psychology can best be viewed as a sort of logical behaviourism: *what it means* to say that someone believes that *p*, is that that person is disposed to behave in certain ways under certain conditions. What ways under what conditions? The ways it would be rational to behave, given the person's other beliefs and desires. The answer looks in danger of being circular, but consider: an account of what it is for an element to have a particular valence will similarly make ineliminable reference to the valences of other elements. What one is given with valence-talk is a whole system of interlocking attributions, which is saved from vacuity by yielding independently testable predictions.

I have just described in outline a *method* of predicting and explaining the behaviour of people and other intelligent creatures. Let me distinguish two questions about it: (1) is it something we could do and (2) is it something we in fact do? I think the answer to (1) is obviously yes, which is not to say the method will always yield good results. That much one can ascertain by reflection and thought experiment. Moreover, one can recognize that the method is familiar. Although we don't usually use the method self-consciously, we do use it self-consciously on those occasions when we are perplexed by a person's behaviour, and then it often yields satisfactory results. Moreover, the ease and naturalness with which we resort to this self-conscious and deliberate form of problem-solving provide some support for the claim that what we are doing on those occasions is not *switching methods* but simply becoming self-conscious and explicit about what we ordinarily accomplish tacitly or unconsciously.

No other view of folk psychology, I think, can explain the fact that we do so well predicting each other's behaviour on such slender and peripheral evidence; treating each other as intentional systems works (to the extent that it does) because we really are well designed by evolution and hence we *approximate* to the ideal version of ourselves exploited to yield the predictions. But not only does evolution not guarantee that we will always do what is rational; it guarantees that we won't. If we are designed by evolution, then we are almost certainly nothing more than a bag of tricks, patched together by a *satisficing*[11] Nature, and no better than

[11] The term is Herbert Simon's (e.g. 1969).

our ancestors had to be to get by. Moreover, the demands of nature and the demands of a logic course are not the same. Sometimes – even *normally* in certain circumstances – it pays to jump to conclusions swiftly (and even to forget that you've done so), so by most philosophical measures of rationality (logical consistency, refraining from invalid inference) there has probably been some positive evolutionary pressure in favour of 'irrational' methods.[12]

How rational are we? Recent research in social and cognitive psychology suggests we are *minimally* rational, appallingly ready to leap to conclusions or be swayed by logically irrelevant features of situations,[13] but this jaundiced view is an illusion engendered by the fact that these psychologists are deliberately trying to produce situations that provoke irrational responses – inducing pathology in a system by putting strain on it – and succeeding, being good psychologists. No one would hire a psychologist to prove that people will choose a paid vacation to a week in jail if offered an informed choice. At least not in the better psychology departments. A more optimistic impression of our rationality is engendered by a review of the difficulties encountered in artificial intelligence research. Even the most sophisticated AI programmes stumble blindly into misinterpretations and misunderstandings that even small children reliably evade without a second thought.[14] From this vantage point we seem marvellously rational.

However rational we are, it is the myth of our rational agenthood that structures and organizes our attributions of belief and desire to others, and that regulates our own deliberations and investigations. We aspire to rationality, and without the myth of our rationality the concepts of belief and desire would be uprooted. Folk psychology, then, is *idealized* in that it produces its predictions and explanations by calculating in a normative

[12] While in general true beliefs have to be more useful than false beliefs (and hence a system ought to have true beliefs), in special circumstances it may be better to have a few false beliefs. For instance it might be better for beast B to have some false beliefs about whom B can beat up and whom B can't. Ranking B's likely antagonists from ferocious to pushover, we certainly want B to believe it can't beat up all the ferocious ones, and can beat up all the obvious pushovers, but it is better (because it 'costs less' in discrimination tasks and protects against random perturbations such as bad days and lucky blows) for B to extend 'I can't beat up x' to cover even some beasts it can in fact beat up. *Erring on the side of prudence* is a well recognized good strategy, and so Nature can be expected to have valued it on occasion when it came up. An alternative strategy in this instance would be to abide by the rule: avoid conflict with penumbral cases. But one might have to 'pay more' to implement that strategy than to implement the strategy designed to produce, and rely on, some false beliefs.

[13] See, e.g. Tversky and Kahneman 1974; and Nisbett and Ross 1978.

[14] Roger Schank's (1977; Schank and Abelson 1977) efforts to get a computer to 'understand' simple but normally gappy stories is a good illustration.

system; it predicts what we *will* believe, desire, and do, by determining what we *ought* to believe, desire, and do.[15]

Folk psychology is *abstract* in that the beliefs and desires it attributes are not – or need not be – presumed to be intervening distinguishable states of an internal behaviour-causing system. (The point will be enlarged upon later.) The role of the concept of belief is like the role of the concept of a centre of gravity, and the calculations that yield the predictions are more like the calculations one performs with a parallelogram of forces than like the calculations one performs with a blueprint of internal levers and cogs.

Folk psychology is thus *instrumentalistic* in a way the most ardent realist should permit: people really do have beliefs and desires, on my version of folk psychology, just the way they really have centres of gravity and the earth has an Equator.[16] Reichenbach distinguished between two sorts of referents for theoretical terms: *illata* – posited theoretical entities – and *abstracta* – calculation-bound entities or logical constructs.[17] Beliefs and desires of folk psychology (but not all mental events and states) are *abstracta*.

