Expanded version of a chapter in *The Oxford Handbook of Contemporary Philosophy*, edited by Frank Jackson and Michael Smith (Oxford: Oxford University Press, 2005).

An Approach to Philosophy of Cognitive Science

MARTIN DAVIES

I Foundations

1. A Little History: Psychology, Philosophy, and Physiology

A century ago, psychology was in the process of separating from philosophy and becoming an experimental science, while also distinguishing itself from physiology. It is usually said that the father of experimental psychology was Wilhelm Wundt, who began conducting experiments in a laboratory in Leipzig in 1879. The domain of Wundt's investigation was that of conscious mental states and processes and his experimental method was skilled introspection under controlled circumstances. The scope of his experiments was relatively narrow, for introspection as Wundt used the notion was not a matter of retrospective and discursive description of phases of one's mental life. Experiments principally concerned sensations rather than higher mental processes and participants – who were usually the same people that conducted the experiments – were typically asked to report the onset of a sensation or to say whether two sensations were the same or different (Danziger, 1980, Thomas, 2001).

Wundt's project was taken to the United States by E.B. Titchener who completed his B.A. degree at Oxford – classics, philosophy, and a year of physiology – and, after doctoral research with Wundt, became professor of psychology at Cornell in 1882. Titchener and Oswald Külpe – another of Wundt's students – extended the experimental method of introspection far beyond the bounds that Wundt had set for it, so as to include higher mental processes of thinking within its scope. Titchener claimed that mental images were, in effect, the bearers of content or meaning in thinking and in the early twentieth century he and Külpe were involved in an extended controversy over the possibility of imageless thought.

During this period, while Wundt worked at Leipzig and Titchener at Cornell, William James held a series of positions in physiology, philosophy and psychology at Harvard. There are differences between James's teleological functionalist approach and the structuralism of Wundt and Titchener. But both sides were agreed that the emerging discipline of experimental psychology was to be the science of the conscious mind.

The experimental approach to psychology was taken up more rapidly in the United States than in Britain. But as early as 1892, the Scottish philosopher and psychologist Alexander Bain said, in a paper delivered to the International Congress of Experimental Psychology held in London (1893, p. 42):

> The resources at our disposal, in imparting to Psychology a scientific character, are now numerous. At the head, must still remain Introspection, or the self-consciousness of each individual working apart. . . . In the enumeration of means now available for the study are included observations (and experiments) directed upon Infants, upon Abnormal and Exceptional minds, upon Animals, and upon the workings of Society, or collective humanity. To these are added Physiology, and, last but not least, Psycho-physical experiments.

Bain repeatedly recurs to the primacy of introspection, 'the alpha and omega of psychological inquiry' (p. 42), but he also offers some brief indications of the role for

'psycho-physical' experiments in investigating the senses, 'where mind and body are most palpably associated' (p. 44). Introspection is crucial for qualitative analysis, because outward expressions do not reveal 'the full sequence of the mental movements' (p. 47). But measurement of outward signs – as, for example, by the 'reaction-time apparatus' that Hermann von Helmholtz had used to investigate the velocity of electrical impulses in nerves – is vital for the quantitative analysis that is characteristic of 'a science in the proper sense' (p. 48).

The University of Oxford took its first step towards developing psychology in 1898, appointing G.F. Stout – a philosopher and psychologist, though not an experimenter – to a position in 'mental philosophy'. Stout was also the editor of *Mind* and the journal's subtitle, 'A Quarterly Review of Psychology and Philosophy', served from its foundation by Bain in 1876 even until 1973 as a reminder of the historical links between the two disciplines. When, after five years, Stout moved to the chair of Logic and Metaphysics at St. Andrews, his successor in the mental philosophy position was William McDougall, an experimental psychologist of broad theoretical interests – including the explanation of behaviour in mentalistic terms – and one of the founders of the British Psychological Society. McDougall moved to the United States to be professor of psychology at Harvard, from 1920, and subsequently at Duke University. Back in Oxford, an Institute of Experimental Psychology was established in 1935, though the first professor of psychology was not appointed until 1947 and for another twenty years or more psychology in the undergraduate programme had to be accompanied either by philosophy or by physiology.

1.1 Behaviourism and the 'cognitive revolution'

By the 1920s, and especially in the United States, introspectionism had given way to the behaviourism of J.B. Watson and later B.F. Skinner. This was not merely methodological behaviourism, a restriction on admissible evidence and a rejection of the deliverances of introspection, but a radical redefinition of psychology's subject matter. Psychology was no longer to be the science of the conscious mind but, instead, the science of behaviour with stimulus and response, rather than sensation and feeling, its central theoretical notions. Whereas Bain had regarded the phenomena under investigation as two-sided, the 'objective and experimental' side was now elevated to supremacy, while the side of 'subjective consciousness' was banished altogether.

It is quite widely held that the restoration of the mind as the proper subject matter of psychology came in the 1950s and amounted to a revolutionary end to the behaviourist era. But behaviourism was more dominant in America than elsewhere and, even there, other approaches were still pursued – especially, but not only, in the psychology of perception. Internationally, there was continuing work on mental phenomena including memory, attention, and thinking and, of course, Gestalt theory offered an alternative theoretical framework to behaviourism (Hatfield, 2002). The description of psychology from the 1920s to the 1950s is further complicated by the influence of the operationalist insistence that theoretical notions should be defined in terms of, or at least tied very closely to, observable data. For this, taken together with methodological behaviourism, was apt to have consequences similar to radical behaviourism and, more generally, to block the construction of theories with real explanatory depth.

In any case, the so-called 'cognitive revolution' (Gardner, 1985) in American psychology owed much to developments in adjacent disciplines. The foundational work of Noam Chomsky in theoretical linguistics and research by John McCarthy, Marvin

Minsky, Allen Newell, and Herbert Simon in computer science, building on the pioneering research of Alan Turing, were of particular importance. Indeed, the cognitive revolution brought forth, not only a change in the conception of psychology, but also an inter-disciplinary approach to understanding the mind, involving philosophy, anthropology and neuroscience along with computer science, linguistics and psychology. George Miller (2003), a participant, and many commentators agree in dating the conception of this inter-disciplinary approach, cognitive science, to 11 September 1956, the second day of a symposium on information theory held at MIT. Over the next twenty years or so, cognitive science developed an institutional presence through research centres, conferences, journals, and a substantial infusion of funds from the Alfred P. Sloan Foundation.

2. Hamilton and Brentano on Unconscious Mental States

The most striking difference between introspectionist psychology before behaviourism and post-1956 cognitive psychology is the latter's appeal to unconscious mental states and processes. It is true that the notion of unconscious inference goes back at least to Helmholtz (1867). But, for the most part, the appeal to unconscious states and processes in late nineteenth or early twentieth century psychology was an appeal to physiological states and processes conceived as furnishing enabling conditions for the operation of the conscious mind.

2.1 Hamilton, the cognitive unconscious, and Chomsky

An argument by William Hamilton (1859) does, however, provide a fascinating antecedent for the contemporary appeal to the 'cognitive unconscious' (Manson, 2000). We can approach Hamilton's argument by first returning to Bain, who says that even in cases where outward signs disclose the steps of a 'truly mental operation', there is still something missing (1893, p. 47): 'Outward expression, however close and consecutive, is still hop, skip and jump. It does not supply the full sequence of mental movements. This entire unbroken sequence is revealed solely to Introspection.' In order to see how mental operations exemplify 'the primary or highest laws' we need to fill in all the intermediate links, and this can be done only 'by reference to inner consciousness' (p. 48).

Now, Hamilton focuses on cases in which introspection reveals an anomalous train of thought, a sequence that does not fit the (presumed) laws. And he proposes, not that such anomalies should be explained in terms of some failure of physiological enabling conditions, but rather that we should postulate unconscious mental states that fall within the scope of the same laws that apply to conscious thoughts. Only with such unconscious intermediate steps inserted can the sequence of mental states be seen to conform to the laws. So Hamilton would have agreed with Chomsky's remark, more than a century later, that a subject's conscious beliefs constitute only 'a scattered subpart of the full cognitive structure' (1976, p. 163).

Hamilton's argument was critically discussed by John Stuart Mill (1865), who recommended appeal to unconscious physiological states rather than unconscious mental states, and by Franz Brentano (1874, pp. 101–37), who attacked the very idea of unconscious mentality ('unconscious consciousness', as he called it) at some length. Concerning Hamilton's line of argument in particular, Brentano says (1874, p. 110):

> Like Hamilton, many philosophers have deduced the hypothesis of unconscious ideas from the fact that, when an earlier train of ideas is recalled, sometimes a

> whole series of intermediate steps appears to be skipped over. This fact would undoubtedly be reconciled with the laws of association if we were to assume that the intermediate steps in question had intervened on this occasion but without appearing in consciousness. Neither Hamilton nor others, however, have shown, or have even tried to show, that this is the only possible method of explanation.

And more generally (p. 137), 'The question, "Is there unconscious consciousness?" in the sense in which we have formulated it, is, therefore, to be answered with a firm, "No."'

2.2 Brentano, intentionality, and Searle

In some of the most discussed sentences in the philosophy of mind, Brentano claimed that intentionality is the distinctive mark of mental phenomena (1874, pp. 88–9):

> Every mental phenomenon is characterized by what the Scholastics of the Middle Ages called the intentional (or mental) inexistence of an object, and what we might call, though not wholly unambiguously, reference to a content, direction towards an object (which is not to be understood here as meaning a thing), or immanent objectivity. Every mental phenomenon includes something as object within itself, although they do not all do so in the same way. In presentation something is presented, in judgement something is affirmed or denied, in love loved, in hate hated, in desire desired and so on.
>
> This intentional in-existence is characteristic exclusively of mental phenomena. No physical phenomenon exhibits anything like it.

Brentano's claim clearly has two parts. First, *all* mental phenomena exhibit intentionality; and second, *only* mental phenomena exhibit intentionality. The first part has frequently been contested (for example, by Hamilton), especially as it applies to sensations and feelings; but versions of the first part of the claim continue to be defended (e.g. Crane, 2001). The second part, taken together with the point that mental phenomena 'are only perceived in inner consciousness' (1874, p. 91), rules out unconscious intentionality and so appears to challenge the very idea of the cognitive unconscious.

Brentano's notion of intentionality is not easy to understand. But it seems to include, even if it is not exhausted by, the idea that mental states can be about things ('something is presented') and can represent things as being the case ('something is affirmed or denied'). In short, intentionality seems to include the representational properties of mental states and so the second part of Brentano's claim leads to the doctrine that there are no unconscious representational states. This doctrine may seem to be open to clear counterexamples. For thoughts are sometimes expressed in written words that are about things and that represent the world as being one way or another. So words, like thoughts, may exhibit intentionality; yet there is no consciousness in the paper or the ink. But Brentano's doctrine can be refined so as to avoid this kind of problem, since words plausibly inherit their representational powers from the thoughts of speakers and hearers. The intentionality of language is perfectly genuine; but it seems to be a kind of *derived* intentionality. The critic of cognitive science who maintains the spirit of Brentano can say that no unconscious states are non-derivatively representational. Thus, for example, John Searle echoes Brentano's firm 'No' when he says (1990, p. 338): 'There are brute, blind, neurophysiological processes and there is consciousness; but there is nothing else. If we are looking for phenomena which are intrinsically [non-derivatively] intentional but inaccessible in principle to consciousness there is nothing there.' Indeed, Searle inserts

Brentano's firm 'No' into a list of what would usually be regarded as the glories of cognitive science: 'no rule-following, no mental information processing, no unconscious inferences, no mental models, no primal sketches, no $2\frac{1}{2}$D images, no three-dimensional descriptions, no language of thought and no universal grammar' (p. 338).

3. Personal and Subpersonal Levels of Description

From 1956 and through the 1960s, discussion in analytic philosophy of mind concerned competing theses about the metaphysics of mind – Gilbert Ryle's behaviourism, the materialism of U.T. Place, Jack Smart, and David Armstrong, Hilary Putnam's machine functionalism, and the rather different version of functionalism developed by David Lewis.[1] The 1960s also saw the publication of books by three of the most major figures in the philosophy of cognitive science. Chomsky's *Aspects of the Theory of Syntax* was published in 1965, with its striking claim about cognitive states that are inaccessible to consciousness, states of tacit knowledge of syntactic rules (1965, p. 8): 'Obviously, every speaker of a language has mastered and internalized a generative grammar that expresses his knowledge of his language. This is not to say that he is aware of the rules of the grammar or even that he can become aware of them.' In 1968, Jerry Fodor published his first book, *Psychological Explanation: An Introduction to the Philosophy of Psychology*, along with an important paper, 'The appeal to tacit knowledge in psychological explanations'. And Daniel Dennett's *Content and Consciousness*, with its distinction between personal and subpersonal levels of description, appeared in 1969.

3.1 Dennett's distinction

We may say that 'a person pulled his hand away from the stove . . . because it hurt' (Dennett, 1969, p. 91). But, Dennett says, we cannot, at this *personal* level of description, elaborate an account of the processes that led the person to remove his hand. This is because (p. 94): 'The only sort of explanation in which "pain" belongs is non-mechanistic.' Similarly, explanations that advert to mental phenomena exhibiting intentionality in Brentano's sense, such as the explanation of an act of trying in terms of a subject's desire, 'are not causal explanations in the more or less Humean sense of the term' (p. 35).

The general picture is that, at the personal level, we talk about persons as such – as experiencing, thinking subjects and agents. We describe what people feel and what people do, and we explain what people do in terms of their sensations, desires, beliefs and intentions. These personal-level explanations are of a distinctive, not-straightforwardly-causal, kind and they do not work by elaborating accounts of mental processes. Still less do they work by postulating physical mechanisms underpinning the activities of persons. An account of the physical mechanisms that are involved when a person withdraws his hand from a hot stove belongs at a quite different level of description and explanation. We abandon 'the explanatory level of people and their sensations and activities' and shift to 'the *sub-personal* level of brains and events in the nervous system' (p. 93). At this

[1] On behaviourism, see Ryle, 1949, though Ryle did not offer reductive behaviourist analyses. For a limited mind-brain identity theory, see Place, 1956. Place's 'inner process story' was developed and defended by Smart, 1959. Armstrong, 1968, is a thoroughgoing statement of central state materialism. On machine functionalism, see Putnam, 1967. Block and Fodor, 1972, pointed out that Putnam's machine functionalism lacks the resources to account for the fact that a creature may be in more than one psychological state at a time. Functionalism received its canonical modern formulation from Lewis, 1966, 1970, 1972.

subpersonal level of description and explanation, the kinds of occurrences that are described receive causal explanations in purely mechanistic terms. But, according to Dennett, these occurrences are not to be identified with the sensations and actions of persons. Indeed, Dennett does not assume that there are 'physical events, states or processes which deserve to be called thoughts, ideas, mental images and so forth' (p. 19).[2]

3.2 Intentional systems: An apparent tension and two ways to resolve it

Descriptions of persons in terms of beliefs and desires – *intentional* descriptions – exhibit the logical property of intensionality; descriptions that figure in the physical sciences are, in contrast, extensional. When Dennett asks (p. 40), 'Could there be a system of internal states or events, the extensional description of which could be upgraded into an Intentional description?', we might expect that he would answer, as Brentano would, in the negative. That is what the distinction between personal and subpersonal levels of description might seem to suggest. But in fact Dennett says (p. 40): 'The answer to this question is not at all obvious, but there are some promising hints that the answer is Yes.'

His strategy for developing these hints makes use of the notion of an intentional system and the related notion of the *intentional stance*. These notions are central in Dennett's later work in philosophy of mind. But the character of the strategy for attributing intentionality to physical phenomena is already apparent in *Content and Consciousness* (pp. 78, 80):

> [T]he relation between Intentional descriptions of events, states or structures (as signals that carry certain messages or memory traces with certain contents) and extensional descriptions of them is one of *further interpretation*.
>
> The ideal picture, then, is of content being ascribed to structures, events and states in the brain on the basis of a determination of origins in stimulation and eventual appropriate behavioural effects, such ascriptions being essentially a heuristic overlay on the extensional theory.

There appears to be a tension in Dennett's position here. The interpretative strategy of 'heuristic overlay' could be adopted towards a system that is not a person. Yet intentional descriptions were supposed to belong at the personal level, which is the level of 'people and their sensations and activities'. As Jennifer Hornsby points out (2000, pp. 17–18): 'Dennett's continued insistence on the importance of his personal/sub-personal distinction becomes hard to fathom when properties visible at the personal level are meant to be the products of a stance that is equally appropriately adopted towards sub-personal things.'

To the extent that this apparent tension is genuine, there seem to be two ways to resolve it. One way would be to hold hard to what is distinctive of persons and to deny that personal-level intentionality can be literally attributed to subpersonal-level systems. The other way would be to take a more relaxed view of the distinction between the personal and subpersonal levels of description and to allow that personal-level

[2] Dennett's metaphysics of mind is closer to Ryle's behaviourism than to the identity theory or functionalism. In its later development (e.g. 1971, 1987), it is naturally described as a kind of 'supervenient behaviourism'; attributions of beliefs and other propositional attitudes are made true by patterns in behaviour.

intentionality is the product of adopting a stance that can just as well be adopted towards subpersonal-level systems.

The first way is apt to lead to a worry about the theoretical foundations of cognitive science. For it may seem that the appeal to unconscious representations and tacitly known rules involves a kind of category mistake in which distinctively personal-level notions, such as representation and rule, are applied at a subpersonal level of description. I shall consider responses to this worry in the next section.

If we take the second way of resolving the apparent tension in Dennett's position then no such worry arises. For we allow that the intentionality of people's conscious mental states is of a piece with the representational properties that we attribute to neural states on the basis of their causes, their effects, and some relation of harmony (appropriateness, in Dennett's terminology) between inputs and outputs. Indeed, people's conscious thoughts and the states of brains, computers, and thermostats can all be regarded as intentional or representational states in just the same sense.

This broadly reductionist approach to intentionality and representation has, in fact, been dominant in recent philosophy of mind and cognitive science and there is no denying that it has some clear advantages, of which I mention just two. Suppose that the notion of a representational state, and so of a representational mental state, can be elucidated without essential appeal to the notion of consciousness. Then, first, the residual mystery of consciousness need not be regarded as attaching to the core business of cognitive science. And, second, we can appeal to the notion of representation in trying to explain how it is that some mental states are gilded with consciousness. Indeed, we might be so bold as to hope that consciousness can be adequately explained in terms of representation – perhaps as meta-representation, that is, representation of representation (Rosenthal, 2002).

4. Intentionality and the Foundations of Cognitive Science

The aim of this section is to show that, in order to avoid foundational worries about cognitive science, it is not essential that we should adopt a reductionist view of personal-level intentionality

In 'Artificial intelligence as philosophy and as psychology' (1978), Dennett's concern is with an apparent problem posed by the fact that the notion of a representation goes along with that of an interpreter (1978, p. 122): 'something is a representation only *for* or *to* someone'. So cognitive science's appeal to internal representations requires an appeal, also, to internal interpreters of those representations – that is, to homunculi. And this looks like the beginning of a regress. Whether or not the threat of regress is genuine,[3] Dennett's account seems to offer a way of moving the notion of representation to a subpersonal level of description without committing a category mistake. For it involves populating the subpersonal level with little persons.

4.1 Homunculi, and 'as if' intentionality

Dennett's solution to the problem of the regress begins from the thought that the performance of a cognitive task can be secured by having subsystems perform parts of the task. These subsystems are to be thought of initially as intelligent homunculi whose

[3] See Fodor, 1975, p. 74, n. 14: 'the regress never needs to start'. Fodor is responding to Dennett, 1969, p. 87.

functions are in turn discharged by sub-subsystems and so on. But the crucial point is that these are ever less intelligent homunculi with ever simpler tasks to perform. So, in the end, the simplest tasks can be performed by mere mechanical devices, the homunculi are discharged, and the threat of an infinite regress is avoided (see also Block, 1995a).

In this account, talk about internal representations goes along with talk about little people, just as we should expect of a personal-level notion. But, of course, the talk of homunculi is metaphor. There are no little people there, just little mechanisms. And, once the homunculi are discharged, the whole story can be retold, this time literally rather than metaphorically, without any talk of intentionality, representations or rules. Personal-level notions enter the subpersonal-level account only as a metaphorical staging post *en route* to the non-metaphorical neurophysiological truth.

