Models of memory: Wittgenstein and cognitive science

DAVID G. STERN
Department of Philosophy, 269 English–Philosophy Building, University of Iowa, Iowa City, IA 52242, USA

ABSTRACT *The model of memory as a store, from which records can be retrieved, is taken for granted by many contemporary researchers. On this view, memories are stored by memory traces, which represent the original event and provide a causal link between that episode and one's ability to remember it. I argue that this seemingly plausible model leads to an unacceptable conception of the relationship between mind and brain, and that a non-representational, connectionist, model offers a promising alternative. I also offer a new reading of Wittgenstein's paradoxical remarks about thought and brain processes: as a critique of the cognitivist thesis that information stored in the brain has a linguistic structure and a particular location. On this reading, Wittgenstein's criticism foreshadows some of the most promising contemporary work on connectionist models of neural functioning.*

Since the time of the ancient Greeks, our memory has been compared to writing on a wax tablet or a storehouse which contains many written records; more recently, analogies have been drawn with information storage in a computer. But while a conventional digital computer will always retrieve an item from its memory in the same form, our recall of an event such as a family gathering long ago is often a much more fluid affair: a great deal depends on the context in which one tries to remember, and how much prompting goes on. Recalling a distant event is often a reconstructive activity, one in which there is no sharp dividing line between factual recall and piecing together the most plausible and relevant story (Neisser, 1982; Neisser & Winograd, 1988, chs 3–5). Our memory for faces is another example that does not sit well with the storehouse model: most people are much better at recognizing an acquaintance than remembering his or her appearance, let alone describing that person on the basis of free recall.

Advocates of the storehouse analogy treat these cases as anomalies which can be accomodated by postulating regulations which control access to the storehouse. But these are, I believe, telling instances of the limitations of that analogy. For memory is, after all, an ability to think and act in certain ways; we do not have to assume that these actions must be explained in terms of memory traces, or that the words we utter are a "translation of something that was there before" (Wittgenstein, 1980a, p. 736). [1] Storehouse theorists take these traces for granted, for they think

of the trace as performing two vital functions: it both represents the original event and provides a causal link between that episode and my current capacity to remember it. In this paper, I argue that these seemingly plausible commitments create serious problems for contemporary versions of the storehouse model, and that these problems can best be avoided by giving up the model. This leads to a discussion of some of Wittgenstein's remarks about brain processes and memory traces and their implications for the current debate over cognitivist and connectionist theories of neural functioning. Wittgenstein's treatment of memory is valuable because he questions the widespread assumption that memory must be a matter of *storing something*: he suggests that memory also fulfils other purposes, and that these other purposes may be quite compatible with an account which dispenses with the notion of storage altogether:

> Memory can be compared with a storehouse only so far as it fulfills the same purpose. Where it doesn't, we couldn't say whether the things stored up may not constantly change their nature and so couldn't be stored at all. (Wittgenstein, 1935–6, p. 17)

The principal advocates of the storehouse model today are cognitivists. Cognitivism takes many forms; its proponents work in fields such as philosophy of mind, psychology, artificial intelligence and neurophysiology. They share a commitment to the goal of explaining the operation of the mind as a system of mental representations, representations which are realized by certain structures in the brain. Here is a clear and explicit statement of this thesis from Fodor's *The Language of Thought:*

> [W]hat one tries to do in cognitive psychology is to explain the propositional attitudes of the organism by reference to its (hypothetical) computational operations . . . So, for example, assume that remembering *P* is one of the relations that a reasonable psychological theory might acknowledge between an organism and (the proposition) *P*. Suppose, too, that storing *F* is one of the computational relations that a reasonable psychological theory might acknowledge between an organism and the internal formula *F*. It would then be (at best) a contingent truth—precisely the kind of contingent truth that cognitive psychology seeks to formulate—that the organism remembers *P* if, and only if, the organism stores *F* . . . We have no *a priori* guaranteed that all the cognitive states of an organism *can* be explained by reference to the special subset which consists of relations between the organism and formulae of its internal representational system. All we know *a priori* is that such cognitive psychology as is currently available assumes that this is true. (1975, pp. 76, 77, fn. 17)