This view of folk psychology emerges more clearly in contrast to a diametrically opposed view, each of whose tenets has been held by some philosopher, and at least most of which have been espoused by Fodor:

> Beliefs and desires, just like pains, thoughts, sensations and other episodes, are taken by folk psychology to be real, intervening, internal states or events, in causal interaction, subsumed under covering laws of causal stripe. Folk psychology is not an idealized, rationalistic calculus but a naturalistic, empirical, descriptive theory, imputing causal regularities discovered by extensive induction over experience. To suppose two people share a belief is to suppose them to be ultimately in some structurally similar internal condition, e.g. for them to have the same words of Mentalese written in the functionally relevant places in their brains.

I want to deflect this head-on collision of analyses by taking two steps. First, I am prepared to grant a measure of the claims made by the opposition. *Of course* we don't all sit in the dark in our studies like mad

[15] It tests its predictions in two ways: action predictions it tests directly by looking to see what the agent does; belief and desire predictions are tested indirectly by employing the predicted attributions in further predictions of eventual action. As usual, the Duhemian thesis holds: belief and desire attributions are under-determined by the available data.

[16] Michael Friedman's 'Theoretical Explanation' (in this volume) provides an excellent analysis of the role of instrumentalistic thinking within realistic science. Scheffler (1963) provides a useful distinction between *instrumentalism* and *fictionalism*. In his terms I am characterizing folk psychology as instrumentalistic, not fictionalistic.

[17] Reichenbach 1938: 211–12. 'Our observations of concrete things confer a certain probability on the existence of *illata* – nothing more . . . Second, there are inferences to *abstracta*. These inferences are . . . equivalences, not probability inferences. Consequently, the existence of abstracta is reducible to the existence of concreta. There is, therefore, no problem of their objective existence; their status depends on a convention.'

Leibnizians rationalistically excogitating behavioural predictions from pure, idealized concepts of our neighbours, nor do we derive all our readiness to attribute desires from a careful generation of them from the ultimate goal of survival. We may observe that some folks seem to desire cigarettes, or pain, or notoriety (we observe this by hearing them tell us, seeing what they choose, etc.) and without any conviction that these people, given their circumstances, ought to have these desires, we attribute them anyway. So rationalistic generation of attributions is augmented and even corrected on occasion by empirical generalizations about belief and desire that guide our attributions and are learned more or less inductively. For instance, small children believe in Santa Claus, people are inclined to believe the more self-serving of two interpretations of an event in which they are involved (unless they are depressed), and people can be made to want things they don't need by making them believe that glamorous people like those things. And so forth in familiar profusion. This folklore does not consist in *laws* – even probabilistic laws – but some of it is being turned into science of a sort, e.g. theories of 'hot cognition' and cognitive dissonance. I grant the existence of all this naturalistic generalization, and its role in the normal calculations of folk psychologists – i.e. all of us. People do rely on their own parochial group of neighbours when framing intentional interpretations. That is why people have so much difficulty understanding foreigners – their behaviour, to say nothing of their languages. They impute more of their own beliefs and desires, and those of their neighbours, than they would if they followed my principles of attribution slavishly. Of course this is a perfectly reasonable shortcut for people to take, even when it often leads to bad results. We are in this matter, as in most, satisficers, not optimizers, when it comes to information gathering and theory construction. I would insist, however, that all this empirically obtained lore is laid over a fundamental generative and normative framework that has the features I have described.

My second step away from the conflict I have set up is to recall that the issue is not what folk psychology as found in the field truly is, but what it is at its best, what deserves to be taken seriously and incorporated into science. It is not particularly to the point to argue against me that folk psychology is *in fact* committed to beliefs and desires as distinguishable, causally interacting *illata*; what must be shown is that it ought to be. The latter claim I will deal with in due course. The former claim I *could* concede without embarrassment to my overall project, but I do not concede it, for it seems to me that the evidence is quite strong that our ordinary notion of belief has next to nothing of the concrete in it. Jacques shoots his uncle dead in Trafalgar Square and is apprehended on the spot by Sherlock; Tom reads about it in the *Guardian* and Boris learns of it in

Pravda. Now Jacques, Sherlock, Tom and Boris have had remarkably *different* experiences – to say nothing of their earlier biographies and future prospects – but there is one thing they share: they all believe that a Frenchman has committed murder in Trafalgar Square. They di l r ot all *say* this, not even 'to themselves'; *that , ro,osition* did not, we can suppose, 'occur to' any of them, and even if it had, it would have had entir·ly different import for Jacques, Sherlock, Tom and Boris. Yet they all be'ieve that a Frenchman committed murder i: Trafalgar Square. This is a shared property that is, as it were, visible only from one very limited point of view – the point of view of folk psychology. Ordinary folk psych logists have no difficulty imputing such useful but elusive commonalities to people. If they then insist that in doing s they are postulating a similarly structured object, as it were, in each head, this is a gratuitous bit of misplaced concreteness, a regrettable lapse in ideology.