The overall picture that emerges is this. When we stress what is distinctive about persons, we make literal use of personal-level notions. Literal application of those same notions to pieces of cognitive machinery would be illegitimate, just as the foundational worry suggests. But the metaphorical use of personal-level notions in subpersonal-level psychological descriptions is conceptually unproblematic as, of course, is the purely biological description of neural mechanisms. So the personal level is distinguished from two legitimate subpersonal levels, one making use of '*as if*' intentional descriptions and the other making use of literal biological descriptions.[4]

We might wonder whether this picture differs significantly from Searle's. For, so far as the literal truth goes, it seems to be agreed that 'There are brute, blind, neurophysiological processes and there is consciousness; but there is nothing else' (Searle, 1990, p. 338). But it is crucial to observe that the notion of 'as if' intentionality might be offered in either a more constructive or a more critical spirit. Hornsby (2000) regards the notion as having serious explanatory potential and, in a similar vein, John McDowell says (1994, p. 199): 'To insist that the attribution of content at this sub-personal . . . level is "as if" talk is in no way to debunk it. . . . And it is surely clear, at least in a general way, how content-attribution that is only "as if" can even so pull its weight in addressing a genuine explanatory need.' Searle, in contrast, uses the idea of 'as if' attributions of intentionality in an argument that is much more critical of cognitive science.

4.2 'As if' intentionality and Searle's critique of cognitive science

McDowell regards 'as if' attributions as 'not irresponsible' and as 'constrained by the physiological facts' (1994, p. 199). But Searle says that 'as if' intentionality is to be found everywhere (1992, p. 156): 'Everything in the universe follows laws of nature, and for that reason everything behaves with a certain degree of regularity, and for that reason everything behaves *as if* it were following a rule, trying to carry out a certain project, acting in accordance with certain desires, etc.' That is the first main part of his argument: 'as if' intentionality is trivial.

The second part begins from the point that clear cases of intentional mental states, such as beliefs, have intrinsic or non-derivative intentionality and present their objects,

[4] In this account, the shift from the personal to subpersonal levels is also a shift to talk about subsystems. But a level of description is not the same as a level of aggregation and the primary use of the term 'subpersonal' is not to indicate parts of a person. Nevertheless, the primary use of the term allows that a proper part of a person, lacking the properties that are distinctive of persons, may be the subject of subpersonal-level descriptions. See further, Hornsby, 1997, pp. 161–7.

for example, a planet or a kind of stuff, under aspects – the Morning Star aspect or the Evening Star aspect, the water aspect or the H_2O aspect. From there the argument moves, in a series of steps, to the conclusion that 'all unconscious mental states are in principle accessible to consciousness' (the Connection Principle; 1992, p. 156).[5] So intentionality and mentality extend from occurrent thoughts to beliefs that are not in the forefront of my mind at the moment and even to repressed desires and other states that belong to the Freudian unconscious. But they do not extend to the processes in the early stages of my visual system or to my tacit knowledge of a generative grammar for my language. According to Searle, the only intentionality that can be attributed to most of the states and processes that figure in cognitive scientific theories is the trivial 'as if' intentionality that they share with 'thirsty' lawns and falling stones that 'want' to reach the centre of the earth.

Those cognitive scientists and commentators who are already committed to a broadly reductionist approach to intentionality are likely to be unmoved by Searle's argument. Indeed, Searle himself allows that the argument is not absolutely compelling. But if we are inclined to stress what is distinctive of persons then we may well judge that there is something right and deep in the vicinity of Searle's connection principle. And, according to Searle, the connection principle reveals a serious flaw in the theoretical foundations of cognitive science. So, granting for the purpose of the argument that Searle is right about the connection principle, we should ask whether this negative consequence for cognitive science really follows.

What surely does follow is that the unconscious states and processes that figure in cognitive scientific theories do not share the kind of intentionality that belongs to conscious thinking and, more generally, to propositional attitudes, whether occurrent, dispositional, or repressed. They do not share what we might call *attitude aboutness*. But, in order to move from this point to a conclusion that threatens the explanatory claims of cognitive science, we would need to make a further assumption along the lines that attitude aboutness is the only kind of non-derivative representationality that escapes the triviality of mere 'as if' intentionality.

Such an assumption would already be rejected by those who, like McDowell and Hornsby, agree verbally with Searle that attributions of representationality at the subpersonal level are only 'as if', while regarding those attributions as responsible, constrained and explanatory rather than as trivial. But we can also challenge the assumption that Searle's argument needs by making it plausible that there are other notions of non-derivative representationality or aboutness that are not 'as if'.

4.3 Indicator aboutness and subdoxastic aboutness

I noted earlier (section 2.2) that the meaning of words, which we can now call *linguistic aboutness*, is a genuine but derived kind of intentionality. When Paul Grice (1989) sets out to explain exactly how linguistic meaning depends on the thoughts of speakers and hearers he calls it 'non-natural meaning' in order to distinguish it from the natural meaning that we attribute when we say, 'Those spots mean measles' or 'Those clouds mean rain'. For the spots mean or indicate measles and the clouds mean or indicate rain quite independently of anyone's thoughts. As Fred Dretske puts it (1986, p. 18):

[5] For further discussion of the Connection Principle and Searle's argument for it, see Searle, 1990; Davies, 1995, pp. 373–81.

'Naturally occurring signs mean something, and they do so without any assistance from us.' This *indicator aboutness* is closely related to the notion of a signal carrying information. It can be explicated in terms of reliable causal covariation between events of two types – for example, between occurrences of a certain kind of cloud formation and occurrences of rain.

Some notion in the vicinity of indicator aboutness seems to lie at the heart of much research in cognitive science and cognitive neuroscience. For example, when a pattern of neuronal activity is found to be reliably correlated with the instantiation of a particular property by objects presented to an animal, that pattern is taken to be the animal's brain's way of representing that property. But, for at least two reasons, indicator aboutness itself is not satisfactory as a kind of representation. First, as Grice points out, it does not allow for the possibility of misrepresentation. Second, it is too cheap; too many things reliably causally covary with each other. So something must be added and, at this point, many of the theories that have been advanced to account for the representational nature of physical states appeal to some notion of teleological function. The basic idea is that what a type of event (such as a pattern of neuronal activity) represents is not the worldly condition that events of that type actually covary with, but rather the condition that those events are supposed to covary with.

In some cases, a type of event has a function because of the intentions of a designer. Thus, consider the familiar example of a fuel gauge. If the states of the fuel gauge reliably covary with the states of the fuel tank, then the position of the needle, towards the bottom of the scale, indicates that the tank is nearly empty. Since the fuel gauge is doing what is was designed to do, the needle's position not only indicates, but also represents, the tank's being nearly empty. But now suppose that the fuel gauge starts to malfunction, the covariation becomes unreliable, and the needle takes up a position towards the bottom of the scale even when the tank is full. Then the position of the needle no longer indicates that the tank is nearly empty. But it does still represent – it misrepresents – the tank as being nearly empty, since this is what the position of the needle is supposed to indicate. Adding a teleological component seems to allow for the possibility of misrepresentation; and it makes representation less ubiquitous than indication. Of course, if teleological function were always to depend on the intentions of a designer then the resulting notion of representation would be derivative from the intentionality of the designer's thoughts. So it could not, after all, contribute towards a response to Searle's critique. But we can move towards a notion of representation that is not conceptually dependent on personal-level mental notions, such as belief and intention, if we consider teleological functions that are the products of natural, rather than intentional, selection.

Stephen Stich proposes that the unconscious representational states that are invoked in cognitive science – states that 'play a role in the proximate causal history of beliefs, though they are not beliefs themselves' (1978, p. 499) – should be called 'subdoxastic states'. Let us extend that terminology and label the putative representationality of those states *subdoxastic aboutness*. This is the notion that is *sub judice* in the context of Searle's critique of cognitive science. Suppose, for a moment, that Searle is right to insist that states with attitude aboutness must be accessible to consciousness. Or suppose, more generally, that personal-level intentionality cannot be literally attributed to subpersonal-level systems. Then it follows that subdoxastic aboutness is different from the aboutness of propositional attitudes like beliefs and intentions. But it does not follow that an attribution of subdoxastic aboutness is really a metaphorical attribution of attitude

aboutness. *A fortiori*, even granting Searle's argument for the connection principle, it does not follow that attributions of aboutness to states that figure in cognitive scientific theories are trivial.

4.4 Personal-level intentionality and subpersonal-level representation

At the end of section 3, I distinguished two ways of resolving an apparent tension in Dennett's position. In this section, I have explored the consequences of resolving the tension in the first way. If we stress what is distinctive about persons and accept that personal-level intentionality cannot be literally attributed to subpersonal-level systems then what are we to make of subpersonal-level representation? The overall picture that emerged from Dennett's appeal to, and ultimate discharge of, homunculi had the personal level of description distinguished from two subpersonal levels. At one of these, the level of information-processing psychology, we use 'as if' personal-level intentional descriptions; at the other, the level of neuroscience, we use literal biological descriptions. But now we see that an alternative picture is available, in which information-processing psychology makes use of descriptions that are literal rather than metaphorical. For it is plausible that a notion of subdoxastic aboutness can be elucidated in terms of causal covariation plus evolutionary function, natural meaning plus natural selection.

Those who adopt the second way of resolving the apparent tension in Dennett's position take a more relaxed view of the distinction between the personal and subpersonal levels of description. They regard the intentionality of people's conscious mental states as of a piece with the representational properties of the states of neural and cognitive systems. So their hope is that attitude aboutness itself can be elucidated in terms of causal covariation and evolutionary function – perhaps with some further elaboration but without appeal to consciousness.[6] As I mentioned earlier, this broadly reductionist approach to intentionality and representation has been the dominant one in both philosophy of mind and cognitive science (Stich and Warfield, 1994).

It would be fair to say, however, that none of the theories proposed within this dominant approach has been accepted as providing a fully satisfying account of the intentionality of ordinary conscious mental states like beliefs (Pietroski, 2000). A critic of the approach, looking for a pattern in the failure, might suggest that the inadequacy of the reductionist approach results from its severing the connection between intentionality and distinctive features of a person's mental life such as consciousness (Searle, 1992). So there might seem to be some pressure towards rejecting the reductionist approach and assigning to cognitive scientific descriptions a metaphorical status at best. But I have been arguing that we do not face the stark choice between adopting the reductionist approach and rejecting the theoretical credentials of cognitive science. An intermediate or hybrid position is available. The broadly reductionist programme may be adequate to account for subpersonal-level representation even if it does not provide a fully satisfying account of personal-level intentionality.

[6] The theoretical resources in a theory of representation may also include the internal functional role of a type of representational state (Block, 1986). This promises a more fine-grained notion of representation than some possible accounts that start from the basic idea of indication. So it might help to capture something of Searle's idea that objects are presented under aspects, but without having to bring consciousness into the account.

5. Inter-disciplinary Relations: Philosophy, Psychology and Physiology Again

By distinguishing between several varieties of aboutness, I have described a position that is intermediate between two extremes. Towards one end of the spectrum there is a broadly reductionist view about intentionality. Towards the other end of the spectrum there is a negative view about cognitive science.

5.1 Philosophy and cognitive science

When we consider the relationship between philosophy of mind and cognitive science more generally, we find a similar pattern. Towards one end of the spectrum there is cognitive scientism – the view that the proper business of the philosophy of mind is simply to hand the substantive questions over to cognitive science. Towards the other end of the spectrum there is philosophical isolationism – the view that, even if cognitive science is not built on a category mistake, it still has little or nothing to contribute to the philosopher's project of plotting the contours of our conceptual scheme. An intermediate position has it that, contrary to cognitive scientism, philosophy makes a distinctive theoretical contribution using a methodology different from that of the empirical sciences while, contrary to philosophical isolationism, philosophical theory cannot be insulated from the findings of empirical research.

These views about the inter-disciplinary relationship go along naturally with views about the relationship between the personal-level descriptions that are of primary interest to philosophy of mind and the subpersonal-level descriptions that figure in cognitive scientific theorising. Cognitive scientism goes with a broadly reductionist view of personal-level descriptions. Philosophical isolationism goes with the view that the relationship between personal-level and subpersonal-level descriptions is one of relative independence.

Once again, an intermediate position is available. As against the reductionist view, cognitive science may well not provide fully satisfying explanatory accounts of personal-level phenomena such as free agency or the intentionality of thoughts. This first aspect of the intermediate position is a generalisation of the familiar idea that cognitive scientific accounts of conscious experience leave an explanatory gap (Levine, 1983). But as against the claim of independence, philosophical theorising may itself reveal that personal-level descriptions, cast in terms of experience, thought, and agency, impose subpersonal-level requirements. When philosophical theorising systematises our conception of ourselves as persons, that conception may turn out to have built into it commitments about what the underpinning information-processing machinery must be like. So the overall picture, according to the intermediate position, is one of downward inferences from the personal level to the subpersonal level, but upward explanatory gaps.[7]

5.2 Cognitive science and neuroscience

The question of the relationship between psychology and physiology, or cognitive science and neuroscience, can be focused on claims about autonomy. Some theorists – for example, Fodor – apparently maintain, while others – for example Paul and Patricia Churchland – deny, that information-processing psychology is autonomous from neurobiology. But there is more than one thing that may be meant by 'autonomous'.

[7] For further discussion of the intermediate view of the inter-level relationship (interaction without reduction), see Davies, 2000a, 2000b.

Fodor's (1974; 1998, p. 9) autonomy claim is a denial of the possibility of inter-theoretic reduction. Psychology cannot, in this sense, be reduced to neuroscience. But an autonomy claim might amount to the rejection, not just of reduction, but of any relation of government or constraint. It is in this sense that the Churchlands (1996, p. 220) deny that psychology is autonomous from neuroscience. The reductionist view that Fodor opposes lies towards one end of a spectrum. The no-constraint view that the Churchlands oppose lies towards the opposite end. Clearly intermediate positions are available (Stone and Davies, 1999).

The strong autonomy claim that psychological theories are not constrained at all by developments in neuroscience is deeply implausible. Some theorists certainly hold that, in inter-disciplinary research, psychology has some kind of priority over neuroscience (Coltheart and Langdon, 1998). But even the strongest claims for the theoretical and practical priority of investigations at the information-processing level over those at the biological level are still consistent with the acknowledgement that psychological theory is, in principle, answerable to neurobiological data (Shallice, 1988, p. 214).

As we move away from the implausible claim of strong autonomy, we allow that cognitive psychology and neuroscience are mutually constraining disciplines. This relationship of mutual constraint is one aspect of what Patricia Churchland speaks of as co-evolution of theories (1986, p. 284): '[T]heories at distinct levels often co-evolve . . . as each informs and corrects the other.' Churchland links co-evolution with reduction (p. 284): '[T]he discoveries and problems of each theory may suggest modifications, developments, and experiments for the other, and thus the two evolve towards a reductive consummation.' But it does not seem obligatory that inter-disciplinary interaction should aim at 'reductive consummation'. It is open to us to join the Churchlands in denying that psychology is autonomous in the second, stronger, sense while still maintaining, with Fodor, that it is autonomous in the first, more modest, sense.

5.3 Persons and their brains

According to the intermediate positions that I have described already, the personal-level descriptions that are so important for philosophy of mind carry commitments that relate to information-processing underpinnings; and cognitive psychology, in turn, is constrained by neuroscience. So findings in neuroscience, as in cognitive science, may impact on philosophy of mind. But inter-level constraint is one thing, and reduction is another. The idea of upward explanatory gaps is no less plausible in the case of personal-level phenomena and neurobiology than in the case of personal-level phenomena and cognitive science. When we consider the inter-level relationship between personal-level descriptions and neurobiological descriptions, or the inter-disciplinary relationship between philosophy and neuroscience, our options once again go beyond the extremes of reduction and independence, or scientism and isolationism.

In *Content and Consciousness*, Dennett already considered the possibility of a purely biological account of the causes of behaviour (1969, p. 78): 'There should be possible some scientific story about synapses, electrical potentials and so forth that would explain, describe and predict all that goes on in the nervous system. If we had such a story we would have in one sense an extensional theory of behaviour.' And, we might add, the true account of personal-level phenomena could scarcely be independent from this scientific story. But, there is a problem with a biological account (p. 79): 'A solely biological, non-Intentional theory of behaviour . . . would be mute on the topic of the actions (as opposed to motions), intentions, beliefs and desires of its subjects.' Switching modalities, we can

say that persons and their thoughts, plans, and actions would not be made visible by the scientific story. A literally true biological subpersonal-level account would not provide a satisfying explanation of personal-level phenomena (Pietroski, 2000).

II An Approach

6. Tacit Knowledge and Psychological Reality

Chomsky's claim that ordinary speakers possess tacit knowledge of a generative grammar for their language stands as the canonical example of appeal to the cognitive unconscious. This extension of mental representation beyond the conscious mind – anticipated by Hamilton and attacked, early and late, by Brentano and Searle – is practically definitive of cognitive science. As I turn from historical and foundational issues to the dominant approach in cognitive science over the last half century, it is natural to begin with Chomsky's work in theoretical linguistics. But philosophical discussion of tacit or implicit knowing (or believing) has ranged over many different notions, some of which are remote from Chomskyan tacit knowledge. So, before describing Chomsky's project I need briefly to consider some of the notions of tacit or implicit knowledge from which Chomsky's must be distinguished.

In a definition of *explicit* knowledge that seems to lend itself to a contrast with tacit knowledge, Michael Dummett says (1991, p. 96): 'Someone has explicit knowledge of something if a statement of it can be elicited from him by suitable enquiry or prompting.' And (p. 97):

> A body of knowledge, however explicit, is obviously not continuously before our consciousness, being a store of items available, save when our memory betrays us, for use when needed. How the storage is effected is of no concern to philosophy: what matters to it is how each item is presented when summoned for use.

So, in Dummett's usage, explicitness is a matter of the subject being able to present information in linguistic form, and is not a matter of how the information is stored in between presentations. Explicit knowledge is *ipso facto* accessible knowledge – 'save when our memory betrays us'.

Suppose that someone knows, in the ordinary sense of that term, the axioms of a theory. Provided that this knowledge can be articulated it counts as explicit knowledge. Now consider some theorem that can be derived from the axioms. The person who knows the axioms may well, with some suitable enquiry and prompting, be able to see that the theorem follows from the axioms and to state it. So at least some of the *as-yet-undrawn consequences* of propositions that are explicitly known are classified by Dummett's account as being explicitly known as well.

In contrast, Dennett says (1983, p. 216):

> [L]et us have it that for information to be represented *implicitly*, we shall mean that it is *implied* logically by something that is stored explicitly.

So, on Dennett's account of the explicit versus implicit distinction, the as-yet-undrawn consequences of propositions that count for Dummett as explicitly known would be classified as implicit knowledge (see also Lycan, 1986, on 'tacit belief'). In fact, relative to any given notion of explicit knowledge or representation it is possible to define a whole family of notions of implicit knowledge or representation where the members of the family differ over the resources that can be used in drawing out consequences. Thus, for example, we might focus on logico-mathematical inferences that can be carried out by the person or system in question; we might allow any valid logico-mathematical inference; we might go beyond strict logico-mathematical inference and consider

deductions that make use of some specified set of background assumptions; we might go further and allow methods of induction or rule extraction, with or without a restriction to methods that are available to the person or system in question.

So far, we have seen that Dummett's 'explicit' overlaps with Dennett's 'implicit'. But how does Dummett himself use the term 'implicit knowledge'? He says (1991, p. 96):

> A piece of *implicit* knowledge may perhaps be attributed to someone who has only implicit grasp of the concepts involved.

And he also suggests (p. 96):

> [W]e may credit the speaker with an implicit knowledge of that rule, provided that, when he understands the statement of it, he acknowledges it as accurately describing his existing practice.

It is not altogether easy to construe these remarks. But it is plausible that Dummett's conditions for explicit knowledge and for implicit knowledge are independent of each other and that neither includes Chomsky's tacit knowledge. First, the conditions for explicit knowledge ('a statement of it can be elicited from him by suitable enquiry or prompting') and for implicit knowledge ('when he understands the statement of it, he acknowledges it as accurately describing his existing practice') seem to be logically independent. It seems to be possible to imagine someone having explicit knowledge of a rule in the sense that he or she could be prompted or coaxed into stating it, yet not being such as to *recognise the rule as correct* if presented with it without any coaxing or coaching. And conversely, it seems possible to imagine someone being able to recognise the rule if presented with it, yet not able to generate a statement of the rule for him- or herself.