On this account, psychology has to postulate that every thought corresponds to an underlying computational process. Fodor takes the notion of a computational process quite literally: he thinks of the brain as a computer processing internal formulae, written in an innate language of thought. The formulae are realized by certain patterns of physical activity in the brain. This 'internal representational system' plays the same role in Fodor's account of the brain as a 'machine language'

in a computer: it is the language in which the elementary parts of thought are implemented by the brain. He holds that we must assume that any proposition *P* we might entertain corresponds to some formula *F* in the language of thought.

In the passage just quoted, Fodor emphasizes the need for two distinct levels of analysis. First, there is the intentional level of everyday psychology, in which we talk about people's beliefs, memories, desires and other such intentional states—what Fodor calls "the propositional attitudes of the organism". Second, there is the computational level, postulated by cognitive psychology, the internal program which is supposed to explain what goes on at the first, intentional, level. This program is specified as a series of operations which could be instantiated by any physical system with a sufficient degree of complexity, whether it be a human brain, a von Neumann computer, or an exotic alternative such as a new computer architecture or an extraterrestrial brain. So, in addition to the two levels which Fodor stresses, there is also an implicit third—physical—level of analysis which concerns itself with how the system in question is constructed. Fodor does not stress this level of analysis, for he holds that it makes no difference which causally efficacious physical system implements the program.

While these programmatic commitments are widely accepted—Pylyshyn calls the assumption that there are intentional, computational and physical levels of explanation *"the basic assumption of cognitive science"* (Pylyshyn, 1984, p. 131, italics in original)—their dualistic implications are rarely recognized. For cognitivists think of themselves as sophisticated materialists, scientific philosophers who work on problems in the foundations of psychology and artificial intelligence. But the problem of connecting the causally effective physical level of explanation with the cognitively significant intentional level is a restatement of the problem of connecting mind and matter, the central problem of Cartesian dualism, rephrased in the idiom of 'levels of explanation'. Descartes argued that there is a real distinction between mind and matter: the mind is a thinking non-spatial thing, matter is spatial and does not think. While Descartes recognized that my mind is so "tightly bound to my body and so 'mixed up' with it that we form a single thing," (1986, p. 159) his sharp conceptual separation between mind and matter made it impossible for him to explain how this intimate relationship is possible. If minds are not in space, how can they affect or be affected by bodies in space? In reply, Descartes proposed that messages from every part of the body converge on the pineal gland where they interact with the mind. But postulating a place where mind meets matter is a *deus ex machina* which does not solve the problem, for it does nothing to explain how this interaction is possible.

Cognitive science is committed to a parallel division between the physical and the intentional, and corresponding difficulties in putting the two together again. For the physical level of explanation is a direct descendant of Descartes' project of constructing scientific laws which will allow us to predict the behaviour of matter in motion. While contemporary physics, chemistry and biology are very different from the geometric physics Descartes envisaged, they conform to his conviction that physical science should not invoke psychological properties in the construction of its theories. Similarly, the intentional level of explanation is a direct descendant of

Descartes' conception of the mind. It concerns itself with what is, or could be, present to consciousness: our thoughts, desires and intentions. Of course, the two dualisms are significantly different: Descartes distinguished between two sorts of substances, while the cognitivist sets out a methodological distinction between intentional and physical levels of explanation. But what these positions do have in common is that they both treat mind and matter as conceptually independent. While few cognitivists follow Descartes in maintaining that the mind could survive without any physical embodiment, they do look at the mind as a program which could be instantiated by any machine which can instantiate that program. As Searle puts it, this is a view on which "what it is specifically mental about the mind has no intrinsic connection with the actual properties of the brain" (1980, p. 424). The point, then, of the computational level of explanation is to provide a way of bringing together the intentional and the physical, mind and brain, by incorporating elements of each. It is supposed to show how the smallest significant units of the programs which make up the mind are mechanically implemented. As Stoutland has observed, "the point of this third [computational] level seems to be, precisely, to show that . . . the computational level can link up these disparate levels. It is, one might say, the pineal gland of contemporary mentalism" (1988, p. 48). Just as Descartes invoked the pineal gland as the meeting place for mind and matter, cognitivism enlists the language of thought and the computational level of explanation to link up the physical and intentional levels of explanation. Ultimately, it is this unrecognized dualism which is responsible for the dogma of the language of thought and a distorted conception of mind and brain.