But in any case there is no doubt that folk psychology is a mixed bag, like folk productions generally, and there is no reason in the end not to grant that it is much more complex, variegated (and in danger of incoherence) than my sketch has made it out to be. The *ordinary* notion of belief no doubt does place beliefs somewhere midway between being *illata* and being *abstracta*. What this suggests to me is that the concept of belief found in ordinary understanding, i.e. in folk psychology, is unappealing as a scientific concept. I am reminded of Anaxagoras' strange precursor to atomism: the theory of seeds. There is a portion of everything in everything, he is reputed to have claimed. Every object consists of an infinity of seeds, of all possible varieties. How do you make bread out of flour, yeast and water? Flour contains bread seeds in abundance (but flour seeds predominate – that's what makes it flour), and so do yeast and water, and when these ingredients are mixed together, the bread seeds form a new majority, so bread is what you get. Bread nourishes by containing flesh and blood and bone seeds in addition to its majority of bread seeds. Not good theoretical entities, these seeds, for as a sort of bastardized cross between properties and proper parts they have a penchant for generating vicious regresses, and their identity conditions are problematic to say the least.

Beliefs are rather like that. There seems no comfortable way of avoiding the claim that we have an infinity of beliefs, and common intuition does not give us a stable answer to such puzzles as whether the belief that 3 is greater than 2 is none other than the belief that 2 is less than 3. The obvious response to the challenge of an infinity of beliefs with slippery identity conditions is to suppose these beliefs are not all 'stored separately'; many – in fact *most* if we are really talking about infinity – will be stored *implicitly* in virtue of the *explicit* storage of a few (or a few million)

– the *core beliefs*.[18] The core beliefs will be 'stored separately', and they look like promising *illata* in contrast to the *virtual* or *implicit* beliefs which look like paradigmatic *abstracta*. But although this might turn out to be the way our brains are organized, I suspect things will be more complicated than this: there is no reason to suppose the core *elements*, the concrete, salient, separately stored representation-tokens (and there must be some such elements in any complex information processing system), will explicitly represent (or *be*) a subset of our *beliefs* at all. That is, if you were to sit down and write out a list of a thousand or so of your paradigmatic beliefs, *all* of them could turn out to be virtual, only implicitly stored or represented, and what was explicitly stored would be information (e.g. about memory addresses, procedures for problem-solving, or recognition, etc.) that was entirely unfamiliar. It would be folly to prejudge this empirical issue by insisting that our core representations of information (whichever they turn out to be) are beliefs *par excellence*, for when the facts are in our intuitions may instead support the contrary view: the least controversial self-attributions of belief may pick out beliefs that from the vantage point of developed cognitive theory are invariably virtual.[19]

In such an eventuality what could we say about the *causal* roles we assign ordinarily to beliefs (e.g. 'Her belief that John knew her secret caused her to blush')? We could say that whatever the core elements were in virtue of which she virtually believed that John knew her secret, they, the core elements, played a direct causal role (somehow) in triggering the blushing response. We would be wise, as this example shows, not to tamper with our *ordinary* catalogue of beliefs (virtual though they might all turn out to be), for these are predictable, readily understandable, manipulable regularities in psychological phenomena in spite of their apparent neutrality with regard to the explicit/implicit (or core/virtual) distinction. What Jacques, Sherlock, Boris and Tom have in common is probably only a virtual belief 'derived' from largely different explicit stores of information in each of them, but virtual or not, it is their sharing of *this* belief that would explain (or permit us to predict) in some imagined circumstances their all taking the same action when given the same new information. ('And now for one million dollars, Tom [Jacques, Sherlock, Boris], answer our jackpot question correctly: has a French citizen ever committed a major crime in London?')

At the same time we want to cling to the equally ordinary notion that beliefs can cause not only actions, but blushes, verbal slips, heart attacks and the like. Much of the debate over whether or not intentional explanations are causal explanations can be bypassed by noting how the

[18] See my 'Brain Writing and Mind Reading', 1975. See also Fodor 1975, and Field 1978.
[19] See Field 1978: 55, n. 12 on 'minor concessions' to such instrumentalistic treatments of belief.

core elements, *whatever they may be*, can be cited as playing the causal role, while belief remains virtual. 'Had Tom not believed that p and wanted that q, he would not have done A.' Is this a causal explanation? It is tantamount to this: Tom was in some one of an indefinitely large number of structurally different states of type B that have in common just that each one of them licenses attribution of belief that p and desire that q in virtue of its normal relations with many other states of Tom, and this state, whichever one it was, was causally sufficient, given the 'background conditions' of course, to initiate the intention to perform A, and thereupon A was performed, and had he not been in one of those indefinitely many type B states, he would not have done A. One can call this a causal explanation because it talks about causes, but it is surely as unspecific and unhelpful as a causal explanation can get. It commits itself to there being some causal explanation or other falling within a very broad area (i.e. the intentional interpretation is held to be supervenient on Tom's bodily condition), but its true informativeness and utility in actual prediction lie, not surprisingly, in its assertion that Tom, however his body is currently structured, has a particular set of these elusive intentional properties, beliefs and desires.