Second, it seems clear that Chomsky's tacit knowledge falls outside the bounds of either explicit or implicit knowledge as Dummett construes those notions. Most ordinary speakers lack the concepts that figure in linguistic theory and so will not be in a position – even after suitable enquiry or prompting – to state the principles which, according to Chomsky, they tacitly know. And speakers who learn the technical concepts of linguistics may all too easily fail to recognise tacitly known principles as corresponding to anything in their linguistic practice. Thus, Chomsky says (1986, p. 269):

> [T]here is no way for one to determine by introspection that the rules and principles hold. One cannot become aware that one knows, or cognizes, these rules and principles. If presented with these principles as part of a theory of grammar, we may become convinced that they are correct, but we do so 'from the outside', as we may be convinced that a theory of fusion correctly explains the emission of light from the sun.

Chomsky's tacit knowledge is not Dummett's implicit knowledge (indeed, it falls outside Dummett's classificatory scheme) and it does not correspond to implicit knowledge in Dennett's sense either. For implicit knowledge is attributed on the basis of what could be derived from a stock of knowledge already certified as explicit. Commitment to these as-yet-undrawn logical consequences is already implicit in the thinker's commitment to the explicit beliefs and if the consequences were actually to be drawn then the resulting beliefs would be causally downstream from the already explicit beliefs. In contrast, tacit knowledge is supposed to be causally antecedent to a battery of more everyday pieces of knowledge. States of tacit knowledge are, in Stich's (1978) terminology, subdoxastic states. That is, they are states that can be described in

representational or semantic terms and that 'play a role in the proximate causal history of beliefs, though they are not beliefs themselves'.

6.1 Chomsky's project

A grammar is, in the first instance, something abstract; it is a collection of rules that generate a set of structural descriptions of sentences. For example, a grammar might include phrase structure rules such as the rule, S → NP VP, which says that a sentence can be made up of a noun phrase plus a verb phrase. Taken together with a lexicon specifying that 'Charlene' is a noun and 'sleeps' is a verb, and rules saying that a noun by itself is a noun phrase, and a verb by itself is a verb phrase, the phrase structure rules license a structural description $_S[_{NP}[\text{Charlene}]_{VP}[\text{sleeps}]]$. Early generative grammars were made up of a base component – lexicon plus phrase structure rules – and a transformational component, where a transformation is a rule that maps a whole structural description to another structural description in which some of the constituents have been moved (Chomsky, 1965).

Modern generative grammars take a rather different form. But the present point is that, while a grammar is a collection of rules, a specification of a grammar may be put forward as a partial account of what a particular language user (tacitly) knows; that is, as a partial psychological description of that language user. Thus, in recent versions, statements of a grammar are said to be about I-language (internalised language) conceived as the attained state of the language faculty (Chomsky, 1986, pp. 22–3; 1995, p. 18).

The aim of the theoretical linguist is to provide a correct account of the body of specifically linguistic tacit knowledge possessed by a language user. This body of knowledge is the speaker's *competence*, and linguistic competence is contrasted with linguistic *performance*, which is the actual use of language – and so the use of this body of knowledge – in concrete situations. Performance includes using a sentence to say something, interpreting someone else's utterance, or silently thinking in words. It also includes making judgements about whether a sentence is grammatical, whether a sentence is ambiguous, and so on. In language use, linguistic knowledge is drawn on by cognitive processes whose operation is subject to all the usual constraints of time and space and to the possibility of malfunction. Also, the use of language in concrete situations draws on knowledge that falls outside the domain of theoretical linguistics, including knowledge about people's beliefs and other mental states, knowledge about conversational practices, and – as the interpretation of metaphors illustrates – an open-ended mass of knowledge about the workings of the non-linguistic world. So the relationship between competence and performance may be relatively indirect consistently with the idea that competence is somehow drawn on in performance.

Whether a sentence is *grammatical* is a question about competence. It is a question about which structural descriptions are generated by a tacitly known collection of rules. But whether a sentence is *judged* by language users to be grammatical, or is judged to be *acceptable* or to sound right, is a question about performance. Clearly, there could be countless reasons why a sentence that is grammatical might be judged – reflectively or unreflectively – to be unacceptable. There could be countless reasons why a sentence that is ambiguous might be judged to have just one meaning or why a sentence that literally means one thing might be interpreted as meaning something quite different. It would involve multiple misreadings of Chomsky to think that the aim of the theoretical linguist

is to produce a smoothed-out systematisation of unreflective judgements by language users (see Quine, 1953; Stich, 1972; Fodor, 1981).

If a grammar provides a correct description of a language user's tacit knowledge then the grammar is said to be *descriptively adequate*. So, the aim of the theoretical linguist is to specify descriptively adequate grammars; that is, true accounts of the I-language – the attained state of the language faculty – of individual speakers. But there is then a pressing question how the attained state of tacit knowledge is arrived at. Chomsky's view here is that it is not possible to explain this attainment without postulating a substantial body of tacit knowledge as an innate endowment; that is, as the initial state of the language faculty.

In early versions of the theory, the innate endowment was supposed to be a body of knowledge about the forms that rules, and so grammars, could take; it was an innate linguistic theory. Any particular hypothesis about this innate linguistic theory was subject to two kinds of constraint. It was constrained from above, so to speak, by the formal requirement that the linguistic theory, together with information available in the environment of the young child learning a language, should correctly determine a single grammar as the grammar of which the child would acquire tacit knowledge. A linguistic theory meeting this formal requirement was said to be *explanatorily adequate*. But, in addition to the requirement of 'adequacy-in-principle', there was also a constraint from below; namely that the processes required to effect the transition from innate endowment to attained state should be 'feasible' given the 'constraints of time and access', as Chomsky (1965, p. 54) puts it. Of this constraint from below, Chomsky says (p. 62): 'This requirement of "feasibility" is the major empirical constraint on a theory, once the conditions of descriptive and explanatory adequacy are met.' Since the early days of generative grammar, many of the developments in Chomsky's view, both of the innate endowment and of the attained state of tacit knowledge, have been driven by this feasibility constraint (see e.g. Chomsky, 1986, p. 55; 1995, p. 7).

6.2 Psychological reality

The project of theoretical linguistics is not committed to any specific views about the way in which linguistic knowledge (competence) is drawn on in language use (performance). But it is, at least initially, an attractive hypothesis that the relationship between competence and performance should be as direct as the limitations of space and time and the role of non-linguistic knowledge permit. The idea would be that the processing of a heard or seen sentence should track the derivation of the sentence's structural description in the tacitly known grammar sufficiently closely that complexity of processing should vary with the number of rules involved in the derivation. This is the *derivational theory of complexity* (Fodor, Bever and Garrett, 1974, p. 320; see also Smith, 2004, pp. 110–2).

In early versions of generative grammar the derivation of a passive sentence involved, first, the derivation (from the lexicon plus phrase structure rules) of an active sentence and, then, the derivation of the passive form from the active form (underwritten by the passive transformation rule). So one way to evaluate the derivational theory of complexity was to investigate the question whether passive sentences are more difficult to process than the corresponding active forms. For example, do subjects take longer to assess the truth or falsity of passive sentences ('13 is preceded by 7') than active sentences ('7 precedes 13')? While some experimental results were consistent with the derivational theory of complexity, others were not.

One example of a problematic result makes use of the distinction between *reversible* and *irreversible passives*. The passive sentence 'The lion was chased by the tiger' makes equally good sense if the two noun phrases are reversed. But the passive sentence 'The ice cream was eaten by the boy' does not make equally good sense if the noun phrases are reversed. In an experiment where subjects were asked whether a picture is correctly described by a sentence, responses to reversible passives were slower than to corresponding active sentences – as predicted by the derivational theory of complexity. But the effect was not present for irreversible passives. So, it was concluded, the results that were apparently favourable to the derivational theory of complexity could not be regarded as revealing additional processing complexity that was attributable to the presence of the passive construction – and so the passive transformation – as such. Rather, it was suggested, the additional time taken to process some passive sentences was a reflection of the subjects' adoption of a processing strategy; namely, initially treating the first noun phrase as the subject (referring to the agent) of the following verb. This flawed strategy could be rapidly abandoned in the light of non-linguistic knowledge in the case of an irreversible passive, such as 'The ice cream was eaten by the boy', since ice creams do not eat things. But the strategy would be abandoned only later in the case of a reversible passive.

Overall, the derivational theory of complexity was not regarded as supported by the experimental results and Jerry Fodor, Tom Bever and Merrill Garrett (1974, p. 323) concluded that 'we must postulate a far more abstract relation between the grammar and the sentence recognizer'. Some psycholinguists regarded the *psychological reality* of a grammar as consisting in there being a very close relationship between sentence processing and the formal derivations of structural descriptions. Given that use of the term, Fodor, Bever and Garrett's conclusion could be reported as saying that grammars are not psychologically real. But the terminology is apt to be misleading. It would certainly not be right to regard the experimental results as somehow in conflict with the basic idea that the theoretical linguist is making empirical psychological claims.

Over the last thirty years, there have been many changes in theories about tacitly known grammars and in theories about sentence processing. (For a review of research on sentence processing, see J.D. Fodor, 1995.) But Chomsky's project – now as then – allows for the possibility that tacit knowledge of linguistic rules is very directly implicated in sentence processing and also for the possibility that there is a quite indirect ('abstract') relationship between the body of linguistic knowledge (competence) and the processes involved in language use (performance).

7. Tacit Knowledge Challenged and Defended

There are disputes in cognitive science about whether we need to attribute tacit knowledge of linguistic rules to language-users and especially about the claim that a substantial body of linguistic knowledge must be regarded as innate. But, within cognitive science, it is not denied that the notion of tacit knowledge is a legitimate one. Even someone who claims that there is no need to appeal to tacit knowledge of linguistic rules in order to account for sentence processing, or no need to appeal to innate knowledge in order to account for language acquisition, allows that the notion of tacit knowledge makes sense. But within philosophy the attitude towards tacit knowledge has been much more critical. I shall consider two kinds of criticism that go beyond general resistance to the idea of the cognitive unconscious.

7.1 Wittgensteinian worries and Quine's challenge

There are criticisms of Chomsky's project that arise from the later philosophy of Ludwig Wittgenstein. For example, a sense that there is something wrong with the appeal to tacit knowledge might be based on a dominant theme in Wittgenstein's philosophy, namely that, at some point, justifications and explanations in terms of reasons give out. It might be thought that Chomsky commits a double philosophical error by first seeking a justification where none is needed and then postulating a kind of justification by unconscious consultation of linguistic rules. But while this would indeed be a serious error, it is not an error that Chomsky makes.

A similar line of criticism might be developed from the thought that the notion of a rule of language belongs with the idea of a normative practice. Rules determine (logically rather than causally) what is correct or incorrect and participants in the practice advert to rules to justify, criticise or excuse their actions. The proper response to this line of criticism is to distinguish the notion of a tacitly known rule from the notion of a rule that figures in a normative practice. Ordinary language users do not, of course, advert to tacitly known rules to justify their judgements about whether a sentence is grammatical or what a sentence means. Rules of language as Chomsky conceives them 'are not normative in this sense' (2000, p. 98). But states of tacit knowledge of rules can still play a causal-explanatory role in accounting for the particular pieces of knowledge expressed in those judgements.

In a wide-ranging Wittgensteinian critique of Chomsky's project, Gordon Baker and Peter Hacker (1984) anticipate the response on behalf of Chomsky that tacit knowledge of a rule is constituted by the presence of a mechanism that ensures conformity to the rule. About this idea they raise two important questions. One is the question how, given this kind of account of tacit knowledge, we can distinguish between tacit *guidance* by a rule and mere *conformity* to the rule (Baker and Hacker, 1984, p. 298). This, in essence, is the challenge that W.V.O. Quine (1972) posed. I shall review Quine's challenge now and return to Baker and Hacker's second question later (see below, section 7.3).

A subject can behave in a way that conforms to a rule without using the rule to guide his behaviour for, as Quine uses the notion of guidance, it requires explicit verbalisable knowledge. Chomsky's tacit knowledge is supposed to be an intermediate notion. While it requires less than explicit knowledge, it cannot be equated with mere conformity. In fact, conformity to rules is neither necessary nor sufficient for tacit knowledge of those rules. It is not necessary, since the presence of tacit knowledge of rules does not guarantee perfect deployment of that knowledge in actual performance. It is not sufficient, since a tacit knowledge claim is not offered as a summary description of behaviour but as a putative explanation of behaviour. In principle, there will always be alternative sets of rules that require just the same behaviour for conformity – that are 'extensionally equivalent' in Quine's terminology.

The notion of extensional equivalence as introduced by Quine is understood in terms of rules allowing strings of words as well-formed sentences. In fact, the notion of a well-formed string plays no theoretical role in Chomsky's project. But Chomsky does allow, at least as a theoretical possibility, that two sets of rules, R and R´, might generate the same set of structural descriptions (1995, p. 15): 'Two distinct I-languages might, in principle, have the same structure [that is, might generate the same set of structural descriptions], though as a matter of empirical fact, human language may happen not to permit this option.' In such a case, R and R´ would be extensionally equivalent. Also, two sets of rules generating different sets of structural descriptions could be extensionally equivalent

because they allow the same strings but assign different structural descriptions to some strings. For example, R and R′ might assign the following different structural descriptions to the same string 'Charlene ate the carrot':

$_{S}[_{NP}$[Charlene]$_{VP}[_{V}$[ate]$_{NP}$[the carrot]]

$_{S}[_{NP}$[Charlene]$_{V}$[ate]$_{NP}$[the carrot]]

In short, Chomsky's project clearly allows for the possibility that a speaker's linguistic performance is correctly explained in terms of tacit knowledge of R and not correctly explained in terms of tacit knowledge of R′, even though just the same judgements about grammaticality conform to the rules of R as conform to the rules of R′.

It is at this point that Quine poses his challenge (1972, p. 444):

> Implicit guidance is a moot enough idea to demand some explicit methodology. If it is to make sense to say that a native was implicitly guided by one system of rules and not by another extensionally equivalent system, this sense must link up somehow with the native's dispositions to behave in observable ways in observable circumstances. These dispositions must go beyond the mere attesting to the well-formedness of strings, since extensionally equivalent rules are indistinguishable on that score.

He insists that, if an attribution of tacit knowledge is an empirical claim that goes beyond a summary of conforming behaviour, then it should be possible to indicate what kinds of evidence would count in favour of or against that empirical claim. Quine also insists that this evidence should involve the subject's behaviour. To this latter point, it is reasonable to reply that there can be no *a priori* limit on the kinds of evidence that might be relevant to an empirical claim. So it is not legitimate to restrict evidence to the behaviour of the very subject to whom the attribution of tacit knowledge is being made. But the more general point about evidence can be allowed as a fair one and we may even be able to go some way towards meeting Quine's challenge on its own terms.

More fundamental, however, than the question of what evidence would support an attribution of tacit knowledge is the question what the correctness of such an attribution would consist in.

7.2 Evans's response and an account of tacit knowledge

While Chomsky himself focuses on tacit knowledge of syntax, much of the philosophical literature considers tacit knowledge of semantic theories such as truth-conditional theories of meaning. Responding to a version of Quine's challenge for the case of tacit knowledge of a semantic theory (Wright, 1981), Gareth Evans proposed a constitutive account of tacit knowledge in terms of dispositions (1981, p. 124):

> I suggest that we construe the claim that someone tacitly knows a theory of meaning as ascribing to that person a set of dispositions – one corresponding to each of the expressions for which the theory provides a distinct axiom.

For this purpose, it is vital that attributing a disposition to a person is not just describing a regularity in that person's behaviour.

Thus, consider a semantic theory that includes an axiom saying that a particular name '*a*' refers to the dog Fido and another axiom saying that a particular predicate 'F' is satisfied by things that bark. Tacit knowledge of this semantic theory requires having a disposition corresponding to '*a*', and a disposition corresponding to 'F', and so on for each of the other expressions for which the theory includes an axiom. But having these

dispositions is not just a matter of treating each sentence containing '*a*' as meaning something about Fido and treating each sentence containing 'F' as meaning something about barking. For someone might exhibit those regularities as a result of learning the meanings of a large number of complete sentences from a phrasebook, yet without having any sensitivity to the way in which the sentences are built up from expressions like '*a*' and 'F'. Rather, says Evans (1981, p. 125), an attribution of tacit knowledge of the semantic theory 'involves the claim that there is a single state of the subject which figures in a causal explanation of why he reacts in this regular way to all the sentences containing the expression'. Tacit knowledge of the semantic theory requires that all the instances of the regularity involving sentences containing '*a*' should have a common causal explanation, all the instances of the regularity involving sentences containing 'F' should have a common causal explanation, and so on. This requirement would *not* be met in the case of someone who had acquired only phrasebook knowledge. For in a case of phrasebook knowledge, the explanations of the subject's reactions to different sentences may be quite separate. The explanation of why the subject treats 'F*a*' as meaning that Fido is barking need have nothing in common with the explanation of the subject's reaction to another sentence containing '*a*' or with the explanation of the subject's reaction to another sentence containing 'F'.

According to Evans's account, tacit knowledge of a semantic theory requires the presence of a battery of causal-explanatory states, one state corresponding to each axiom of the theory. Each such state functions as a causal common factor in explaining instances of a pattern in language use, just as the axiom to which it corresponds figures as a derivational common factor in the proofs of theorems that specify the meanings of complete sentences in which a particular expression occurs. Different batteries of states might causally yield the same language use just as different sets of axioms might derivationally yield the same theorems. But this fact does not point to a problem for the notion of tacit knowledge. An attribution of tacit knowledge is not a mere summary of performance; it is offered as part of an explanation of performance.

Quine's challenge highlights the fact that evidence from patterns in performance may not itself help us decide between competing attributions of tacit knowledge. As the example of phrasebook knowledge illustrates, a pattern of performance on the task of assigning meanings to sentences might be shared by one subject who possesses, and another subject who lacks, tacit knowledge of a semantic theory with separate axioms for '*a*', 'F', and so on. But given the kind of constitutive account just sketched, we can readily imagine empirical evidence that would confirm, and other evidence that would disconfirm, an attribution of tacit knowledge of one or another syntactic or semantic theory. Furthermore, much of this evidence would be manifested in the behaviour of the very subject to whom the attribution of tacit knowledge was being made. So we can go some way towards meeting Quine's challenge, even on its own terms.

Evans mentions three broad families of evidence that would certainly be relevant to attributions of tacit knowledge: evidence from language acquisition, evidence from sentence perception, and evidence from loss of linguistic abilities. Many examples of these three kinds of evidence could be provided from studies of the normal course of language development, from experiments on normal adult language processing, and from descriptions of language impairments following brain injury. To these families of evidence we could also add evidence from neural imaging.

In the case of sentence perception, Evans says (1981, pp. 128–9):

> There is a clear difference between perceiving a sentence which does in fact contain the expression *a*, and perceiving a sentence *as* containing the expression *a*. Consequently, we can regard as relevant to the decision between the two models [one involving tacit knowledge of a compositional semantic theory, the treating sentences as unstructured] the various psychological tests which have been devised for identifying perceived acoustic structure, for example, the click test originally devised by Ladefoged and Broadbent [1960].

The purpose of the Click Test was to investigate the constituent structure of sentences by measuring the perceived displacement of the position of clicks occurring during the auditory presentation of a sentence (for reviews, see Fodor, Bever and Garrett, 1974; Smith, 2004, pp. 101–7). For example, the following two sentences have very different constituent structures:

> As a direct result of their new invention's influence †the company was given an award.
>
> The retiring chairman whose methods still greatly influence the company †was given an award.

When the sentences were presented auditorily, with the presentation acoustically identical from 'influence' to the end of each sentence, interspersed clicks were heard as displaced towards the major constituent boundary indicated by the dagger (†). Results of this kind were taken as validating the click test, showing that constituent structure has some measurable impact on the perception of sentences in on-line processing. However, subsequent attempts to use the click test to provide evidence about constituent structure in cases that were contested within linguistic theory were not successful and 'no one does click experiments any more' (Smith, 2004, p. 107; see also Chomsky, 2002, pp. 125–7).