The ordinary notion of belief is pulled in two directions. If we want to have *good* theoretical entities, good *illata*, or good logical constructs, good *abstracta*, we will have to jettison some of the ordinary freight of the concepts of belief and desire. So I propose a divorce. Since we seem to have both notions wedded in folk psychology, let's split them apart and create two new theories: one strictly abstract, idealizing, holistic, instrumentalistic – pure intentional system theory – and the other a concrete, micro-theoretical science of the actual realization of those intentional systems – what I will call sub-personal cognitive psychology. By exploring their differences and interrelations, we should be able to tell whether any plausible 'reductions' are in the offing.

2

The first new theory, intentional system theory, is envisaged as a close kin of – and overlapping with – such already existing disciplines as decision theory and game theory, which are similarly abstract, normative and couched in intentional language. It borrows the ordinary terms, 'belief' and 'desire' but gives them a technical meaning within the theory. It is a sort of holistic logical behaviourism because it deals with the prediction and explanation from belief–desire profiles of the actions of whole systems (either alone in environments or in interaction with other intentional systems), but treats the individual realizations of the systems as black

boxes. The *subject* of all the intentional attributions is the whole system (the person, the animal, or even the corporation or nation)[20] rather than any of its parts, and individual beliefs and desires are not attributable in isolation, independently of other belief and desire attributions. The latter point distinguishes intentional system theory most clearly from Ryle's logical behaviourism, which took on the impossible burden of characterizing individual beliefs (and other mental states) as particular individual dispositions to outward behaviour.

The theory deals with the 'production' of new beliefs and desires from old, *via* an interaction among old beliefs and desires, features in the environment, and the system's actions, and this creates the illusion that the theory contains naturalistic descriptions of internal processing in the systems the theory is about, when in fact the processing is all in the manipulation of the theory, and consists in updating the intentional characterization of the whole system according to the rules of attribution. An analogous illusion of process would befall a naive student who, when confronted with a parallelogram of forces, supposed that it pictured a mechanical linkage of rods and pivots of some kind instead of being simply a graphic way of representing and plotting the effect of several simultaneously acting forces.

Richard Jeffrey (1970), in developing his concept of probability kinematics, has usefully drawn attention to an analogy with the distinction in physics between kinematics and dynamics. In kinematics,

> you talk about the propagation of motions throughout a system in terms of such constraints as rigidity and manner of linkage. It is the physics of position and time, in terms of which you can talk about velocity and acceleration, but not about force and mass. When you talk about forces – *causes* of accelerations – you are in the realm of dynamics (172).

Kinematics provides a simplified and idealized level of abstraction appropriate for many purposes – e.g. for the *initial* design development of a gearbox – but when one must deal with more concrete details of systems – e.g. when the gearbox designer must worry about friction, bending, energetic efficiency and the like – one must switch to dynamics for more detailed and reliable predictions, at the cost of increased complexity and diminished generality. Similarly one can approach the study of belief (and desire and so forth) at a highly abstract level, ignoring problems of realization and simply setting out what the normative demands on the design of a believer are. For instance, one can ask such questions as 'What must a system's epistemic capabilities and propensities be for it to survive in environment A?'[21] or 'What must this system already know in order for

[20] See my 'Conditions of Personhood' (1976).

[21] Cf. Campbell 1973, and his William James lectures (Harvard U.P., forthcoming).

it to be able to learn B?' or 'What intentions must this system have in order to mean something by saying something?'[22]

Intentional system theory deals just with the performance specifications of believers while remaining silent on how the systems are to be implemented. In fact this neutrality with regard to implementation is the most useful feature of intentional characterizations. Consider, for instance, the role of intentional characterizations in evolutionary biology. If we are to explain the evolution of complex behavioural capabilities or cognitive talents by natural selection, we must note that it is the intentionally characterized capacity (e.g. the capacity to acquire a belief, a desire, to perform an intentional action) that has survival value, however it happens to be realized as a result of mutation. If a particularly noxious insect makes its appearance in an environment, the birds and bats with a survival advantage will be those that come to believe this insect is not good to eat. In view of the vast differences in neural structure, genetic background and perceptual capacity between birds and bats, it is highly unlikely that this useful trait they may come to share has a common description at any level more concrete or less abstract than intentional system theory. It is not only that the intentional predicate is a projectible predicate in evolutionary theory; since it is more general than its species-specific counterpart predicates (which characterize the successful mutation just in birds, or just in bats), it is preferable. So from the point of view of evolutionary biology, we would not want to 'reduce' all intentional characterizations even if we knew in particular instances what the physiological implementation was.

This level of generality is essential if we want a theory to have anything meaningful and defensible to say about such topics as intelligence in general (as opposed, say, to just human or even terrestrial or natural intelligence), or such grand topics as meaning or reference or representation. Suppose, to pursue a familiar philosophical theme, we are invaded by Martians, and the question arises: do they have beliefs and desires? Are they that much *like us*? According to intentional system theory, if these Martians are smart enough to get here, then they most certainly have beliefs and desires – in the technical sense proprietary to the theory – no matter what their internal structure, and no matter how our folk-psychological intuitions rebel at the thought.

This principled blindness of intentional system theory to internal structure seems to invite the retort:[23] but there has to be *some* explanation of the *success* of intentional prediction of the behaviour of systems. It isn't

[22] The questions of this variety are familiar, of course, to philosophers, but are now becoming equally familiar to researchers in artificial intelligence.

[23] From Ned Block and Jerry Fodor, *inter alia*, in conversation.

just magic. It isn't a mere coincidence that one can generate all these *abstracta*, manipulate them *via* some version of practical reasoning, and come up with an action prediction that has a good chance of being true. There must be some way in which the internal processes of the system mirror the complexities of the intentional interpretation, or its success would be a miracle.

Of course. This is all quite true and important. Nothing without a great deal of structural and processing complexity could conceivably realize an intentional system of any interest, and the complexity of the realization will surely bear a striking resemblance to the complexity of the instrumentalistic interpretation. Similarly, the success of valence theory in chemistry is no coincidence, and people were entirely right to expect that deep microphysical similarities would be discovered between elements with the same valence, and that the structural similarities found would explain the dispositional similarities. But since people and animals are unlike atoms and molecules not only in being the products of a complex evolutionary history, but also in being the products of their individual learning histories, there is no reason to suppose that individual (human) believers that *p* – like individual (carbon) atoms with valence 4 – regulate their dispositions with *exactly* the same machinery. Discovering the constraints on design and implementation variation, and demonstrating how particular species and individuals in fact succeed in realizing intentional systems is the job for the third theory: sub-personal cognitive psychology.

3

The task of sub-personal cognitive psychology is to explain something that at first glance seems utterly mysterious and inexplicable. The brain, as intentional system theory and evolutionary biology show us, is a *semantic engine*; its task is to discover what its multifarious inputs *mean*, to discriminate them by their significance and 'act accordingly'.[24] That's what brains *are for*. But the brain, as physiology or plain common sense shows us, is just a *syntactic engine*; all it can do is discriminate its inputs by their structural, temporal, and physical features, and let its entirely mechanical activities be governed by these 'syntactic' features of its

[24] More accurately if less picturesquely, the brain's task is to come to produce internal mediating responses that reliably vary in concert with variation in the actual environmental significance (the natural and non-natural meanings, in Grice's (1957) sense) of their distal causes and independently of meaning-irrelevant variations in their proximal causes, and moreover to respond to its own mediating responses in ways that systematically tend to improve the creature's prospects in its environment if the mediating responses are varying as they ought to vary.

inputs. That's all brains *can do*. Now how does the brain manage to get semantics from syntax? How could *any* entity (how could a genius, or an angel, or God) get the semantics of a system from nothing but its syntax? It couldn't. The syntax of a system doesn't determine its semantics. By what alchemy, then, does the brain extract semantically reliable results from syntactically driven operations? It cannot be designed to do an impossible task, but it could be designed to *approximate* the impossible task, to *mimic* the behaviour of the impossible object (the semantic engine) by capitalizing on close (close enough) fortuitous correspondences between structural regularities – of the environment and of its own internal states and operations – and semantic types.

The basic idea is familiar. An animal needs to know when it has satisfied the goal of finding and ingesting food, but it settles for a friction-in-the-throat-followed-by-stretched-stomach detector, a mechanical switch turned on by a relatively simple mechanical condition that *normally* co-occurs with the satisfaction of the animal's 'real' goal. It's not fancy, and can easily be exploited to trick the animal into either eating when it shouldn't or leaving off eating when it shouldn't, but it does well enough by the animal in its normal environment. Or suppose I am monitoring telegraph transmissions and have been asked to intercept all *death threats* (but only death threats in English – to make it 'easy'). I'd like to build a machine to save me the trouble of interpreting semantically every message sent, but how could this be done? No machine could be designed to do the job perfectly, for that would require defining the semantic category *death threat in English* as some tremendously complex feature of strings of alphabetic symbols, and there is utterly no reason to suppose this could be done in a principled way. (If somehow by brute-force inspection and subsequent enumeration we could list all and only the English death threats of, say, less than a thousand characters, we could easily enough build a filter to detect them, but we are looking for a principled, projectible, extendable method.) A really crude device could be made to discriminate all messages containing the symbol strings

. . . I will kill you . . .

or

. . . you . . . die . . . unless . . .

or

. . . (for some finite disjunction of likely patterns to be found in English death threats).

This device would have some utility, and further refinements could screen the material that passed this first filter, and so on. An unpromising beginning for constructing a sentence understander, but if you want to get semantics out of syntax (whether the syntax of messages in a natural

language or the syntax of afferent neuron impulses), variations on this basic strategy are your only hope.[25] You must put together a bag of tricks and hope nature will be kind enough to let your device get by. Of course some tricks are elegant, and appeal to deep principles of organization, but in the end all one can hope to produce (all natural selection can have produced) are systems that *seem* to discriminate meanings by actually discriminating things (tokens of no doubt wildly disjunctive types) that co-vary reliably with meanings.[26] Evolution has designed our brains not only to do this but to evolve and follow strategies of self-improvement in this activity during their individual lifetimes.[27]

It is the task of sub-personal cognitive psychology to propose and test models of such activity – of pattern recognition or stimulus generalization, concept learning, expectation, learning, goal-directed behaviour, problem-solving – that not only produce a simulacrum of genuine content-sensitivity, but that do this in ways demonstrably like the way people's brains do it, exhibiting the same powers and the same vulnerabilities to deception, overload and confusion. It is here that we will find our good theoretical entities, our useful *illata*, and while some of them may well resemble the familiar entities of folk psychology – beliefs, desires, judgments, decisions – many will certainly not.[28] The only similarity we can be sure of discovering in the *illata* of sub-personal cognitive psychology is the