Variants of Evans's constitutive account of tacit knowledge have been cast in terms of loci of systematic revision (Davies, 1981a, 1981b), the role of certain states in differential explanation (Davies, 1981b, p. 160), or a relation of isomorphism between a pattern of causal-explanatory structure in a processing system and a pattern of derivational structure in a semantic theory (Davies, 1986, 1987). Closely related accounts have also been cast in terms of a causal notion of an algorithm or mechanism drawing on the information carried by a state (Peacocke, 1986, 1989; see Miller, 1997, for a review). There is no guarantee, of course, that any one of these versions will, in all its details, suit the purposes of cognitive scientists. But the coherence of these accounts strongly suggests that there is nothing conceptually wrong with Chomsky's invocation of tacit knowledge in linguistic theory.

7.3 Wittgenstein again: The rule-following considerations

We should not, however, leave the issue without returning to Baker and Hacker's critique and, in particular, to the second of the two questions that they raise about accounts of tacit knowledge cast in terms of mechanisms whose presence causally explains conformity to a rule. Although tacitly known rules do not play a justificatory role, the body of tacit knowledge (competence) determines grammaticality. But language users make mistakes and mechanisms malfunction. So competence is not perfectly reflected in performance; for example, grammaticality is not perfectly reflected in judgements of acceptability. The worry, then, is that, to give an account of tacit knowledge in terms of a mechanism's explanation of patterns in performance, we need some account of what

would constitute an error in the operation of the mechanism. For otherwise, every feature of performance will have to be reflected in the rules that are tacitly known.

Here, with an apparent threat to the distinction between something's seeming right and being right (Wittgenstein, 1953, §202), we are in the vicinity of Wittgenstein's rule-following considerations, especially as they are developed by Kripke (1982; see also Wright, 1981, 1989; Boghossian, 1989; Miller and Wright, 2002). Indeed, it might be thought that the Kripke-Wittgenstein problem arises in a particularly acute way for the notion of tacit knowledge. For many commentators believe that a response to the problem must involve appeal to a community; yet Chomsky's notion of tacit knowledge is supposed to be applicable to an individual without reference to a community. But, however this may be, we should certainly acknowledge that a fully satisfying account of tacit knowledge must make room for the fact that there is a gap between seeming right and being right – that performance is an imperfect guide to competence.

8. Tacit Knowledge in Philosophy of Language

I begin this section with a review of the main ideas in the constitutive account of tacit knowledge (section 7.2). The example to be used does not involve rules of syntax or semantics, but very simple letter-sound rules of the kind that could be employed in reading words aloud.

Suppose that one of these rules states that if a word begins with the letter 'B' then its pronunciation begins with the sound /B/. If a subject's pronunciation behaviour *conforms* to this rule then it displays a pattern. Whenever a presented word begins with 'B', the subject's pronunciation begins with /B/. But the transitions that concern us are not these transitions from presentation to pronunciation. Rather, we focus on transitions amongst beliefs or states of information. Whenever the subject starts out with the information that the presented word begins with 'B', the subject ends up with the information that the word's pronunciation begins with /B/.

If these states were beliefs, then the subject's pattern of transitions from belief state to belief state would be accounted for if the subject possessed explicit knowledge of the 'B'-to-/B/ rule. This piece of explicit knowledge would figure as a common factor in the causal explanations of the subject's transitions from belief to belief. In contrast, there would be no such common factor if the subject had merely memorised the pronunciation of each of a large number of words beginning with 'B'. The difference between having explicit knowledge of the rule and having an independent piece of knowledge for each of the instances that fall under the rule corresponds to a difference in causal-explanatory structure.

In fact, we should not assume that the subject has beliefs about words and their pronunciations; the transitions may involve states of the kind that figure in information-processing psychology. Nor should we assume that the subject either has explicit knowledge of pronunciation rules or else has explicitly memorised the pronunciations of words. But we can still make use of the idea of causal-explanatory structure and, in particular, the idea of a common factor in the causal explanations of transitions. An attribution of tacit knowledge of the 'B'-to-/B/ rule can be construed as the claim that there is a single state of the subject that figures in the causal explanations of the various particular transitions that the subject makes from input representations of words as beginning with 'B' to output representations of pronunciations as beginning with /B/ (Evans, 1981).

In general, a state of tacit knowledge is a state that figures as a common factor in causal explanations of certain transitions amongst states of information (or beliefs). Tacit knowledge of a rule requires more than just conformity to the rule. (In fact, strictly speaking conformity to a rule is neither necessary nor sufficient for tacit knowledge of the rule.) There are always different sets of rules that require the same transitions for conformity. But the attribution to a subject of tacit knowledge of a particular set of rules is made correct by the presence of a particular structure in the causal explanations of the subject's rule-conforming transitions.

Once an account of tacit knowledge in terms of causal-explanatory structure has been given, it is a relatively straightforward matter to give examples of empirical evidence that would confirm the attribution to a subject of tacit knowledge of a particular set of rules such as a grammar or a semantic theory. Indeed, some of this evidence meets Quine's (1972) additional requirement of concerning the behaviour of the subject to whom the attribution of tacit knowledge is being made.

Earlier (section 5.1), I outlined an account of the relationship between philosophy and cognitive science that is intermediate between cognitive scientism and philosophical isolationism. This view of the inter-disciplinary relationship goes naturally with a view about the relationship between the personal-level descriptions that are of primary interest to philosophy and the subpersonal-level descriptions that figure in cognitive scientific theorising. According to that view, the inter-level relationship is one of downward inferences from the personal level to the subpersonal level, but also of upward explanatory gaps. I now consider, specifically, the relationship between personal-level descriptions cast in terms of linguistic understanding or knowledge of meaning and subpersonal-level descriptions cast in terms of tacit knowledge of semantic theories.

8.1 Compositionality and meaning without use: A downward inference

It is one thing to show that the notion of tacit knowledge is conceptually in good order; it is another thing to show that it is important for philosophy. The most natural point of contact is provided by philosophical work on compositional theories of meaning – theories that show how the meanings of whole sentences depend on the meanings of their parts. But many philosophers of language maintain that philosophy prescinds from questions about actual psychological processes and concerns itself instead with more abstract or normative conceptions of linguistic structure. Thus, for example, we find Dummett saying (1991, p. 92):

> A meaning-theory should not . . . aspire to be a theory giving a *causal* account of linguistic utterances, in which human beings figure as natural objects, making and reacting to vocal sounds and marks on paper in accordance with certain natural laws.

It may be suggested that the philosopher of language cast in the role of semantic theorist can simply choose whether to have more or less involvement with empirical research. He or she might aim to capture what actual speakers tacitly know about the semantic structure of their language or might – no less well – opt for the project of systematising semantic knowledge along the lines of a hypothetical ideally rational subject. However, we can argue that it is not so easy for the philosopher of language to opt for isolation from cognitive science.

Suppose that I understand on first hearing a sentence, S, of my language, built from familiar words in familiar ways. It is part of our ordinary conception of linguistic

meaning that it is not my hearing the sentence S that imbues it with meaning. A sentence that I never hear or use may have a determinate meaning in my language. So if we suppose, not that I hear S and understand it, but that I never hear, or use, or even think about it, then still S may have a determinate meaning in my language. What could determine this meaning? This is the question posed by *the problem of meaning without use*.

The natural answer to this question is that the meaning of S in my language is fixed by some kind of extrapolation from the meanings of the sentences that I do use. But there are two ways to develop this answer. On the first way, the projection of meaning from used sentences to unused sentences follows the contours of the semantic theory that I tacitly know. On the second way, the projection of meaning follows the inductive, abductive, and deductive reasoning of a hypothetical ideally rational subject.

Indeed, quite generally, the project of constructing *compositional* semantic theories can be constrained in either of two ways. The common starting point is that we should not commit ourselves to the idea that speakers of the language actually *know* (in the ordinary sense of that term) the facts stated by the axioms of a compositional semantic theory. Attributing such knowledge would over-intellectualise ordinary language use. From that starting point, one possible move is to appeal, not to ordinary knowledge, but to *tacit knowledge*. Another possible move is to appeal, not to knowledge that speakers actually possess, but to *knowledge that would suffice* for understanding the language. Thus, the first kind of constraint on compositional semantic theories says that the axioms of the theory should be tacitly known by the speaker or speakers whose language is under investigation. The derivational structure of the theory should *mirror* the causal-explanatory structure in those speakers. The second kind of constraint says that a semantic theory for a language should display the compositional structure that is present in the language, and should display it as structure that could be used by an idealised rational subject. (See also the discussion of the Mirror Constraint and the Structural Constraint on semantic theories in Davies, 1981a, chapter 3, and Wright, 1987.)

A compositional semantic theory meeting the second kind of constraint shows how knowledge of meaning is possible. It shows how systematic mastery of a language – a mastery marked by the ability to move from understanding of some sentences to understanding of others (in principle, of infinitely many others) – could be a rational achievement. But it does not bring with it any account of how ordinary speakers actually arrive at their knowledge of the meanings of hitherto unused sentences. If, however, compositional semantic theories are subject to the first kind of constraint than an account of the epistemology of understanding is naturally suggested. By the requirement of tacit knowledge, there are cognitive structures and processes that determine the meaning of unused sentences. So, we may suppose, when a hitherto unused sentence is heard for the first time, those same cognitive structures and processes whose presence has provided the sentence with its meaning come into play to underpin the speaker's assignment to the sentence of that very meaning.

This epistemological difference between the two approaches to compositionality might already count in favour of the first approach, with its appeal to tacit knowledge. But there is also a second difference that is more metaphysical than epistemological. In one kind of case, the two ways of conceiving meaning for unused sentences deliver importantly different results.

Suppose that I learn the meanings of some sentences from a phrasebook and that I remain blind to the semantic structure that an ideally rational subject would see in those

sentences. The semantic theory that I tacitly know is not a compositional theory but a theory with a separate axiom specifying the meaning of each of the sentences that I looked up in the phrasebook. It does not determine any meaning at all for sentences that fall outside my corpus. So the first way of developing the idea of extrapolating meanings says that no unused sentence has a determinate meaning in my language. However, an ideally rational subject would see patterns to which I am blind and would be able to assign determinate meanings to some sentences that I never looked up. So the second way of developing the idea of extrapolating meanings says that some unused sentences do have determinate meanings in my language.

Some philosophers of language argue that the second way is definitely wrong. They say that, if someone has mere phrasebook knowledge of some sentences, then we should not attribute to that person a language in which additional sentences, of which the person knows nothing, have determinate meanings (Schiffer, 1993). If these philosophers of language are right then we should take the first way and we should agree with Brian Loar when he says (1981, p. 259) that 'the Chomskyan idea of the internalization [tacit knowledge] of the generative procedures of a grammar has got to be invoked to . . . make sense of literal meaning'.

8.2 Dummett on understanding: An upward explanatory gap

If the argument that I have just sketched is correct then philosophical theorising may itself reveal that personal-level descriptions cast in terms of linguistic understanding impose subpersonal-level requirements of tacit knowledge. This is an example of a downward inference from the personal level to the subpersonal level of information-processing psychology. However, it does not follow that the notion of tacit knowledge can provide a fully satisfying explanatory account of linguistic understanding. I now turn to considerations that suggest an upward explanatory gap.

In his early paper, 'What is a theory of meaning?', Dummett remarks (1975, p. 112):

> It is one of the merits of a theory of meaning which represents mastery of a language as the knowledge not of isolated, but of deductively connected, propositions, that is makes due acknowledgement of the undoubted fact that *a process of derivation* of some kind is involved in the understanding of a sentence.

However, as we have seen, this remark is not to be interpreted as favouring the idea that semantic theories are causal theories about human beings as 'natural objects'. Dummett also says (1991, p. 103): '[A] meaning-theory aims at providing, not a faithful representation of a speaker's linguistic knowledge, but a systematisation of it.' It is not immediately clear what the distinction between 'faithful representation' and 'systematisation' amounts to and, consequently, it is not immediately clear what attitude Dummett takes towards empirical questions about what actual speakers actually know or about how that knowledge is put to use. But one thing that is clear is that, in Dummett's view, it is a mistake to think that all that is required of a philosopher of language is to state 'what it is that a speaker knows' (p. 105). For a philosophical theory about meaning must also 'explain . . . what counts as a manifestation of [the speaker's] linguistic knowledge. This may be vividly expressed as the requirement that we say in what that knowledge consists' (p. 104).

For a given language, L, and a given individual, whether that individual meets the requirements for being an L-speaker is clearly an empirical question. But the question of what those requirements are is a philosophical or constitutive question. It is also a prior

question; in order even to ask the empirical question, we need already to have answered the constitutive one. So, part of Dummett's point is that an account of knowledge of language, or understanding, must include a component that is not empirical and causal but philosophical and constitutive.

To acknowledge the importance of this philosophical question is not, of course, to deny that there must be some empirical account of how language users meet the requirements that a constitutive account of understanding imposes on them. Indeed, Dummett considers, and does not rule out, the possibility that Chomsky's notion of tacit knowledge may figure in this empirical account of how the requirements for understanding are met (pp. 96–7):

> For [Chomsky], a speaker's competence consists in his knowing a complete syntactical and semantic theory . . . unconsciously; even presentation of an explicit statement of its contents may well not serve to bring this knowledge to consciousness. Chomsky puts this forward not as a philosophical explanation but as a psychological hypothesis; and it is as such that it must be evaluated.

But Dummett also notes that an appeal to tacit knowledge does not itself yield any insight into the way in which linguistic knowledge is presented to us when it is actually deployed in use (p. 97): 'The important question about a body of knowledge possessed by a subject is, however, the form in which it is delivered, and of this Chomsky tells us little. . . . When we ask in what kind of knowledge our understanding of our language consists, we are asking in what form it is delivered.'

There is more than one point that Dummett is making here – and so more than one issue that might be identified as the *delivery problem*.[8] One point is that an empirical, causal theory (a psychological hypothesis) does not answer the questions that a philosophical, constitutive theory about meaning and understanding seeks to address. Correspondingly, the delivery problem could be identified as the basic philosophical question about the nature of understanding: In what does understanding consist?

But there is a second point that is at least suggested by Dummett's comments – a point that would not turn so simply on distinguishing empirical questions from constitutive questions. For suppose that we had a philosophical account of the requirements for being an L-speaker, and that our interest was in the question how a person can meet those requirements. Dummett seems to allow that an empirical cognitive scientific theory could make some contribution to answering that question. But there is an intuition that cognitive science would not provide a wholly satisfying answer to the question how the requirements for understanding are met. On the basis of this second point, the delivery problem could be identified as the question: How do subpersonal-level computational processes deliver the personal-level phenomenon of linguistic understanding? The idea that there is an upward explanatory gap here would go along with Dummett's own stress on the fact that the use of language is a 'conscious rational activity' (1991, p. 91).[9]

---

[8] For Chomsky's response to what Dummett says here, see Chomsky (1995, p. 34).

[9] A further reason why, on Dummett's account, there would be an explanatory gap here is that the information-processing story would contribute to an account of 'a language as known by a single individual'. But, according to Dummett (1991, p. 106): 'If we isolate [the individual] in thought from his society, there ceases to be any right or wrong in his use of his personal language; and consequently all meaning evaporates from it.'

9. The Language of Thought Hypothesis

As I shall understand it here, Fodor's language of thought hypothesis is a hypothesis about internal representational states that figure in subpersonal-level psychological structures and processes. Many of these representational states are clear examples of subdoxastic states and of the cognitive unconscious. But the language of thought hypothesis is, of course, also supposed to apply to states of thinking and, in particular, to occurrent thoughts – states or events that are neither subdoxastic nor unconscious. It may be tempting to suppose that, when it is thoughts that are at issue, the hypothesis is an answer to the question whether people think 'in language' – whether, in conscious thinking, sentences of one's natural language come silently before the mind. But, in fact, the language of thought hypothesis does not concern the phenomenology of conscious thinking and so is quite distinct from the 'thinking in natural language' hypothesis (Carruthers, 1996, 1998; Fodor, 1998, pp. 63–74).

9.1 Intentional realism and syntactic properties

We can begin from the assumption that personal-level events of conscious thought are underpinned by occurrences of physical configurations belonging to types that figure in the science of information-processing psychology. These physical configurations can be assigned the contents of the thoughts that they underpin. So we assume that, if a person consciously or occurrently thinks that *p*, then there is a state that has the representational content that *p* and is of a type that can figure in subpersonal-level psychological structures and processes. This assumption is what Fodor (1985, 1987) calls *intentional realism*. We do not assume that the properties of these underpinning states, other than their representational contents, are evident to the thinking subject's introspection. Then the language of thought hypothesis says, first, that these states that underpin thoughts have *syntactic properties* and, second, that the same goes for other states in the domain of information-processing psychology.

Fodor (1987, pp. 16–21) imposes three conditions on syntactic properties. First, a syntactic property is a physical property (though this is not intended to require that a syntactic property should be a property that figures in fundamental physics). Second, a syntactic property is correlated with a semantic property. Third, a syntactic property is a determinant of causal powers and so of causal consequences. Fodor says that shape, which is an intrinsic property, is the right sort of property to be a syntactic property. So we can take it that the three conditions are intended to have the consequence that a syntactic property is an intrinsic property of a representation. This helps to explain why semantic or representational properties themselves do not qualify as syntactic properties. For accounts of the semantic properties of representations typically appeal to causal relational properties, on both the input side and the output side. These are certainly not intrinsic properties.

Clearly, the language of thought hypothesis is very far from being trivially true. In principle, a physical configuration with the representational content that Fido barks might be syntactically unstructured. It might not have two causally potent physical properties, one correlated with the semantic property of being about Fido and the other correlated with the semantic property of being about barking.

9.2 Compositionality and inference to the best explanation

Whether the language of thought hypothesis is true is a substantive empirical question that cannot be settled by introspection alone. In *The Language of Thought* (1975),

Fodor's argument in favour of the hypothesis begins from the idea that the best cognitive psychological theories postulate internal representations and processes that manipulate those representations. This is surely enough to motivate the claim that there are internal states with both semantic properties and intrinsic causally potent properties. But why should the intrinsic causal properties be syntactic properties? Why should they be correlated with the semantic properties?

Here, Fodor takes over from philosophy of language the notion of *compositionality*: the meanings of whole sentences depend on the meanings of their parts. Compositionality provides an explanation of semantic *productivity*. For where a sentence's constituent parts – or, more generally, its intrinsic properties – are correlated with semantic properties, the recombination of these parts or properties allows the construction of further sentences with different but related semantic properties. For example, the syntactic structure in 'Fido barks' and in 'Fiona sings' guarantees that we can recombine constituent parts in order to express the thoughts (both false, let us suppose) that Fido sings and that Fiona barks.[10] In contrast, where intrinsic properties are not correlated with semantic properties there is no reason to expect semantic productivity. A language might contain two syntactically unstructured sentences, one meaning that Fido barks and the other that Fiona sings, but not provide for the expression of any other thoughts at all.

Where we find semantic productivity, we naturally postulate compositionality as its explanation; and that involves postulating that representations have syntactic properties. So Fodor can construct an argument in support of the language of thought hypothesis – an inference to the best explanation – by showing that the schemes of representation postulated by cognitive psychological theories exhibit a measure of semantic productivity. He does this (1975, chapter 1) by pointing to examples in which a cognitive process needs to range over representations of states of affairs drawn from an open-ended domain.

Not all the cognitive processes that operate, according to this line of argument, over syntactically structured representations are thought processes. But when we consider the case of thoughts the line of argument seems particularly compelling. For semantic productivity seems to be at the very heart of thinking. If someone is able to frame the thought that Fido barks and the thought that Fiona sings then that person is in a position to frame (even if not to believe) the thought that Fido sings and the thought that Fiona barks.[11]

9.3 Tacit knowledge of rules and syntactically structured representations

Semantic productivity provides one line of argument in support of the language of thought hypothesis; but there is also a quite general connection between tacit knowledge of rules and syntactic properties of representations. We can illustrate this connection with a very simple example.