[25] One might think that while *in principle* one cannot derive the semantics of a system from nothing but its syntax, *in practice* one might be able to cheat a little and exploit syntactic features that don't *imply* a semantical interpretation, but strongly suggest one. For instance, faced with the task of deciphering isolated documents in an entirely unknown and alien language, one might note that while the symbol that *looks like* a duck doesn't *have* to mean 'duck', there is a good chance that it does, especially if the symbol that looks like a wolf seems to be eating the symbol that looks like a duck, and not *vice versa*. Call this *hoping for hieroglyphics* and note the form it has taken in psychological theories from Locke to the present: we will be able to tell which mental representations are which (which idea is the idea of *dog* and which of *cat*) because the former will look like a dog and the latter like a cat. This is all very well as a crutch for us observers on the outside, trying to assign content to the events in some brain, but it is of no use to the brain . . . because brains don't know what dogs look like! Or better, this cannot be the brain's fundamental method of eking semantic classes out of raw syntax, for any brain (or brain part) that could be said – in an extended sense – to know what dogs look like would be a brain (or brain part) that had already solved its problem, that was already (a simulacrum of) a semantic engine. But this is still misleading, for brains in any event do not *assign* content to their own events in the way observers might: brains *fix* the content of their internal events in the act of reacting as they do. There are good reasons for positing *mental images* of one sort or another in cognitive theories (see 'Two Approaches to Mental Images' in *Brainstorms* pp. 174–89) but hoping for hieroglyphics isn't one of them, though I suspect it is covertly influential.

[26] I take this point to be closely related to Davidson's reasons for claiming there can be no psycho-physical laws, but I am unsure that Davidson wants to draw the same conclusions from it that I do. See Davidson 1970.

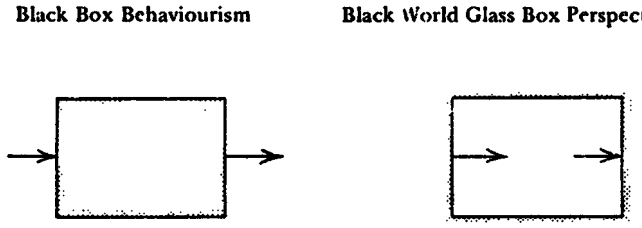
[27] This claim is defended in my 'Why the law of effect will not go away' (1974).

[28] See, for instance, Stephen Stich's (1978) concept of subdoxastic states.

intentionality of their labels.[29] They will be characterized as events with content, bearing information, signalling this and ordering that.

In order to give the *illata* these labels, in order to maintain any intentional interpretation of their operation at all, the theorist must always keep glancing outside the system, to see what normally produces the configuration he is describing, what effects the system's responses normally have on the environment, and what benefit normally accrues to the whole system from this activity. In other words the cognitive psychologist cannot ignore the fact that it is the realization of an intentional system he is studying on pain of abandoning semantic interpretation and hence psychology. On the other hand, progress in sub-personal cognitive psychology will blur the boundaries between it and intentional system theory, knitting them together much as chemistry and physics have been knit together.

The alternative of ignoring the external world and its relations to the internal machinery (what Putnam has called psychology in the narrow sense, or methodological solipsism, and Keith Gunderson lampoons as black world glass box perspectivalism)[30] is not really psychology at all,



but just at best abstract neurophysiology – pure internal syntax with no hope of a semantic interpretation. Psychology 'reduced' to neurophysiology in this fashion would not be psychology, for it would not be able to provide an explanation of the regularities it is psychology's particular job to explain: the reliability with which 'intelligent' organisms can cope with their environments and thus prolong their lives. Psychology can, and should, work towards an account of the physiological foundations of psychological processes, not by eliminating psychological or intentional characterizations of those processes, but by exhibiting how the brain

[29] See my 'Reply to Arbib and Gunderson', in *Brainstorms*, pp. 23–30.

[30] In his reply to Fodor's 'Methodological Solipsism as a Research Strategy in Psychology' at the Cincinnati Colloquium on Philosophy of Psychology, February 1978.

implements the intentionally characterized performance specifications of sub-personal theories.[31]

Friedman, discussing the current perplexity in cognitive psychology, suggests that the problem

> is the direction of reduction. Contemporary psychology tries to explain *individual* cognitive activity independently from *social* cognitive activity, and then tries to give a *micro* reduction of social cognitive activity – that is, the use of a public language – in terms of a prior theory of individual cognitive activity. The opposing suggestion is that we first look for a theory of social activity, and then try to give a *macro* reduction of individual cognitive activity – the activity of applying concepts, making judgments, and so forth – in terms of our prior social theory.[32]

With the idea of macro-reduction in psychology I largely agree, except that Friedman's identification of the macro level as explicitly *social* is only part of the story. The cognitive capacities of non-language-using animals (and Robinson Crusoes, if there are any) must also be accounted for, and not just in terms of an analogy with the practices of us language users. The macro level *up* to which we should relate micro-processes in the brain in order to understand them as psychological is more broadly the level of organism–environment interaction, development and evolution. That level includes social interaction as a particularly important part,[33] but still a proper part.