---

[10] Fodor (2001) argues that the thought that is expressed by a sentence is not determined compositionally by the meanings of the sentences parts and the way they are put together. He puts this by saying that 'language is not compositional' (2001, p. 14) but we could equally well say that compositionally determined linguistic meaning does not fully determine the content of the thought expressed.

[11] See Evans, 1982, p. 104 on the Generality Constraint and Fodor, 1987, Appendix. Fodor draws a distinction between semantic *productivity* (having to do with the potential infinity of thoughts that can be expressed) and *systematicity* (having to do with the fact that being able to express the thought that John loves Mary is 'intrinsically' connected to being able to express the thought that Mary loves John).

Consider again the task of assigning pronunciations to letter strings (section 8). In particular, consider the 125 three-letter strings that can be built from a set of five possible onset consonants, five possible vowels, and five possible coda consonants; and suppose that these strings have pronunciations that conform to regular letter-sound rules. If an information-processing system assigns pronunciations correctly, then there are fifteen patterns in its input-output relation, of which one example would be this: Whenever the input represents a letter string whose onset consonant is 'B', the output represents a pronunciation that begins with the sound /B/.

According to the account of tacit knowledge sketched earlier (section 7.2), tacit knowledge of rules requires the presence of a battery of causal-explanatory states, one state corresponding to each rule. In the present example, each such state would function as a causal common factor in explaining the twenty-five instances of one input-output pattern. One way for an information-processing system to embody tacit knowledge of letter-sound rules would be for it to make use of a stored, syntactically structured representations of those rules. But it is possible for a system to embody tacit knowledge of rules without containing any such stored representations. Suppose, in any case, that the requirement for tacit knowledge is met, and consider the twenty-five input configurations that represent strings beginning with the letter 'B'. These physical configurations need to share some property that will engage or activate the 'B'-to-/B/ component processing mechanism. This property, which we may suppose to be physical and intrinsic, will be a determinant of causal consequences and it will be correlated with the semantic property of representing a string beginning with the letter 'B'. In short, this property will meet Fodor's conditions for being a syntactic property. But the information-processing system also embodies tacit knowledge of fourteen other letter-sound rules including, for example, the 'I'-to-/I/ rule for a vowel and the 'N'-to-/N/ rule for a coda consonant. So, by the same argument, the input representation of the three-letter string 'BIN' must have three syntactic properties correlated with the three semantic properties of representing a string beginning with 'B', representing a string with 'I' in the middle, and representing a string ending in 'N'.

The general connection between tacit knowledge of rules and syntactic properties of representations is thus that processing systems that embody tacit knowledge of rules need to have syntactically structured input representations. There is compositionality here. The representational content of an input configuration is determined by the semantic properties correlated with the three syntactic properties that it instantiates. But, whereas the earlier argument in support of the language of thought hypothesis (section 9.2) was an abductive argument for compositionality as the explanation of semantic productivity, the argument just outlined is a more nearly deductive argument from the involvement of tacit knowledge in cognitive processes.

9.4 Concept possession and the language of thought hypothesis

Just as the argument from semantic productivity seems particularly compelling in the case of thoughts, so also it is plausible that we have tacit knowledge of rules involving thoughts, namely, rules of inference. The reason is that possessing particular concepts involves a thinker in commitments to particular *forms* of inference. Commitment to a form of inference is not just commitment to each of a number of inferences that happen to instantiate that form. Rather, the commitment is to accept or perform those inferences 'in virtue of their form' (Peacocke, 1992, p. 6). The form of the inferences should figure, somehow, in the causal explanation of the thinker's performing those inferences. It is not

obvious what this requirement comes to (Peacocke, 1992, pp. 183–4). But a kind of inference to the best philosophical explanation suggests that performing inferences in virtue of their form involves meeting the conditions for tacit knowledge of the corresponding rule of inference.

Now consider a thinker who thinks a thought in whose content the concept C is a constituent, and suppose that R is a tacitly known rule of inference in whose premise the concept C figures. We have assumed that such a personal-level event of conscious thought is underpinned by the occurrence of a physical configuration to which we can assign the same content. So, just as a physical configuration that represents a letter string beginning with 'B' needs to have a syntactic property that engages the 'B'-to-/B/ component processing mechanism, so also a physical configuration whose content involves the concept C needs to have a syntactic property that engages the R component processing mechanism. And, just as a physical configuration that represents the three-letter string 'BIN' must have three recombinable syntactic properties corresponding to its three semantic properties, so also the physical configuration that underpins a thought whose content has several concepts as constituents must have several recombinable syntactic properties encoding those concepts. For we assume that, for each concept that a thinker possesses, there is at least one form of inference to which the thinker is committed.

If the initial idea about tacit knowledge of rules of inference is correct, then the general connection between tacit knowledge and syntactically structured representations provides a relatively straightforward philosophical argument in support of the language of thought hypothesis (Davies, 1991, 1992). It would surely be overly ambitious to suppose that philosophy, unaided by detailed empirical investigation, could settle the question whether or not the language of thought hypothesis is true of the information processing that takes place inside the heads of human beings. But it is not overly ambitious to suppose that philosophical theory may uncover, within our ordinary conception of ourselves as conscious thinking subjects and agents, commitments to particular kinds of cognitive structures and processes.

10. Computational Psychology and Levels of Explanation

Although the argument for a general connection between tacit knowledge of rules and syntactically structured representations establishes only a one-way dependence, it is clear that syntactic structure and tacit knowledge are made for each other. In typical cases where they go together, the rules that are tacitly known are cast in terms of items in some task domain and the properties of those items – for example, in terms of letter strings and their orthographic and phonological properties. Cognitive processes operate over input representations to generate output representations in ways that are dictated by the rules. And this can be done mechanically because the semantic properties of representations – for example, what letter string is being represented – are encoded by syntactic properties and these, being intrinsic and causally potent, are the sorts of properties (unlike semantic properties) that can engage mechanisms.

10.1 The computational theory of mind

Putting syntactically structured representations and tacitly known rules together yields the computational theory of mind. Representational mental states have, or are underpinned by states that have, syntactic properties; and cognitive processes are computational processes that operate over those representations in virtue of their syntactic properties.

It may seem that the very fact that it is syntactic properties, rather than representational properties, that do the causal work presents a problem for the computational theory of mind. For it is part of our everyday conception of the mind that the representational properties of mental events are crucial to the causal consequences of those events. It is because my belief has the specific content that it does – for example, the content that I am being attacked by a bear rather than the content that I am being approached by someone offering a glass of champagne – that it has the specific causal consequences that it does. The postulation of syntactic properties was supposed to help explain how causal-explanatory claims about thoughts, conceived as representational states, could be true. But now it may seem that the syntactic properties make the representational properties causally irrelevant.[12]

One way of responding to this concern about the computational theory of the mind is to draw a distinction between a more inclusive class of causally *explanatory* properties and a narrower class of causally *efficacious* properties. As Frank Jackson and Philip Pettit say (1988, p. 392): 'Features which causally explain need not cause.' Jackson and Pettit argue that representational properties are not causally efficacious properties because of the highly relational nature of content. But a representational property could still be a causally explanatory property. It could figure in causal explanations and in causal laws.

Simply drawing the distinction between explanatory and efficacious properties is not enough, by itself, to provide a fully satisfying response to the concern about the computational theory of the mind. We need to be assured that the reasons for saying that representational properties are not efficacious do not have the consequence that only fundamental physical properties are efficacious, or that no properties at all are efficacious. Otherwise, syntactic properties will turn out to be no more efficacious than representational properties. But the prospects seem to be good for defending the computational theory of the mind by saying that it is because syntactic properties are causally efficacious that representational properties are causally explanatory.

10.2 Levels of explanation in computational psychology

At the beginning of his book on the computational processes involved in human vision, David Marr (1982) sets out an approach that has been influential within cognitive science and much discussed by philosophical commentators. Marr describes in some detail the grounds for his conviction that, in neurophysiological investigations of vision in the 1970s, 'something was going wrong' (1982, p. 14) and 'something important was missing' (p. 15). This culminates in the following striking claim (p. 19):

> There must exist an additional level of understanding at which the character of the information-processing tasks carried out during perception are analyzed and understood in a way that is independent of the particular mechanisms and structures that implement them in our heads.

[12] Recall from section 9.4 that we wanted it to be true that a thinker is committed to performing inferential transitions in virtue of their form. At that stage of the argument, form was conceived as a representational matter – as a matter of the concepts that are constituents of the contents of thoughts. The requirement was that the form of the inferences should figure, somehow, in the causal explanation of the thinker's performing those inferences. Postulating syntactic properties to encode conceptual constituents was supposed to be a way of meeting that requirement (Peacocke, 1992, pp. 183–4).

Thus Marr was led to propose that information-processing mechanisms have to be understood at three levels. The first is the level of the computational theory that tells us what is being computed, why – given the requirements imposed by the task – this is being computed, and how in principle this might be computed – 'what is the logic of the strategy by which it can be carried out' (p. 25). The second level of understanding involves the specification of a scheme or format of representation and an algorithm by which the computation is to be carried out. And the third level is the level of physical realisation.

In general, experimental research in cognitive psychology is directed towards understanding at the second level. The aim is to develop theories about the representational structures and the computational processes that are actually implicated in performance of an information-processing task. Different algorithms and different representational formats lead to different predictions about the performance of subjects – different predictions about reaction times, for example. But any such theory about cognitive structures and processes is constrained *from above* by the first-level abstract theory of the task and, at least in principle, *from below* by the third-level theory about physical realisation. Marr (p. 27) is explicit that it is the constraint from above that is more crucial and, in this sense, he recommends a top-down, rather than a bottom-up, approach to cognitive scientific research.

The relationship between Marr's first and second, and second and third levels is one-many. There may be several different algorithms and schemes of representation for performing the same computation, considered as a function in extension. And there may be many different physical realisations of the same algorithm and scheme of representation. Christopher Peacocke (1986) proposes that we should interpolate an additional level between Marr's first and second levels. What is specified at Peacocke's level 1.5 is 'the information on which the algorithm draws' (1986, p. 101). This interpolation preserves the pattern of one-many relationships between levels. First, as between Marr's first level and level 1.5, one function in extension might be computed by drawing on different bodies of information. Second, as between level 1.5 and Marr's second level, one body of information might be drawn on by different algorithms using different schemes of representation.

10.3 Marr's levels and Chomsky's competence-performance distinction

Marr says that the distinction between his first and second levels is the difference between what is computed and how it is computed, and he suggests that it corresponds to Chomsky's distinction between competence and performance. Part of Marr's point here is that developing a theory of how something is computed – a theory of representation and algorithm – is 'a completely different endeavour' (1982, p. 28) from formulating the theory of what is being computed – the 'computational theory' in Marr's terminology. Chomsky's theory of competence is not a theory of 'how grammatical structure might actually be computed from a real English sentence' (p. 28) and he is not committed to any specific account of how sentence processing might be achieved. (Recall the discussion of the derivational theory of complexity; section 6.2.)

But it is not clear that locating a Chomskyan theory of competence at Marr's first level takes full account of the fact that a theory of competence is an empirical psychological theory. For, as we have construed it, Marr's computational theory is a relatively abstract theory about how, in principle, a function in extension might be computed. A computational theory might tell us that a set of structural descriptions could,

in principle, be generated by a particular collection of rules, or that the same set of structural descriptions could be generated by several different collections of rules. But a theory of competence is supposed to describe the body of specifically linguistic tacit knowledge that is actually possessed by a language user. It is supposed to specify the rules that are tacitly known – that is, a generative procedure conceived as a function in intension – and not just the set of structural descriptions generated by those rules – that is, the extension of that generative procedure (Chomsky, 1995, p. 15).

So a Chomskyan theory of competence does not really belong at Marr's first level as we have construed it. But, as Marr rightly points out, it does not belong at the second level. A theory of representation and algorithm would be a theory about an actual computational process. It would be a theory of sentence processing and so part of the theory of performance, rather than competence. An empirical theory of competence – a theory of the knowledge that is drawn on by the computational processes that subserve linguistic performance – properly belongs at Peacocke's level 1.5.

A theory at level 1.5, like a theory at Marr's second level, is constrained from above and from below. From above, it is constrained by a formal requirement of adequacy in principle. The postulated body of tacit knowledge should at least generate the right set of structural descriptions. From below, it is constrained by the requirements of computational implementation – algorithm and representational format – and physical realisation. We saw the same pattern of constraint earlier (towards the end of section 6.1) when we considered theories about the initial state of the language faculty, for those theories also belong at level 1.5. That pattern of constraint from above and from below remains intact when, in later developments, the innate endowment is conceived rather differently and the process of language acquisition becomes the process of parameter setting (Chomsky 1986, 1995; Roeper and Williams, 1987; Pinker, 1995).

11. Informational Encapsulation and the Modularity of Mind

Our discussion of the first of two Fodorian contributions to cognitive science – the language of thought hypothesis (section 9) – led into a more general consideration of computational theories (section 10). I now turn to a second theme from Fodor, namely, modularity.

The generic idea of modularity is familiar both from everyday life and from science. In everyday life, furniture and stereo systems are 'modular'. They are built from components, each of which makes a relatively independent contribution to the functionality or the performance of the system as a whole. In science, the law of gravitational attraction allows us to calculate the gravitational force exerted by one body on another and so, by way of the connection between force and acceleration ($F = ma$), yields predictions about motion. If the two bodies in question are a proton and an electron then, because of the force between charged bodies, the motion of the electron will not actually be as predicted from the law of gravitational attraction. Yet the law of gravitational attraction is still true under idealisation or 'all else equal' (*ceteris paribus*). This does not mean that the motion of an electron *would* be as predicted from the law of gravity *if only* protons and electrons had no charge. We scarcely know how to interpret that counterfactual conditional. Rather, the law of gravity is true all else equal in the sense that departures from it can be explained, as in the case of the proton and electron, in terms of independent factors; that is, in terms of 'interference from independent systems' (Pietroski and Rey, 1995, p. 87). The system of gravitational attraction and the system of

attraction and repulsion between charged bodies are independent systems – or, as we might say, 'modules' of the natural order – each system with its own laws.

We can take the generic idea of modularity and consider its application at each of Marr's three levels of explanation and also at the interpolated level 1.5.

11.1 Modularity and levels of explanation

At Marr's first level – the level of the abstract theory of what is being computed – we have the idea of a modular *task analysis*: the overall task to be performed is analysed into sub-tasks. This decomposition is close to what Robert Cummins (1983) calls 'functional analysis' and it can be represented in a box-and-arrow diagram – a flow chart. But, as Cummins stresses, the boxes in a flow chart do not represent components of a system that performs the task (1983, p. 29): 'A cook's capacity to bake a cake analyzes into other capacities of the "whole cook".'

At Marr's second level – the level of algorithm and representational format – we have the idea of modularity in processing. A *processing module* is a relatively independent or autonomous component of a larger information-processing system. The decomposition of a system into component sub-systems can, once again, be represented by a box-and-arrow diagram. But we need to guard against the too-simple assumption that there will be a tidy mapping between these sub-systems and the sub-tasks that figure in the flow chart at the first level – a 'direct form-function correlation', as Cummins puts it (p. 29).

At Marr's third level – the level of physical realisation – we have the idea of modularity in neuroanatomy. Suppose that a functional analysis of some cognitive task has revealed two sub-tasks and suppose that, as a matter of empirical fact, an information-processing system performs the cognitive task by having *inter alia* separate components that perform those two sub-tasks. This would still leave open the further empirical question whether the *anatomical regions of the brain* that subserve performance of the two sub-tasks coincide, or overlap, or are disjoint from each other.

At Peacocke's interpolated level 1.5, we have the idea of modularity in a body of information or system of knowledge. A *knowledge module* is a relatively independent component for storing information. In order to avoid trivialising the idea, we must not regard a difference in content as sufficient to justify a claim about separate modules (Fodor, 2000, p. 58). A single knowledge module may, in principle, contain information about different topics or different domains. But bodies of knowledge about different domains might be regarded as different knowledge modules if, for example, they were drawn on by distinct processing modules.

11.2 Fodor's modularity thesis

If it is the generic idea of modularity that is in play then it is relatively uncontroversial that the mind, conceived as an information-processing system, is modular. But Fodor's modularity thesis goes beyond this uncontroversial claim.

The background to the thesis is provided by a threefold taxonomy of cognitive mechanisms. First, there are transducers, whose outputs are the representations from which mental information processing begins. What is represented at this initial stage is something proximal – typically, the pattern of stimulation at a sensory surface. Second, there are input systems that perform inference-like transitions from these initial representations of the proximal stimulus to representations of the properties and distribution of distal objects. The processing in these systems is computational and the function of these systems is 'to so represent the world as to make it accessible to thought'

(Fodor, 1983, p. 40). The outputs of the input systems are representations of worldly objects, properties, events, and states of affairs. Third, there are central cognitive systems that subserve thinking, problem solving, planning, and the fixation of belief. Given this background, Fodor's modularity thesis says that input systems share theoretically important properties that are different from the properties of central systems. Input systems typically exhibit the marks of modularity; central systems do not.

11.3 The essence of Fodorian modularity

Fodor lists nine marks of modularity of which six are relatively straightforward. Modules are domain specific; that is, they are specialised for tasks like the analysis of spoken words and sentences or the perception of faces.[13] The operation of modules is mandatory and fast. Modules are associated with fixed neural architecture, they exhibit characteristic and specific patterns of breakdown, and their ontogeny exhibits a characteristic pace and sequencing. The remaining three marks require more comment.

First, there is only limited central access to the mental representations that modules compute. That is, the information at the various intermediate stages of a module's computation is not generally available to the subject; the states of affairs represented at those stages are not thereby accessible to thought. Second, the final outputs of modules, which are available to central systems, are 'shallow' (1983, p. 86). What Fodor means by this is that the interface between modules and central systems comes relatively early. In the case of sentence processing, for example, the output of the module might specify which sentence was uttered and it might specify the literal meaning of that sentence. But it would not specify whether the sentence was intended ironically or metaphorically – nor, more generally, what overall message the speaker was trying to communicate.

The final mark is the one that Fodor describes as 'perhaps the most important aspect of modularity' (p. 37) and even as its 'essence' (p. 71). Modules are *informationally encapsulated*; that is, the information that is available within a module is considerably less than all the relevant information that is represented within the organism. In particular, the processes in a module do not draw on all that the subject knows or believes. The canonical illustration of informational encapsulation is provided by visual illusions. In the Ames room illusion, even though I know perfectly well that the adult is taller than the child, perception still presents the child as taller. Fodor (1989) argues that having perceptual processes that are informationally encapsulated and draw only on a very restricted body of information is a way of giving due weight to both observational adequacy and conservatism in the fixation of belief.

Fodor's modularity thesis has a positive part and a negative part. Input systems exhibit the marks of modularity; central systems do not. Arguments about the positive part mainly concern the claim that input systems are informationally encapsulated (Garfield, 1987; Farah, 1994). Defence of that claim is facilitated by the idea that the outputs of an input system are shallow, so that the interface between input systems and central systems comes relatively early. It is more challenging when we bring other ideas to the foreground. Thus, for example, it was part of the initial picture that input systems represent worldly states of affairs so as to make them accessible to thought. Indeed, Fodor says (1983, p. 136, n. 31): 'It seems to me that we want a notion of perceptual

[13] Coltheart, 1999, argues that domain specificity goes to the heart of the notion of modularity; cf. Fodor, 2000, pp. 58–61.

process that makes the deliverances of perception available as the premises of *conscious* decisions and inferences'. These ideas favour the assignment of more, rather than less, processing to input systems, with the interface between input systems and central systems coming correspondingly later.

So long as the negative part of Fodor's modularity thesis is not called into question, arguments about the positive part can be regarded as revealing conflicting pressures on the location of the interface between modules – conceived as informationally encapsulated input systems – and central systems. But, in fact, the negative part of the thesis has been called into question.

11.4 Central systems and the limits of modularity

One of the marks of modularity is domain specificity, and it is natural to think of modules as solving problems of specific types: What kind of object is this? Whose face is this? What word is this? What sentence, with what structural description, is this? If a type of problem is solved by a module then there is a mechanical procedure – an algorithm – for solving problems of that type. More accurately, there is an empirically feasible mechanical procedure for solving problems of that type tolerably well (no less well than the module solves them). The other side of the modularity coin is that, if there is no empirically feasible mechanical procedure for solving problems of some particular type, then problems of that type must be solved by the central systems if they are solved at all.

Fodor suggests that we think about fixation of beliefs by analogy with the process of confirmation of hypotheses in science and he points to two features of hypothesis confirmation. First, confirmation is *isotropic* (1983, p. 105): 'the facts relevant to the confirmation of a scientific hypothesis may be drawn from anywhere in the field of previously established empirical (or, of course, demonstrative) truths'. Second, confirmation is *Quinean* (p. 107): 'the degree of confirmation assigned to any given hypothesis is sensitive to properties of the entire belief system.' If Fodor is right about the analogy between belief fixation in individual thinkers and hypothesis confirmation in science then it seems clear that the processes of belief fixation cannot be informationally encapsulated and that there is no mechanical procedure for deciding what to believe. Thus, about the Quinean feature of confirmation, Fodor says (1987, p. 63): 'it's hard even to imagine a mechanism whereby the whole cognitive background can contribute to determining the local tactics of problem solving'.[14]

If belief fixation cannot be the business of a module then, given the taxonomy of cognitive mechanisms that is the background to Fodor's modularity thesis, solving the problem of what to believe must be done by the central systems. But the same features that make belief fixation ill-suited to modularity also make it extremely difficult to understand and, in particular, difficult to understand in terms of the computational theory of mind. More generally, according to Fodor, the prospects for an empirically feasible computational theory of central cognitive processes are dim (1983, p. 107): 'the more global (e.g., the more isotropic) a cognitive process is, the less anybody understands it. *Very* global processes, like analogical reasoning, aren't understood at all'. Fodor dubs this claim 'Fodor's First Law of the Nonexistence of Cognitive Science' (p. 107).

[14] The process of scientific *discovery* – the formation, rather than confirmation, of hypotheses and theories – seems even more clearly unencapsulated and, intuitively, there is no mechanical procedure for making a scientific breakthrough. These points about scientific practice also have their analogues in the case of belief fixation. See Fodor, 1983, pp. 106–7.

The overall situation appears to be this. The arguments for syntactically structured representations are particularly compelling in the case of thoughts (section 9). Syntactic structure and tacit knowledge of rules are made for each other and the computational theory of mind is the result of bringing them together (section 10). Yet the application of the computational theory of mind in the domain of thought is problematic. *Modus ponens* inferences fit the computational theory well enough; but inference to the best explanation fits the theory less well, perhaps even to the point of intractability.

The limitations of the computational theory of mind have been a recurrent theme in Fodor's work, at least since the final chapter of *The Language of Thought*, where he says (1975, p. 200): 'There seem to be some glaring facts about mentation which set a bound to our ambitions.' But someone might hope to bring central cognitive processes within the scope of the computational theory of mind by rejecting the negative part of Fodor's modularity thesis and maintaining instead that the mind is modular through and through – *massively modular* (Sperber, 2002). The idea would be that central cognitive processes, like input processes, are subserved by modules; not one module for thinking, one for problem solving, one for planning, and one for belief fixation, but a host of modules, each dedicated to the solution of a particular, and perhaps quite idiosyncratic, type of problem. The task for someone wanting to make use of this idea is, of course, to show how the features of human thought that seem problematic for the computational theory of mind could emerge from a massively modular cognitive architecture (Carruthers, 2003a, 2003b, 2004).

The massive modularity hypothesis draws some support from evolutionary psychology (Barkow, Cosmides and Tooby, 1992; Cosmides and Tooby, 1994), from examples of domain specificity in cognitive development (Hirschfeld and Gelman, 1994), and from dissociations between impairments in the performance of central cognitive tasks (Shallice, 1988, part 4). Fodor argues against it in *The Mind Doesn't Work That Way* (2000).[15] I cannot review the debate here, but perhaps it is enough to note that there are serious open questions about the scope and limits of the approach to cognitive science that we have been describing.

12. Modules and Cognitive Neuropsychology

Research in cognitive neuropsychology has two complementary aims. One is to use theories about normal cognitive processes to help understand disorders of cognition that result from stroke or head injury. The other is to use data from people with acquired disorders to test and further develop theories of normal cognition (Coltheart, 1985; Humphreys, 1991). This programme of research is based on a number of assumptions of which the first is that the mind is modular in the sense that there are relatively independent processing and storage components that can be selectively damaged. The second assumption is that the modular structure or *functional architecture* of the mind as a whole, and of the systems responsible for the performance of particular tasks, is the same for all normal (neurologically intact) subjects. Alfonso Caramazza (1986, p. 49) calls this the assumption of *universality*. The third assumption is that, when one component is damaged, this does not bring about massive reorganisation of the prior

[15] The title of Fodor, 2000, is directed at Pinker, 1997, who says (p. x): 'the mind is a system of organs of computation designed by natural selection to solve the problems faced by our evolutionary ancestors in their foraging way of life'. For further discussion, see Pinker, 2005a; Fodor, 2005; Pinker, 2005b.

modular structure. Rather, the undamaged components continue to operate as before, so far as this is compatible with the impaired operation of the damaged component. Caramazza (p. 52) calls this the assumption of *transparency*.

When we study normal subjects, the assumption of universality licenses the averaging of data across groups of subjects in order to assess hypotheses about the normal information-processing system. But when we study brain-damaged subjects, we cannot antecedently assume that the information-processing systems of different patients have been damaged in identical ways – even if the patients have been given the same clinical diagnosis. Rather, we reach hypotheses about damage to the normal system as putative explanations of specific patterns of impaired performance. So cognitive neuropsychology typically proceeds by the study of single cases. A series of single-case studies yields, via the assumption of transparency, multiple constraints on theories about the normal functional architecture.

12.1 The dual-route theory of reading aloud

To see the methodology of cognitive neuropsychology at work, consider the task of reading single words aloud (Coltheart, 1985). So far as this is an information-processing task, it calls for transitions from representations of orthography to representations of phonology. One way of carrying out the task would involve, for each orthographic input representation, a direct mapping to a phonological output representation, drawing on *lexical* information about the orthography and phonology of a single word. Another way would involve, for each orthographic input representation, the assembly of a phonological output representation, drawing on *non-lexical* information about regular letter-sound correspondences. The *dual-route* theory of the processes involved in mature reading aloud of single words starts from the idea of a lexical route and a non-lexical route from print to speech. In the case of regular words, both routes would deliver the same correct pronunciation. In the case of irregular words like ‘PINT’ or ‘YACHT’, the lexical route would be vital for a correct pronunciation, while in the case of pronounceable non-word letter strings like ‘SLINT’ or ‘VIB’, only the non-lexical route would deliver a pronunciation.

Since the dual-route theory of reading aloud involves two relatively autonomous processing systems, we can consider predictions about the consequences of selective damage to one route or the other. If the lexical route were damaged while the non-lexical route continued to operate unimpaired then the predicted pattern of performance would be preserved reading of regular words and non-words but regularisation errors on irregular words (for example, ‘PINT’ pronounced to rhyme with ‘MINT’). If the non-lexical route were damaged while the lexical route continued to operate unimpaired then the predicted pattern of performance would be preserved reading of both regular and irregular words but impaired pronunciation of non-words.

In fact, each of these patterns of performance is found in patients with acquired disorders of reading. The first is surface dyslexia; the second is phonological dyslexia. So the dual-route theory of normal reading promises to help us understand these acquired disorders. We can explain them in terms of selective damage to some components of the normal reading system while other components continue to operate as before. To the extent that these are not just good explanations of surface dyslexia and phonological dyslexia but the best explanations, the dual-route theory of reading is supported and competing theories are disconfirmed.

12.2 Double dissociation arguments

People with surface dyslexia and people with phonological dyslexia instantiate a *double dissociation* of reading impairments. People with surface dyslexia show impaired reading of irregular words but intact reading of non-words while people with phonological dyslexia show the reverse pattern – impaired reading of non-words but intact reading of irregular words. The dual-route theory accounts for this double dissociation of impairments in terms of damage to separate component systems or processing modules that are implicated in reading irregular words (the lexical route) and in reading non-words (the non-lexical route).

The general pattern here is that a double dissociation between impairments in the performance of two tasks supports theories that postulate separate processing modules that are responsible for, or at least distinctively implicated in, performance of those two tasks (Shallice, 1988, part 3). Thus, for example, suppose that we find people with impaired recognition of faces but intact recognition of visually presented objects and other people with impaired recognition of visually presented objects but intact recognition of faces. This double dissociation of impairments would support theories that postulate separate modules implicated in face processing and in visual object processing.[16]

Double dissociation arguments occupy a central position in the practice of cognitive neuropsychology and it is sometimes said that evidence of associations or of one-way dissociations is of less value than evidence of double dissociations. First, evidence of associations is said to be of less value than evidence of dissociations because associations of impairments might just reflect facts about neuroanatomy. Even if separate modules are responsible for two tasks, the locations of the neural regions that subserve the tasks might make it virtually impossible for one module to be damaged while the other is spared. Second, evidence of one-way dissociations is said to be of less value than evidence of double dissociations because, even if two tasks are performed by a single module, damage to that system may result in performance of the more difficult task being impaired while performance of the easier task remains intact.

It is correct that arguments from associations to shared modules, or from one-way dissociations to separate modules, must address the possibility of alternative explanations of the data. And it is correct that these particular kinds of alternative explanation are not clearly available in the case of double dissociations. But none of this should be allowed to obscure the fact that double dissociation arguments, like arguments throughout normal science, are abductive. They work by inference to the best explanation (Coltheart and Davies, 2003).

[16] See Coltheart, 1999, for discussion and references. On double dissociation arguments, see also Shallice, 1988, part 3; Dunn and Kirsner, 2003, and the commentaries thereon.

III Challenges and Prospects

13. Four Challenges from Connectionism: Brains, Rules, Learning, Dissociations

While the basic ideas behind connectionist models of cognitive processes have a long history, contemporary research on connectionist, parallel distributed processing, or neural network models owes much to the appearance in 1986 of two major volumes by David Rumelhart, Jay McClelland and the PDP Research Group.[17]

Connectionist modelling of cognitive processes has captured the imagination of both cognitive scientists and philosophers at least in part because it seems to support many different challenges to the dominant classical approach to cognitive science. In this section, I consider four such challenges. First, because connectionist networks are 'brain-like', they seem to enjoy an advantage of plausibility over classical information-processing systems. Second, because networks are said to work without syntactically structured representations or tacitly known rules, connectionism seems to favour an alternative to the computational theory of mind. Third, because networks 'learn', connectionism seems to offer support to those who reject nativism. Fourth, because networks without modular architectures are said to show double dissociations of impairments after damage, connectionism seems to undermine the methodology of cognitive neuropsychology.

13.1 Units, connections, and the brain

The formal or numerical description of a connectionist network speaks of units and of connections between units. Each unit has a level of activation between zero and one; each connection has a weight that can be any real number, positive or negative. The level of activation of an individual unit is determined by an activation function given the input that the unit receives as a result of activation at units that are connected to it and the weights on those connections. The level of activation, a_i, of a unit, u_i, is the result of applying an activation function to the sum $\Sigma_j\, a_j\, w_{ij}$, where u_j is a unit connected to u_i and w_{ij} is the weight on the connection.

In most connectionist networks, units are organised into layers: a layer of input units and a layer of output units with one or more layers of hidden units in between. Suppose that we impose a pattern of activation on the input units of a layered network. Given the activation function and the weights on the connections, this determines a pattern of activation over the hidden units and that in turn determines a pattern of activation over the output units.

The basic ideas of connectionism are neurally inspired, with units and their activation levels, connections and their weights being simplified analogues of neurons and their firing rates, synapses and their strengths. In the brain, a neuron receives signals from other neurons by way of synaptic connections between the axons of other neurons and its dendrites. If the sum of the incoming signals is sufficiently high then the neuron fires, sending a signal along its axon to the dendrites of other neurons. For this reason, it is sometimes suggested that connectionist cognitive science enjoys an advantage of

[17] Rumelhart, McClelland and the PDP Research Group, 1986; McClelland, Rumelhart and the PDP Research Group, 1986. McLeod, Plunkett and Rolls, 1998, provides a detailed introduction to connectionist modelling, and the book comes with software for running connectionist simulations. See also Clark, 1989; Churchland, 1990; Bechtel and Abrahamsen, 1991.

plausibility over the classical approach to information processing. But here we need to consider two kinds of case. First, some connectionist networks are offered as models of the operation of real populations of neurons – for example, in the hippocampus or in the parietal cortex.[18] There is a great deal of important and illuminating work here at the neurobiological level, where there is an intimate relationship between representation and algorithm, on the one hand, and physical realisation, on the other. But, second, in the case of many connectionist models of cognitive processes there is no suggestion that the units and connections in the models correspond to real neurons and synapses (Smolensky, 1988, pp. 32–3). These models belong at a level of description that presumably supervenes on the neurobiological and, ultimately, on the fundamental physical. But it is not clear why, as putative descriptions of cognitive processes, they should be reckoned more plausible for being 'brain-like' (see further, McLaughlin, 1993; McLaughlin and Warfield, 1994).

13.2 Representations, rules, and learning

In connectionist networks, patterns of activation over units are the vehicles of representation. For any given pattern of activation considered as a representation, the property of containing a particular sub-pattern of activation is an intrinsic and causally potent property. But is it a syntactic property?

Consider, for example, patterns of activation that represent three-letter strings (section 9.3). If each three-letter string is represented by an entirely separate pattern of activation then these patterns are syntactically unstructured representations even if they are distributed over several units. If each three-letter string is represented by a pattern of activation made up of separate sub-patterns representing the onset, vowel, and coda then the patterns are syntactically structured representations with a compositional semantics. But, in between these extremes, there is a third possibility. The patterns of activation representing three-letter strings beginning with 'B', for example, might not have a sub-pattern strictly in common, yet might still be similar. (This might be, for example, because activation at individual units represents 'microfeatures' of hand-written letters, reflecting differences in the way that 'B' is written in different contexts.) In this case, it can be at most approximately true that the representations of three-letter strings are syntactically structured.[19]

Rumelhart and McClelland (1986, p. 218) suggest that connectionist networks 'may provide a mechanism sufficient to capture lawful behaviour, without requiring the postulation of explicit but inaccessible rules.' The intended contrast here is with tacitly known rules that are inaccessible to consciousness but are *explicit* in the sense that they are represented in a format with syntactic structure and stored in a way that requires additional processes of search and access before the knowledge can be used. In the case of connectionist networks, rules are not explicit in this sense, for knowledge is stored in

[18] For a connectionist model of processing in the hippocampus, see Rolls, 1989; McLeod, Plunkett and Rolls, 1998, chapter 13. For the parietal cortex, see Pouget and Sejnowski, 2001. For reviews, see Churchland and Sejnowski, 1992; Farber, Peterman and Churchland, 2001.

[19] See Smolensky, 1988, pp. 16–7: 'These constituent sub-patterns representing *coffee* in varying contexts are activity vectors that are not identical, but possess a rich structure of commonalities and differences (a family resemblance, one might say).' In this case, it can be at most approximately true that the representations of *cup with coffee*, *can with coffee*, *tree with coffee*, and *man with coffee* are syntactically structured

the weights on connections. Thus (McClelland, Rumelhart and Hinton, 1986, p. 32): 'Using knowledge in processing is no longer a matter of finding the relevant information in memory and bringing it to bear; it is part and parcel of the processing itself.'

Now, it is true that classical cognitive science does, at least sometimes, postulate rules that are explicit in the sense that is relevant here. But the notion of tacit knowledge explained earlier (section 7.2) does not require explicitness. It allows that tacit knowledge of a rule might be directly embodied in a processing mechanism. So far, then, there is no reason why a connectionist network should not embody tacit knowledge of rules.[20] But the general connection between tacit knowledge of rules and syntactically structured input representations ensures that, if it is at most approximately true that the input representations are syntactically structured, then it can be at most approximately true that the network embodies tacit knowledge of rules.

Paul Smolensky (1988, p. 11) says that the reason for the departure from syntactic structure is a 'dimensional shift' between the concepts used in a classical task analysis and the semantics of individual units in the network. But there is nothing about connectionist modelling as such that requires schemes of input representations that are so obliquely related to a classical description of the task domain. And, where there is no dimensional shift, there is no reason of principle why a network should not embody tacit knowledge of rules that are cast in the same terms that figure in a classical task analysis. Thus suppose that a connectionist model of reading aloud makes use of input representations with syntactic properties (sub-patterns of activation) that are correlated with representation of individual letters. Then we can intelligibly ask whether the model generates pronunciations of regular words and pronounceable non-words by drawing on regular letter-sound rules and the fact that the model is a connectionist one does not rule out the possibility that the answer to the question might be affirmative.

In connectionist networks, patterns of activation over units are relatively transitory vehicles of input and output representations. More abiding knowledge about the task domain is embodied in weights on the connections. Networks are said to 'learn' in the sense that there are algorithms for adjusting the weights on connections in order to bring input-output performance more closely into conformity with a training set of input-output pairs. In the case of feed-forward networks with hidden units, the training procedure is back-propagation of error (Rumelhart, Hinton and Williams, 1986; McLeod, Plunkett and Rolls, 1998, chapter 5).

A network's progress through the epochs of a training regime is sometimes taken as a model of a process of cognitive development. Thus, for example, in a pioneering contribution to developmental connectionism Rumelhart and McClelland (1986) offered a model of learning the past tense of English verbs.[21] But the idea of algorithms for extracting from a training set information about patterns or regularities is not exclusive to connectionism. There is a body of classical cognitive scientific research on rule induction and there are studies comparing the performance of classical algorithms and algorithms

[20] See Fodor and Pylyshyn, 1988, p. 60: '[O]ne should not confuse the rule-implicit/rule-explicit distinction with the distinction between Classical and Connectionist architecture.'

[21] For a critique, see Pinker and Prince, 1988. For more recent work on the past tense, see Marcus, 1995; Plunkett and Marchman, 1993, 1996.

used for training connectionist networks. Connectionism does not offer any special support to those who reject nativism.[22]

13.3 Modularity and dissociations in networks

Connectionist cognitive science is not opposed to the generic idea of modularity. Thus, Hinton, McClelland and Rumelhart (1986, p. 79) say that 'different modules would be devoted to things as different as mental images and sentence structures'. But the way in which knowledge is stored in the weights on connections opens up the possibility that there may be less modularity in a network than we might expect given a classical analysis of the task.

In an influential paper, William Ramsey, Stephen Stich, and Joseph Garon (1990) investigate the way in which a simple feed-forward network might depart from *propositional modularity*, which is the claim that (1990, p. 504) 'propositional attitudes are *functionally discrete*, *semantically interpretable*, states that play a *causal role* in the production of other attitudes, and ultimately in the production of behavior'. (Propositional modularity is thus similar to Fodor's intentional realism.) A network was trained by back-propagation of error to generate as output the correct verdict ('yes' or 'no') on each of sixteen propositions, such as 'Dogs have fur' and 'Cats have gills', that were encoded by patterns of activation across the input units. Knowledge of the verdicts on all sixteen propositions was embodied in the weights on the connections in the network. But there were not sixteen separate processing mechanisms responsible for the sixteen input-output transitions.[23]

In a similar way, there are connectionist models of reading aloud that do not incorporate the modularity to which the dual-route theory is committed. In these models, there are not two separate processing mechanisms corresponding to the lexical route and the non-lexical route. Rather, after the network has been trained on about 3,000 monosyllabic words, the weights on the connections are responsible for producing pronunciations for regular words, irregular words, and pronounceable non-words.[24]

Because connectionist models often exhibit less modularity than their classical counterparts, it is natural to suppose that they may face special challenges from evidence of double dissociations. So, when a network performs two cognitive tasks without containing two separate processing mechanisms, it is important to investigate whether damaging (or 'lesioning') the network can result in a double dissociation of impairments.

One possibility is that, while there are not two component modules or two routes from input to output, the performance of one task depends more heavily on one aspect of the network while the performance of the other task depends more heavily on some other aspect. In such a case, it may be that a double dissociation of impairments can be produced by damaging the network in two different ways; for example, damage to some connections versus damage to other connections (Plaut, 1995).

---

[22] On classical rule induction and connectionist learning, see McLaughlin and Warfield, 1994, for further discussion and references. On connectionism and nativism, see Elman et al., 1996.