There is no way to capture the semantic properties of things (word tokens, diagrams, nerve impulses, brain states) by a micro-reduction. Semantic properties are not just relational but, you might say, super-relational, for the relation a particular vehicle of content, or token, must bear in order to have content is not just a relation it bears to other similar things (e.g. other tokens, or parts of tokens, or sets of tokens, or causes of tokens) but a relation between the token and the whole life – and counter-factual life[34] – of the organism it 'serves' *and* that organism's requirements for survival *and* its evolutionary ancestry.

4

Of our three psychologies – folk psychology, intentional system theory, and sub-personal cognitive psychology – what then might reduce to what?

[31] I treat methodological solipsism in (much) more detail in 'Beyond Belief', in Andrew Woodfield, ed. *Thought and Object*.

[32] Michael Friedman, 'Theoretical Explanation', this volume, pp. 15–16.

[33] See Tyler Burge 1979.

[34] What I mean is this: counterfactuals enter because content is in part a matter of the *normal* or *designed* role of a vehicle whether or not it ever gets to play that role. Cf. Sober (unpublished).

Certainly the one-step micro-reduction of folk psychology to physiology alluded to in the slogans of the early identity theorists will never be found – and should never be missed, even by staunch friends of materialism and scientific unity. A prospect worth exploring, though, is that folk psychology (more precisely, the part of folk psychology worth caring about) reduces – conceptually – to intentional system theory. What this would amount to can best be brought out by contrasting this proposed conceptual reduction with more familiar alternatives: 'type-type identity theory' and 'Turing machine functionalism'. According to type-type identity theory, for every mentalistic term or predicate '*M*', there is some predicate '*P*' *expressible in the vocabulary of the physical sciences* such that a creature is *M* if and only if it is *P*. In symbols:

(1) $(x)(Mx \equiv Px)$

This is reductionism with a vengeance, taking on the burden of replacing, in principle, all mentalistic predicates with co-extensive predicates composed truth-functionally from the predicates of physics. It is now widely agreed to be hopelessly too strong a demand. Believing that cats eat fish is, intuitively, a *functional* state that might be variously implemented physically, so there is no reason to suppose the commonality referred to on the left-hand side of (1) can be reliably picked out by any predicate, however complex, of physics. What is needed to express the predicate on the right-hand side is, it seems, a physically neutral language for speaking of functions and functional states, and the obvious candidates are the languages used to describe automata – for instance. Turing machine language.

The Turing machine functionalist then proposes

(2) $(x)(Mx \equiv$ x realizes some Turing machine k in logical state A)

In other words, for two things both to believe that cats eat fish they need not be physically similar in any specifiable way, but they must both be in a 'functional' condition specifiable in principle in the most general functional language; they must share a Turing machine description according to which they are both in some particular logical state. This is still a reductionist doctrine, for it proposes to identify each mental type with a functional type picked out in the language of automata theory. But this is still too strong, for there is no more reason to suppose Jacques, Sherlock, Boris and Tom 'have the same programme' in *any* relaxed and abstract sense, considering the differences in their nature and nurture, than that their brains have some crucially identical physico-chemical feature. We must weaken the requirements for the right-hand side of our formula still further.

Consider

(3) (x)(x believes that p≡x can be predictively attributed the belief that p)

This appears to be blatantly circular and uninformative, with the language on the right simply mirroring the language on the left. But all we need to make an informative answer of this formula is a systematic way of making the attributions alluded to on the right-hand side. Consider the parallel case of Turing machines. What do two different realizations or embodiments of a Turing machine have in common when they are in the same logical state? Just this: there is a system of description such that according to it both are described as being realizations of some particular Turing machine, and according to this description, which is predictive of the operation of both entities, both are in the same state of that Turing machine's machine table. One doesn't *reduce* Turing machine talk to some more fundamental idiom; one *legitimizes* Turing machine talk by providing it with rules of attribution and exhibiting its predictive powers. If we can similarly legitimize 'mentalistic' talk, we will have no need of a reduction, and that is the point of the concept of an intentional system. Intentional systems are supposed to play a role in the legitimization of mentalistic predicates parallel to the role played by the abstract notion of a Turing machine in setting down rules for the interpretation of artifacts as computational automata. I fear my concept is woefully informal and unsystematic compared with Turing's, but then the domain it attempts to systematize – our everyday attributions in mentalistic or intentional language – is itself something of a mess, at least compared with the clearly defined field of recursive function theory, the domain of Turing machines.

The analogy between the theoretical roles of Turing machines and intentional systems is more than superficial. Consider that warhorse in the philosophy of mind, Brentano's Thesis that intentionality is the mark of the mental: all mental phenomena exhibit intentionality and no physical phenomena exhibit intentionality. This has been traditionally taken to be an *irreducibility* thesis: the mental, in virtue of its intentionality, cannot be reduced to the physical. But given the concept of an intentional system, we can construe the first half of Brentano's Thesis – all mental phenomena are intentional – as a *reductionist* thesis of sorts, parallel to Church's Thesis in the foundation of mathematics.