[23] Ramsey, Stich and Garon put forward an eliminativist argument (1990, p. 500): '*If* connectionist hypotheses . . . turn out to be right, so too will eliminativism about propositional attitudes.' See Clark, 1990, 1993, and Stich, 1996, for further discussion.

[24] See Seidenberg and McClelland, 1989; Seidenberg, 1989; Patterson, Seidenberg and McClelland, 1989; Plaut, McClelland, Seidenberg and Patterson, 1996; McLeod, Plunkett and Rolls, 1998, chapter 8.

Suppose, however, that there is no principled way of damaging the network so as to produce a double dissociation and that the typical result of damage is impaired performance of both tasks. Then – a second possibility – there may still be sufficient variability in the exact levels of performance of the two tasks so that, if the network is damaged many thousands of times, particular patterns of impairment and sparing that instantiate a double dissociation may occur (Juola and Plunkett, 2000; but see also Bullinaria and Chater, 1995; Plaut, 2003).

A third possibility is that damage to a network consistently produces a pattern of impairment and sparing that corresponds to one half of a double dissociation that is found in brain-injured patients and there is no evident way of damaging the network to produce a pattern of performance corresponding to the reverse dissociation. In this kind of case, a defender of the network as providing a model of the cognitive processes by which both tasks are normally performed may call the assumption of transparency (section 12) into question. It may be argued that, if there is near-total damage to the normal processing system, then some different system is pressed into service. The reverse dissociation would then be explained in terms of the performance of this back-up system.

In these ways and others, connectionist cognitive science may offer putative explanations of double dissociations of impairments without postulating separate processing mechanisms that are distinctively implicated in normal performance of the two tasks. These alternative explanations are candidates for being the best explanation and they confront the totality of relevant evidence alongside the putative explanation that appeals to separate modules. But it does not follow that connectionism has somehow revealed that 'double dissociations don't mean much' (Juola and Plunkett, 2000). No evidence ever constitutes a logical guarantee of the truth of the theory that correctly explains it (Coltheart and Davies, 2003).

In a review of work on the cognitive neuropsychology of language, Mark Seidenberg says (1988, p. 405):

> There seems to be basic characteristic of this research that limits its interest, and that is the commitment to explanations framed in terms of the 'functional architecture' of the processing system. One of the main characteristics of the cognitive neuropsychological approach as it has evolved over the past few years . . . is that very little attention is devoted to specifying the kinds of knowledge representations and processing mechanisms involved.

Seidenberg's complaint here is that many models of cognitive processes are presented as box-and-arrow diagrams, with very little detail about either algorithm or representational format. Such models are not explicit enough to be implemented as computer programmes and in this respect they compare unfavourably with connectionist models.

But clearly, explicitness and implementation need not be exclusively associated with connectionist cognitive science. Max Coltheart and his colleagues have developed an implemented version of the classical dual-route model of reading aloud, and they list twenty-seven effects, observed in experiments with normal and brain-damaged subjects, that the model simulates (Coltheart, Rastle, Perry, Langdon and Ziegler, 2001, p. 251). Their claim is that no other presently implemented computational model of reading aloud can match this level of success. David Plaut and his colleagues (1996) also claim advantages for their connectionist model and, of course, there are other models as well, classical, connectionist, and hybrid.

The issue between these models of reading aloud will not be settled simply by appeal to the ostensible benefits of connectionism, such as neural plausibility, departures from syntactic structure and tacitly known rules, learning, and the simulation of dissociations. Rather, to the extent that the issue is settled at all, this will be by the normal method of extended comparison of competing research programmes as they face evidence from a multitude of sources. The same goes, more generally, for the issue between classical and connectionist approaches to cognitive science.

14. Prospects for the Philosophy of Cognitive Science

This chapter has focused on historical and foundational issues (sections 1–5) and then on one approach to cognitive science, the classical computational approach involving tacitly known rules and syntactically structured representations (sections 6–12). These are certainly important elements in analytic philosophy of cognitive science. But other elements, also important, have been neglected.

First, the classical approach to cognitive science faces challenges, not only from connectionism (section 13), but also from neuroscientific reductionism and from approaches that draw on evolutionary psychology (Barkow, Cosmides and Tooby, 1992), robotics (Brooks, 1991), dynamic systems theory (Port and van Gelder, 1995), and artificial life (Boden, 1996) and, more generally, from approaches that stress the idea that cognition as we know it is an activity of minds that are both *embodied* and *embedded* in a worldly environment (Clark, 1997).

Second, according to the view of the relationship between philosophy and cognitive science that was suggested earlier (section 5.1), philosophical theory may reveal that our conception of ourselves as persons has built into it commitments to particular kinds of cognitive structures and processes. Empirical research in cognitive science reveals whether those commitments are met. If they are met, then we learn how an aspect of our personhood is possible (Peacocke, 1992, chapter 7). If they are not met, then we are obliged to revise our philosophical theory or our conception of ourselves.

Given this general view of the inter-disciplinary relationship, we should expect that a host of relatively detailed empirical findings from specific programmes of research in cognitive science would impact on philosophical theory to enrich, refine, or even cast doubt on, aspects of our conception of ourselves as experiencing, thinking subjects and agents. Thus consider, for example, the Kantian line of thought explored by Peter Strawson (1959, chapter 2) and subsequently by Evans (1980). Our conception of the world as a world of objective particulars that exist independently of our experience of them is a conception of a spatial world. We conceive of ourselves as moving through that world so that our experience is explained jointly by the properties of objects and our location. The cognitive science of *spatial representation* helps to explain how this objective conception is possible (Eilan, McCarthy and Brewer, 1993).

Perhaps the cognitive scientific finding that has been most discussed in philosophy is that some patients with damage to primary visual cortex (area V1), who consequently have a blind region in their visual field and who report no visual experience when stimuli are presented in that blind region, are nevertheless able to discriminate stimulus properties when they are asked to guess. This is the phenomenon of *blindsight*: 'visual capacity in a field defect in the absence of acknowledged awareness' (Weiskrantz, 1986, p. 166; see also 1997). Under forced-choice conditions, a blindsight subject may be able to discriminate shape properties between X and O, movement properties between horizontal and vertical, wavelength properties between red and green (Stoerig and

Cowey, 1992), and even affective properties between happy and fearful faces (de Gelder et al., 1999).

People with *prosopagnosia* are unable to recognise and identify familiar faces. If, for example, a patient is asked to classify photographs as being of familiar or unfamiliar faces, he or she may perform at chance levels. Yet the patient's skin conductance responses may discriminate between the familiar and the unfamiliar faces. This is a kind of 'covert recognition'. A patient with prosopagnosia may be unable to classify the faces of famous people according to their occupation (for example, politician or television personality) although he can, of course, correctly assign occupations to the *names* of these people. But, in some cases, presenting the face of a television personality alongside the name of a politician or *vice versa* interferes with the patient's performance when asked to assign an occupation to the name, just as it does in normal subjects (Young, 1998). Information about the occupation of the person whose face is presented affects performance even though it is not available for verbal report.

These neuropsychological phenomena and thought experiments based on them are important for both empirical and philosophical theories about perception and consciousness. The blindsight patient reports no visual experience of the stimulus and, despite being able reliably to guess some of its properties, is not able to make normal use of this information for reporting, reasoning and planning. Ned Block (1995b) suggests that, in part because of the absence of both *phenomenal consciousness* and *access consciousness* in such cases, the two notions of consciousness are liable to be confused. But he argues that they are at least conceptually dissociable and, in particular, that we can make sense of a hypothetical 'super-blindsight' patient who is able to make free use of information about stimuli in the blind region for reporting, reasoning and planning, yet still does not have any visual experience of those stimuli. If this is right then, even if a satisfying explanation of access consciousness could be given in information-processing terms, this would still not be an explanation of phenomenal consciousness.

These neuropsychological phenomena also raise questions about the extent to which normal subjects are authoritative about the workings of their own minds. A normal subject would probably find it compelling to suppose that, when he sees the face of a television personality, his conscious recognition of the face interferes with his classification of a simultaneously presented name as that of a politician. Although this is likely to be correct, the fact that the same pattern of interference is found in patients who are unable to recognise faces raises an alternative possibility. It is at least conceivable that, contrary to what it is so compelling to suppose, the interference is produced in normal subjects in the same way as in people with prosopagnosia, by a process of which the subject is quite unaware (Stone and Davies, 1993).

David Milner and Melvyn Goodale (1995) argue that vision has two functions, representation of the world in *perception* and control of our *action* on the world. They also argue that these two functions are subserved by different neural pathways, the *ventral* and the *dorsal* streams (see also Jacob and Jeannerod, 2003). While the two visual pathways diverge from primary visual cortex they are not affected in the same way by damage to area V1. The ventral stream (vision for perception) depends almost totally on V1 for its inputs and so a patient with damage to V1 has a perceptual deficit. But cortical sites along the dorsal stream (vision for action) continue to receive visual information from sub-cortical structures and so some visual control of action may be preserved even in the absence of visual perception. Milner and Goodale suggest that the blindsight patient's responses may be explained in terms of cues that are provided by activity in

mechanisms of visuomotor control. On this account, blindsight is not perception without awareness but action without perception.

Impairments to the two functions of vision dissociate in both directions in patients with damage to one or the other visual pathway. But a mismatch between visual perception and visually controlled action can also be found in normal subjects who experience visual *illusions* that relate to the relative sizes of objects. It can happen, for example, that two objects are really the same size, one looks larger, yet a subject reaches towards and grasps the objects in just the same way, with a grip aperture that is appropriate to the real size of the objects. It can also happen that two objects are really different sizes, they look the same, yet a subject reaches differently – once again, with a grip aperture that is appropriate to the real size. This finding places some constraints on philosophical theories about representation and, particularly, about the non-conceptualised representational content of perceptual experiences (Peacocke, 2001). It is not obvious how to construct a theory of perceptual representation that is adequate to cases in which how the object really is (on the input side) and how the object is acted on (on the output side) are in harmony with each other, but there is a mismatch between both and the way that the object is perceived to be (Clark, 2001).

People with *unilateral visual neglect* fail to report visually presented stimuli on one side of space (usually the left side of space, following damage to the right hemisphere of the brain). When John Marshall and Peter Halligan (1988) asked a neglect patient to classify pairs of line drawings as same or different, she said that two drawings of a house, in one of which the left side of the house was on fire, were identical. But when, under forced-choice conditions, she was asked which house she would rather live in, she reliably chose the house that was not burning. Here there is a pattern similar to that in blindsight and in prosopagnosia with covert recognition. There is evidence that information is affecting the subject's performance even though the subject is unable to use that information normally for reporting, reasoning and planning. For this reason, blindsight and neglect are sometimes run together in philosophical discussion.

But unilateral neglect patients are not blind. Their problem is not primarily visual but, at least in part, *attentional* (Vallar, 1998; Aimola Davies, 2004). Their deficit is not so much in perception as in exploration. The difference between unilateral neglect and blindsight becomes vivid when we consider another experiment using pictures. When neglect patients are asked to copy line drawings, they typically produce a picture that is incomplete on the left side. When neglect patients are asked to identify and copy a drawing of a house with flames coming from the left side they identify the drawing as simply of a house and produce a picture from which left-side details, including the flames, are omitted. But when they are shown a picture of an object that can only be identified by the information on the left (e.g. a toothbrush or a garden rake with its head towards the left) at least some of the patients identify the object correctly and produce a picture with all the left-side details intact (Maguire, 2000).

Some people with unilateral neglect also have bizarrely false beliefs. Some deny ownership of a limb on the neglected side. Some claim to be able to move a limb that, in reality, is paralysed as a result of their brain damage. These are *delusions*: false beliefs that are firmly maintained despite their massive implausibility in the light of evidence that is available to the subject (Stone and Young, 1997; Davies et al., 2001). Although there are countless further points at which cognitive scientific research impacts on philosophical theory, this will be my final example. The points of contact in this case

include the relationship between experience and belief and the idea that attribution of beliefs is subject to some kind of rationality constraint.

Delusions occur, not only in neglect patients, but also in other cases of brain damage and in people with schizophrenia. The delusion that is most familiar in the philosophical literature is the *Capgras delusion* in which the subject maintains that a close relative (often the spouse) has been replaced by an impostor. Hadyn Ellis and Andy Young (1990) suggest that the Capgras delusion can be explained, at least in part, in terms of a neuropsychological impairment that is the mirror image of prosopagnosia with covert recognition. The subject recognises the presented face as looking just like the face of the spouse, but the affective response, and so the skin conductance response, that would usually accompany perception of a familiar face is absent, and so something seems wrong. The delusional belief is an attempt to make sense of this conflict. The problem with this explanation of the delusion is, of course, that there are much more plausible ways for the subject to make sense of this anomalous experience of the spouse's face. So it seems that we need to appeal to some *second factor* in the aetiology of the delusion – some impairment in the systems responsible for adopting, evaluating and revising beliefs. But it is difficult to say anything illuminating about the cognitive or computational nature of this second factor – and this is hardly surprising if, as Fodor suggests, we do not yet understand the processes of normal belief fixation (section 11.4).

References

Aimola Davies, A.M. 2004: Disorders of spatial orientation and awareness. In J. Ponsford (ed.), *Cognitive and Behavioral Rehabilitation: From Neurobiology to Clinical Practice*. New York: The Guilford Press, 175–223.

Armstrong, D.M. 1968: *A Materialist Theory of the Mind*. New York: Humanities Press.

Bain, A. 1893: The respective spheres and mutual helps of introspection and psychophysical experiment in psychology. *Mind*, 2, 42–53.

Baker, G.P. and Hacker, P.M.S. 1984: *Language, Sense and Nonsense: A Critical Investigation into Modern Theories of Language*. Oxford: Basil Blackwell.

Barkow, J.H., Cosmides, L. and Tooby, J. (eds) 1992: *The Adapted Mind: Evolutionary Psychology and the Generation of Culture*. Oxford: Oxford University Press.

Bechtel, W. and Abrahamsen, A. 1991: *Connectionism and the Mind*. Oxford: Blackwell.

Block, N. 1986: Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, 10, 615–78. Reprinted in Stich and Warfield (eds), 1994, 81–141.

Block, N. 1995a: The mind as the software of the brain. In E.E. Smith and D.N. Osherson (eds), *An Invitation to Cognitive Science, Volume 3: Thinking* (Second Edition). Cambridge, MA: MIT Press, 377–425.

Block, N. 1995b: On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, 18, 227–47.

Block, N. and Fodor, J. 1972: What psychological states are not. *Philosophical Review*, 81, 159–81.

Boden, M. (ed.) 1996: *The Philosophy of Artificial Life*. Oxford: Oxford University Press.

Boghossian, P. 1989: The rule-following considerations. *Mind*, 98, 507–49.

Branquinho, J. (ed.) 2001: *The Foundations of Cognitive Science*. Oxford: Oxford University Press.

Brentano, F. 1874: *Psychology from an Empirical Standpoint*. London: Routledge, 1995.

Brooks, R.A. 1991: Intelligence without representation. *Artificial Intelligence Journal*, 47, 139–59.

Bullinaria, J.A. and Chater, N. 1995: Connectionist modelling: Implications for cognitive neuropsychology. *Language and Cognitive Processes*, 10, 227–64.

Caramazza, A. 1986: On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition*, 5, 41–66.

Carruthers, P. 1996: *Language, Thought and Consciousness*. Cambridge: Cambridge University Press.

Carruthers, P. 1998: Conscious thinking: Language or elimination? *Mind and Language*, 13, 457–76.

Carruthers, P. 2003a: Moderately massive modularity. In A. O'Hear (ed.), *Minds and Persons* (Royal Institute of Philosophy Supplement 53). Cambridge: Cambridge University Press, 67–89.

Carruthers, P. 2003b: On Fodor's problem. *Mind and Language*, 18, 502–23.

Carruthers, P. 2004: The mind is a system of modules shaped by natural selection. In C. Hitchcock (ed.), *Contemporary Debates in the Philosophy of Science*. Oxford: Blackwell, 293–311.

Chomsky, N. 1965: *Aspects of the Theory of Syntax*. Cambridge, MA: MIT Press.

Chomsky, N. 1976: *Reflections on Language*. London: Fontana/Collins.

Chomsky, N. 1986: *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.
Chomsky, N. 1995: *The Minimalist Program*. Cambridge, MA: MIT Press.
Chomsky, N. 2000: *New Horizons in the Study of Language and Mind*. Cambridge: Cambridge University Press.
Chomsky, N. 2002: *On Nature and Language* (edited by A. Belletti and L. Rizzi). Cambridge: Cambridge University Press.
Churchland, P.M. 1990: Cognitive activity in artificial neural networks. In D.N. Osherson and E.E. Smith (eds), *An Invitation to Cognitive Science, Volume 3: Thinking*. Cambridge, MA: MIT Press, 199–227.
Churchland, P.M. and Churchland, P.S. 1996: Replies from the Churchlands. In R.N. McCauley (ed.), *The Churchlands and Their Critics*. Oxford: Blackwell Publishers, 217–310.
Churchland, P.S. 1986: *Neurophilosophy: Toward a Unified Science of the Mind/Brain*. Cambridge, MA: MIT Press.
Churchland, P.S. and Sejnowski, T.J. 1992: *The Computational Brain*. Cambridge, MA: MIT Press.
Clark, A. 1989: *Microcognition*. Cambridge, MA: MIT Press.
Clark, A. 1990: Connectionist minds. *Proceedings of the Aristotelian Society*, 90, 83–102.
Clark, A. 1993: *Associative Engines: Connectionism, Concepts, and Representational Change*. Cambridge, MA: MIT Press.
Clark, A. 1997: *Being There: Putting Brain, Body, and World Together Again*. Cambridge, MA: MIT Press.
Clark, A. 2001: Visual experience and motor action: Are the bonds too tight? *Philosophical Review*, 110, 495–519.
Coltheart, M. 1985: Cognitive neuropsychology and the study of reading. In M.I. Posner and O.S.M. Marin (eds), *Attention and Performance XI*. Hillsdale, NJ: Lawrence Erlbaum Associates, 3–37.
Coltheart, M. 1999: Modularity and cognition. *Trends in Cognitive Sciences*, 3, 115–20.
Coltheart, M. and Davies, M. 2003: Inference and explanation in cognitive neuropsychology. *Cortex*, 39, 188–91.
Coltheart, M. and Langdon, R. 1998: Autism, modularity and levels of explanation in cognitive science. *Mind and Language*, 13, 138–52.
Coltheart, M., Rastle, K., Perry, C., Langdon, R. and Ziegler, J. 2001: DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108, 204–56.
Cosmides, L. and Tooby, J. 1994: Origins of domain specificity: The evolution of functional organization. In L.A. Hirschfeld and S.A. Gelman (eds), *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press, 85–116.
Crane, T. 2001: *Elements of Mind*. Oxford: Oxford University Press.
Cummins, R. 1983: *The Nature of Psychological Explanation*. Cambridge, MA: MIT Press.
Danziger, K. 1980: The history of introspection reconsidered. *Journal of the History of the Behavioral Sciences*, 16, 241–62.
Davies, M. 1981a: *Meaning, Quantification, Necessity: Themes in Philosophical Logic*. London: Routledge and Kegan Paul.