According to Church's Thesis, every 'effective' procedure in mathematics is recursive, that is, Turing-computable. Church's Thesis is not provable, since it hinges on the intuitive and informal notion of an effective procedure, but it is generally accepted, and it provides a very useful reduction of a fuzzy-but-useful mathematical notion to a crisply

defined notion of apparently equal scope and greater power. Analogously, the claim that every mental phenomenon alluded to in folk psychology is *intentional-system-characterizable* would, if true, provide a reduction of the mental as ordinarily understood – a domain whose boundaries are at best fixed by mutual acknowledgment and shared intuition – to a clearly defined domain of entities, whose principles of organization are familiar, relatively formal and systematic, and entirely general.[35]

This reductive claim, like Church's Thesis, cannot be proven, but could be made compelling by piecemeal progress on particular (and particularly difficult) cases – a project I set myself elsewhere (in *Brainstorms*). The final reductive task would be to show not how the terms of intentional system theory are eliminable in favour of physiological terms via sub-personal cognitive psychology, but almost the reverse: to show how a system described in physiological terms could warrant an interpretation as a realized intentional system.

REFERENCES

Block, N. 1978. 'Troubles with functionalism.' *Perception and Cognition: Issues in the Foundations of Psychology*, ed. C. Wade Savage, pp. 261–326. Minnesota Studies in Philosophy of Science, vol. IX. Minneapolis: Minnesota University Press.

Burge, T. 1979. 'Individualism and the mental.' *Midwest Studies in Philosophy*, vol. IV, pp. 73–121.

Campbell, D. 1973. 'Evolutionary epistemology.' *The Philosophy of Karl Popper*, ed. Paul A. Schilpp. La Salle, Illinois: Open Court.

Davidson, D. 1970. 'Mental events.' *Experience and Theory*, ed. L. Foster and J. Swanson, pp. 79–102. Amherst: University of Massachusetts Press.

Dennett, D. C. 1971. 'Intentional systems.' *Journal of Philosophy* 68, 87–106. Reprinted (with other essays on intentional systems) in *Brainstorms*, pp. 3–22.

Dennett, D. C. 1974. 'Why the law of effect will not go away.' *Journal of the Theory of Social Behaviour* 5, 169–187. Reprinted in *Brainstorms*, pp. 71–89.

Dennett, D. C. 1975. 'Brain writing and mind reading.' *Language, Mind and Knowledge*, ed. K. Gunderson. Minnesota Studies in Philosophy of Science, vol. VII. Minneapolis: Minnesota University Press. Reprinted in *Brainstorms*, pp. 39–50.

Dennett, D. C. 1976. 'Conditions of personhood.' *The Identities of Persons*, ed. A. Rorty. Reprinted in *Brainstorms*, pp. 267–85.

Dennett, D. C. 1978. *Brainstorms*. Montgomery, Vermont: Bradford Books; Hassocks, Sussex: Harvester Press.

[35] Ned Block (1978) presents arguments supposed to show how the various possible functionalist theories of mind all slide into the sins of 'chauvinism' (improperly excluding Martians from the class of possible mind-havers) or 'liberalism' (improperly including various contraptions, imagined human puppets, and so forth among the mind-havers). My view embraces the broadest liberalism, gladly paying the price of a few recalcitrant intuitions for the generality gained.

Field, H. 1978. 'Mental representation.' *Erkenntnis* 13, 9–61.
Fodor, J. 1975. *The Language of Thought.* Hassocks, Sussex: Harvester Press; Scranton, Pa.: Crowell.
Grice, H. P. 1957. 'Meaning.' *Philosophical Review* 66, 377–88.
Jeffrey, R. 1970. 'Dracula meets Wolfman: acceptance vs. partial belief.' *Induction, Acceptance and Rational Belief*, ed. Marshall Swain. Dordrecht: Reidel.
Lewis, D. 1974. 'Radical interpretation.' *Synthèse* 23, 331–44.
Nisbett, R. E. and Ross, L. D. 1978. *Human Inference. Strategy and Shortcomings.* Englewood Cliffs, N.J.: Prentice Hall.
Putnam, H. 1975. 'The meaning of "meaning".' *Mind, Language and Reality* (*Philosophical Papers*, vol. II), pp. 215–71. Cambridge: Cambridge University Press.
Reichenbach, H. 1938. *Experience and Prediction.* Chicago: University of Chicago Press.
Ryle, G. 1949. *The Concept of Mind.* London: Hutchinson.
Schank, R. 1977. 'Sam – a story understander.' Research Report 43, Yale University Dept of Computer Science.
Schank, R. and Abelson, R. 1977. *Scripts, Plans, Goals and Understanding.* Hillside, N.J.: Erlbaum.
Scheffler, I. 1963. *The Anatomy of Inquiry.* New York: Knopf.
Simon, H. 1969. *The Sciences of the Artificial.* Cambridge, Mass.: M.I.T. Press.
Sober, E. (unpublished) 'The descent of Mind.'
Stich, S. 1978. 'Belief and subdoxastic states.' *Philosophy of Science* 45, 499–518.
Tversky, A. and Kahneman, D. 1974. 'Judgement under uncertainty: heuristics and biases.' *Science* 185, 1124–31.
Woodfield, A., ed., forthcoming. *Thought and Object.* Oxford: Oxford University Press.