Davies, M. 1981b: Meaning, structure, and understanding. *Synthese*, 48, 135–61.
Davies, M. 1986: Tacit knowledge, and the structure of thought and language. In C. Travis (ed.), *Meaning and Interpretation*. Oxford: Basil Blackwell, 127–58.
Davies, M. 1987: Tacit knowledge and semantic theory: Can a five per cent difference matter? *Mind*, 96, 441–62.
Davies, M. 1991: Concepts, connectionism, and the language of thought. In W. Ramsey, S. Stich and D. Rumelhart (eds), *Philosophy and Connectionist Theory*. Hillsdale, NJ: Lawrence Erlbaum Associates, 229–57.
Davies, M. 1992: Aunty's own argument for the language of thought. In J. Ezquerro and J.M. Larrazabal (eds), *Cognition, Semantics and Philosophy: Proceedings of the First International Colloquium on Cognitive Science*. Dordrecht: Kluwer Academic Publishers, 235–271.
Davies, M. 1995: Consciousness and the varieties of aboutness. In Macdonald and Macdonald (eds), 1995a, 356–92.
Davies, M. 2000a: Persons and their underpinnings. *Philosophical Explorations*, 3, 43–62.
Davies, M. 2000b: Interaction without reduction: The relationship between personal and sub-personal levels of description. *Mind and Society*, 1, 87–105.
Davies, M., Coltheart, M., Langdon, R. and Breen, N. 2001: Monothematic delusions: Towards a two-factor account. *Philosophy, Psychiatry and Psychology*, 8, 133–58.
de Gelder, B., Vroomen, J., Pourtois, G. and Weiskrantz, L. 1999: Non-conscious recognition of affect in the absence of striate cortex. *NeuroReport*, 10, 3759–63.
Dennett, D.C. 1969: *Content and Consciousness*. London: Routledge and Kegan Paul.
Dennett, D.C. 1971: Intentional systems. *Journal of Philosophy*, 68, 87–106. Reprinted in *Brainstorms: Philosophical Essays on Mind and Psychology*. Brighton: Harvester Press, 1978, 3–22.
Dennett, D.C. 1978: Artificial intelligence as philosophy and as psychology. In *Brainstorms: Philosophical Essays on Mind and Psychology*. Brighton: Harvester Press, 109–26.
Dennett, D.C. 1983: Styles of mental representation. *Proceedings of the Aristotelian Society*, 83, 213–26.
Dennett, D.C. 1987: *The Intentional Stance*. Cambridge, MA: MIT Press.
Dretske, F. 1986: Misrepresentation. In R.J. Bogdan (ed.), *Belief: Form, Content and Function*. Oxford: Oxford University Press, 17–36.
Dummett, M. 1975: What is a theory of meaning? In S. Guttenplan (ed.), *Mind and Language*. Oxford: Oxford University Press, 97–138. Reprinted in *The Seas of Language*. Oxford: Oxford University Press, 1993, 1–33.
Dummett, M. 1991: *The Logical Basis of Metaphysics*. Cambridge, MA: Harvard University Press.
Dunn, J.C. and Kirsner, K. 2003: What can we infer from double dissociations? *Cortex*, 39, 1–7.
Eilan, N., McCarthy, R. and Brewer, B. (eds) 1993: *Spatial Representation: Problems in Philosophy and Psychology*. Oxford: Blackwell Publishers (Second Edition, Oxford: Oxford University Press, 1999).
Ellis, H.D. and Young, A.W. 1990: Accounting for delusional misidentifications. *British Journal of Psychiatry*, 157, 239–48.

Elman, J.L., Bates, E., Johnson, M.S., Karmiloff-Smith, A., Parisi, D. and Plunkett, K. 1996: *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Evans, G. 1980: Things without the mind: A commentary upon chapter two of Strawson's *Individuals*. In Z. van Straaten (ed.), *Philosophical Subjects*. Oxford: Oxford University Press, 76–116. Reprinted in *Collected Papers*. Oxford University Press, 1985, 249–90.

Evans, G. 1981: Semantic theory and tacit knowledge. In S. Holtzmann and C. Leich (eds), *Wittgenstein: To Follow a Rule*. London: Routledge and Kegan Paul. Reprinted in *Collected Papers*. Oxford University Press, 1985, 322–42.

Evans, G. 1982: *The Varieties of Reference*. Oxford: Oxford University Press.

Farah, M.J. 1994: Neuropsychological inference with an interactive brain: A critique of the locality assumption. *Behavioral and Brain Sciences*, 17, 43–104.

Farber, I., Peterman, W. and Churchland, P.S. 2001: The view from here: The nonsymbolic structure of spatial representation. In J. Branquinho (ed.), 2001, 55–76.

Fodor, J.A. 1968a: *Psychological Explanation: An Introduction to the Philosophy of Psychology*. New York: Random House.

Fodor, J.A. 1968b: The appeal to tacit knowledge in psychological explanation. *Journal of Philosophy*, 65, 627–40.

Fodor, J.A. 1974: Special sciences. *Synthese*, 28, 77–115. Reprinted in *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Brighton: Harvester Press, 1981.

Fodor, J.A. 1975: *The Language of Thought*. New York: Crowell.

Fodor, J.A. 1981: Some notes on what linguistics is about. In N. Block (ed.), *Readings in Philosophy of Psychology, Volume 2*. London: Methuen, 197–207. Reprinted in Katz (ed.), 1985, 146–60.

Fodor, J.A. 1983: *The Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor, J.A. 1985: Fodor's guide to mental representation. *Mind*, 94, 77–100.

Fodor, J.A. 1987: *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, MA: MIT Press.

Fodor, J.A. 1989: Why should the mind be modular? In A. George (ed.), *Reflections on Chomsky*. Oxford: Blackwell, 1–22.

Fodor, J.A. 1998: *In Critical Condition*. Cambridge, MA: MIT Press.

Fodor, J.A. 2000: *The Mind Doesn't Work That Way*. Cambridge, MA: MIT Press.

Fodor, J.A. 2001: Language, thought and compositionality. *Mind and Language*, 16, 1–15.

Fodor, J.A. 2005: Reply to Steven Pinker 'So how *does* the mind work?'. *Mind and Language*, 20, 25–32.

Fodor, J.A., Bever, T.G. and Garrett, M.F. 1974: *The Psychology of Language*. New York: McGraw-Hill.

Fodor, J.A. and Pylyshyn, Z. 1988: Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3–71.

Fodor, J.D. 1995: Comprehending sentence structure. In L.R. Gleitman and M. Liberman (eds), *An Invitation to Cognitive Science (Second Edition) Volume 1: Language*. Cambridge, MA: MIT Press, 209–46.

Gardner, H. 1985: *The Mind's New Science: A History of the Cognitive Revolution*. New York: Basic Books.

Garfield, J.L. (ed.) 1987: *Modularity in Knowledge Representation and Natural-Language Understanding*. Cambridge, MA: MIT Press.
Grice, H.P. 1989: *Studies in the Way of Words*. Cambridge, MA: Harvard University Press.
Hamilton, W. 1859: *Lectures on Metaphysics and Logic*. Edinburgh: William Blackwood.
Hatfield, G. 2002: Psychology, philosophy, and cognitive science: Reflections on the history and philosophy of experimental psychology. *Mind and Language*, 17, 207–32.
Helmholtz, H. von 1867: *Handbuch der physiologischen Optik*. Leipzig: Leopold Voss.
Hinton, G.E., McClelland, J.L. and Rumelhart, D.E. 1986: Distributed representations. In Rumelhart, McClelland and the PDP Research Group, 1986, 77–109.
Hirschfeld, L. and Gelman, S. (eds), 1994: *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press.
Hornsby, J. 1997: *Simple Mindedness: In Defense of Naive Naturalism in the Philosophy of Mind*. Cambridge, MA: Harvard University Press.
Hornsby, J. 2000: Personal and sub-personal: A defence of Dennett's early distinction. *Philosophical Explorations*, 3, 6–24.
Humphreys, G.W. 1991: Review of Shallice, *From Neuropsychology to Mental Structure*. *Mind and Language*, 6, 202–14.
Jackson, F.C. and Pettit, P. 1988: Functionalism and broad content. *Mind*, 97, 381–400.
Jacob, P. and Jeannerod, M. 2003: *Ways of Seeing: The Scope and Limits of Visual Cognition*. Oxford: Oxford University Press.
Juola, P. and Plunkett, K. 2000: Why double dissociations don't mean much. In G. Cohen, R.A. Johnston and K. Plunkett (eds), *Exploring Cognition: Damaged Brains and Neural Networks: Readings in Cognitive Neuropsychology and Connectionist Modelling*. Hove, East Sussex: Psychology Press, 319–27.
Katz, J.J. (ed.) 1985: *The Philosophy of Linguistics*. Oxford: Oxford University Press.
Kripke, S.A. 1982: *Wittgenstein on Rules and Private Language*. Oxford: Basil Blackwell.
Ladefoged, P. and Broadbent, D.E. 1960: Perception of sequences in auditory events. *Quarterly Journal of Experimental Psychology*, 13, 162–70.
Levine, J. 1983: Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly*, 64, 354–61.
Lewis, D.K. 1966: An argument for the identity theory. *Journal of Philosophy*, 63, 17–25.
Lewis, D.K. 1970: How to define theoretical terms. *Journal of Philosophy*, 67, 427–46.
Lewis, D.K. 1972: Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50, 249–58.
Loar, B. 1981: *Mind and Meaning*. Cambridge: Cambridge University Press.
Lycan, W.G. 1986: Tacit belief. In R.J. Bogdan (ed.), *Belief: Form, Content and Function*. Oxford: Oxford University Press, 61–82.
Macdonald, C. and Macdonald, G. (eds) 1995a: *Philosophy of Psychology: Debates on Psychological Explanation*. Oxford: Blackwell Publishers.
Macdonald, C. and Macdonald, G. (eds) 1995b: *Connectionism: Debates on Psychological Explanation*. Oxford: Blackwell Publishers.
Maguire, A.M. 2000: Reducing neglect by introducing ipsilesional global cues. *Brain and Cognition*, 43, 328–32.

Manson, N. 2000: 'A tumbling-ground for whimsies'? The history and contemporary role of the conscious/unconscious contrast. In T. Crane and S. Patterson (eds), *History of the Mind-Body Problem*. London: Routledge, 148–68.
Marcus, G. 1995: The acquisition of the English past tense in children and multilayered connectionist networks. *Cognition*, 56, 271–9.
Marr, D. 1982: *Vision*. New York: W.H. Freeman and Co.
Marshall, J.C. and Halligan, P.W. 1988: Blindsight and insight in visuo-spatial neglect. *Nature*, 336, 766–7.
McClelland, J.L., Rumelhart, D.E. and Hinton, G.E. 1986: The appeal of parallel distributed processing. In Rumelhart, McClelland and the PDP Research Group, 1986, 3–44.
McClelland, J.L., Rumelhart, D.E. and the PDP Research Group 1986: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition Volume 2: Psychological and Biological Models*. Cambridge, MA: MIT Press.
McDowell, J. 1994: The content of perceptual experience. *Philosophical Quarterly*, 44, 190–205.
McLaughlin, B.P. 1993: The connectionism/classicism battle to win souls. *Philosophical Studies*, 70, 45–72.
McLaughlin, B.P. and Warfield, T.A. 1994: The allure of connectionism reexamined. *Synthese*, 101, 365–400.
McLeod, P., Plunkett, K. and Rolls, E. T. 1998: *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford: Oxford University Press.
Mill, J.S. 1865: *Examination of Sir William Hamilton's Philosophy*. London: Longmans, Green and Co.
Miller, A. 1997: Tacit knowledge. In B. Hale and C. Wright (eds), *A Companion to the Philosophy of Language*. Oxford: Blackwell Publishers, 146–74.
Miller, A. and Wright, C. (eds) 2002: *Rule-Following and Meaning*. Chesham: Acumen Publishing Limited.
Miller, G.A. 2003: The cognitive revolution: A historical perspective. *Trends in Cognitive Sciences*, 7, 141–4.
Milner, A.D. and Goodale, M.A. 1995: *The Visual Brain in Action*. Oxford: Oxford University Press.
Patterson, K.E., Seidenberg, M.S. and McClelland, J.L. 1989: Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In R.G.M. Morris (ed.), *Parallel Distributed Processing: Implications for Psychology and Neurobiology*. Oxford: Oxford University Press, 131–81.
Peacocke, C. 1986: Explanation in computational psychology: Language, perception and level 1.5. *Mind and Language*, 1, 101–23.
Peacocke, C. 1989: When is a grammar psychologically real? In A. George (ed.), *Reflections on Chomsky*. Oxford: Basil Blackwell, 111–30.
Peacocke, C. 1992: *A Study of Concepts*. Cambridge, MA: MIT Press.
Peacocke, C. 2001: Does perception have a nonconceptual content? *Journal of Philosophy*, 98, 239–64.
Pietroski, P. 2000: *Causing Actions*. Oxford: Oxford University Press.
Pietroski, P. and Rey, G. 1995: When other things aren't equal: Saving ceteris paribus laws from vacuity. *British Journal for the Philosophy of Science*, 46, 81–110.

Pinker, S. 1995: Language acquisition. In L.R. Gleitman and M. Liberman (eds), *An Invitation to Cognitive Science, Volume 1: Language* (Second Edition). Cambridge, MA: MIT Press, 135–82.
Pinker, S. 1997: *How the Mind Works*. New York: W.W Norton.
Pinker, S. 2005a: So how *does* the mind work? *Mind and Language*, 20, 1–24.
Pinker, S. 2005b: A reply to Jerry Fodor on how the mind works. *Mind and Language*, 20, 33–8.
Pinker, S. and Prince, A. 1988: On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28, 73–193.
Place, U.T. 1956: Is consciousness a brain process? *British Journal of Psychology*, 47, 44–50.
Plaut, D.C. 1995: Double dissociation without modularity: Evidence from connectionist neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, 17, 291–321.
Plaut, D.C. 2003: Interpreting double dissociations in connectionist networks. *Cortex*, 39, 138–41.
Plaut, D.C., McClelland, J.L., Seidenberg, M.S. and Patterson, K. 1996: Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, 103, 56–115.
Plunkett, K. and Marchman, V. 1993: From rote learning to system building: Acquiring verb morphology in children and connectionist nets. *Cognition*, 48, 21–69.
Plunkett, K. and Marchman, V. 1996: Learning from a connectionist model of the acquisition of the English past tense. *Cognition*, 61, 299–308.
Port, R.F. and van Gelder, T (eds) 1995: *Mind as Motion: Explorations in the Dynamics of Cognition*. Cambridge, MA: MIT Press.
Pouget, A. and Sejnowski, T.J. 2001: Simulating a lesion in a basis function model of spatial representations: Comparison with hemineglect. *Psychological Review*, 108, 653–73.
Putnam, H. 1967: The nature of mental states. Originally titled 'Psychological predicates', in W.H. Capitan and D.D. Merrill (eds), *Art, Mind, and Religion*. Pittsburgh: University of Pittsburgh Press, 37–48.
Quine, W.V.O. 1953: The problem of meaning in linguistics. In *From a Logical Point of View*. New York: Harper and Row, 47–64. Reprinted in Katz (ed.), 1985, 48–61.
Quine, W.V.O. 1972: Methodological reflections on current linguistic theory. In D. Davidson and G. Harman (eds), *Semantics of Natural Language*. Dordrecht: Reidel, 442–54. Reprinted in G. Harman (ed.), *On Noam Chomsky: Critical Essays*. New York: Anchor Press/Doubleday, 1974, 104–17.
Ramsey, W., Stich, S. and Garon, J. 1990: Connectionism, eliminativism and the future of folk psychology. In J.E. Tomberlin (ed.), *Philosophical Perspectives Volume 4: Action Theory and Philosophy of Mind*. Atascadero, CA: Ridgeview Publishing Company, 499–533.
Roeper, T. and Williams, E. (eds) 1987: *Parameter Setting*. Dordrecht: Kluwer Academic Publishers.
Rolls, E.T. 1989 Parallel distributed processing in the brain: Implications of the functional architecture of neuronal networks in the hippocampus. In R.G.M. Morris (ed.), *Parallel Distributed Processing: Implications for Psychology and Neurobiology*. Oxford: Oxford University Press, 286–308.

Rosenthal, D.M. 2002: Explaining consciousness. In D.J. Chalmers (ed.), *Philosophy of Mind: Classical and Contemporary Readings*. Oxford: Oxford University Press, 406–21.

Rumelhart, D.E. and McClelland, J.L. 1986: On learning the past tenses of English verbs. In McClelland, Rumelhart and the PDP Research Group, 1986, 216–71.

Rumelhart, D.E., Hinton, G.E. and Williams, R.J. 1986: Learning internal representations by error propagation. In Rumelhart, McClelland and the PDP Research Group, 1986, 318–62.

Rumelhart, D.E., McClelland, J.L. and the PDP Research Group 1986: *Parallel Distributed Processing: Explorations in the Microstructure of Cognition Volume 1: Foundations*. Cambridge, MA: MIT Press.

Ryle, G. 1949: *The Concept of Mind*. Harmondsworth: Penguin Books.

Schiffer, S. 1993: Actual-language relations. In J.E. Tomberlin (ed.), *Philosophical Perspectives, 7: Language and Logic*, Atascadero, CA: Ridgeview Publishing Company, 231–58.

Searle, J.R. 1990: Consciousness, explanatory inversion, and cognitive science. *Behavioral and Brain Sciences*, 13, 585–96. Reprinted in C. Macdonald and G. Macdonald (eds), 1995a, 331–55. Page references to reprinting.

Searle, J.R. 1992: *The Rediscovery of the Mind*. Cambridge, MA: MIT Press.

Seidenberg, M.S. 1988: Cognitive neuropsychology and language: The state of the art. *Cognitive Neuropsychology*, 5, 403–26.

Seidenberg, M.S. 1989: Visual word recognition and pronunciation: A computational model and its applications. In W. Marslen-Wilson (ed.), *Lexical Representation and Process*. Cambridge, MA: MIT Press, 25–74.

Seidenberg, M.S. and McClelland, J.L. 1989: A distributed, developmental model of word recognition and naming. *Psychological Review*, 96, 523–68.

Shallice, T. 1988: *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press.

Smart, J.C.C. 1959: Sensations and brain processes. *Philosophical Review*, 68, 141–56.

Smith, N.V. 2004: *Chomsky: Ideas and Ideals* (Second Edition). Cambridge: Cambridge University Press.

Smolensky, P. 1988: On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11, 1–74. Reprinted in Macdonald and Macdonald (eds), 1995b, 28–89.

Sperber, D. 2002: In defense of massive modularity. In I. Dupoux (ed.), *Language, Brain and Cognitive Development: Essays in Honor of Jacques Mehler*. Cambridge, MA: MIT Press, 47–57.

Stich, S. 1972: Grammar, psychology and indeterminacy. *Journal of Philosophy*, 79, 799–818. Reprinted in Katz (ed.), 1985, 126–45.

Stich, S. 1978: Beliefs and subdoxastic states. *Philosophy of Science*, 45, 499–518.

Stich, S. 1996: *Deconstructing the Mind*. Oxford: Oxford University Press.

Stich, S.P. and Warfield, T.A. (eds) 1994: *Mental Representation: A Reader*. Oxford: Blackwell.

Stoerig, P. and Cowey, A. 1992: Wavelength sensitivity in blindsight. *Brain*, 115, 425–44.

Stone, T. and Davies, M. 1993: Cognitive neuropsychology and the philosophy of mind. *British Journal for the Philosophy of Science*, 44, 589–622.

Stone, T. and Davies, M. 1999: Autonomous psychology and the moderate neuron doctrine. *Behavioral and Brain Sciences*, 22, 849–50.

Stone, T. and Young, A.W. 1997: Delusions and brain injury: The philosophy and psychology of belief. *Mind and Language*, 12, 327–64.

Strawson, P.F. 1959: *Individuals: An Essay in Descriptive Metaphysics*. London: Methuen.

Thomas, N. 2001: Mental imagery. In E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2001 Edition).
URL = <http://plato.stanford.edu/archives/win2001/entries/mental-imagery/>.

Vallar, G. 1998: Spatial hemineglect in humans. *Trends in Cognitive Sciences*, 2, 87–97.

Weiskrantz, L. 1986: *Blindsight: A Case Study and Implications*. Oxford: Oxford University Press.

Weiskrantz, L. 1997: *Consciousness Lost and Found*. Oxford: Oxford University Press.

Wittgenstein, L. 1953: *Philosophical Investigations*. Oxford: Basil Blackwell (Third Edition, 1967).

Wright, C. 1981: Rule-following, objectivity, and the theory of meaning. In S. Holtzmann and C. Leich (eds), *Wittgenstein: To Follow a Rule*. London: Routledge and Kegan Paul, 99–117.

Wright, C. 1987: Theories of meaning and speakers' knowledge. In *Realism, Meaning and Truth*. Oxford: Blackwell; Second Edition, 1993, 204–38.

Wright, C. 1989: Wittgenstein's rule-following considerations and the central project of theoretical linguistics. In A. George (ed.), *Reflections on Chomsky*. Oxford: Blackwell, 233–64.

Young, A.W. 1998: *Face and Mind*. Oxford: Oxford University Press.