Perhaps the simplest reply to those who still think that we must make Fodorian assumptions is to point to researchers who have already given them up, such as Rumelhart & Norman. Their connectionist model of the brain dispenses with the assumption that information is stored by a formula in an internal code with a specific location.

> Rather than imagining that particular neural nets encode particular pieces of information, this view has it that information is stored in the *relationships* among the units and that each unit participates in the encoding of many, many memories. (1981, p. 3)

Connectionists study mathematical models of neurological functioning in which brain structure is represented by a network of multiply connected nodes. Nodes and connections are initially given small random values on an arbitrarily chosen scale, and a single invariant function determines the value of each node at time $t+1$ in terms of its value at time t and the value of its neighbours and their respective connection strengths at time t. By varying the input to the network at certain assigned input nodes and observing the effect of those variations on the output nodes, one adjusts the values of the connections in order to teach the network to respond to each of a variety of inputs with the appropriate output. Or the network can simply be given an appropriate algorithm with which it can teach *itself.*

One way of understanding this approach is to think of the brain as an electrical network. Each neuron is one node in the network, wired up to other neurons by

axons, long connecting filaments. The strength of these connections varies and this determines the degree to which connected neurons affect each other. The whole system vibrates at many different frequencies at once; retrieving information is analogous to tuning in to a particular frequency. Like a vibrating guitar string, the brain's activity can be analysed into a large number of separate notes. Each memory is a product of the activity of the total neural network:

> Information is not stored anywhere in particular. Rather it is stored everywhere. Information is better thought of as 'evoked' than 'found'. (Rumelhart & Norman, 1981, p. 3)

Connectionism is far from being a unified school of thought, and this is not the only construal of its significance. [2] But on this model, a system can learn to respond correctly to a wide range of inputs without that ability being represented by *any* pattern in the network. For this is an account on which there are *no* representations in the brain and no need for any computational level. Here, I use 'representation' to mean not only the simple formulae which make up the basic units of Fodor's internal computational system and whatever can be constructed out of them, but also any other patterns of activity which are proposed as part of a level of explanation which mediates between the physical and the intentional. Such a dynamic model of memory is certainly not simply a matter of replacing Fodor's formula *F*, a localized representation, by what has been called a 'distributed representation'. But in his recent responses to connectionism, Fodor simply argues that we must give a role to syntax and logical structure at the computational level if we are to explain how thoughts have the content they do; he seems unable to give serious consideration to the idea of dispensing with the computational level altogether (1987, pp. 135–154; Fodor & Pylyshyn, 1988). Thus Fodor's critique of Smolensky's treatment of connectionism ignores his explicitly anti-representationalist statements and instead takes the notion of a computational level for granted (see Dreyfus & Dreyfus, 1988b).

The rules for the operation of the neural net only concern the conditions under which impulses will be transmitted between units; it is the history of the net—the input it has received in the past and the patterns of activity which it has generated—which determines its present characteristics, namely its current connection strengths. While connectionist nets are, at present, typically simulated on digital computers via programs, this is purely a matter of convenience and done because parallel distributed processing is still at an early stage of development. In principle, there is no need to write a program in order to run a network, for the net is established by 'training' the network to respond to certain inputs with the appropriate outputs. As a result, there need be no similarity between the state of the system on different occasions when a particular output is produced over and above the occurrence of the same output. Of course, there are connectionist nets in which a given hidden node or set of nodes is on if and only if the input has a certain feature, and in such a net it may well be appropriate to treat the node or nodes in question as representing that feature. But other nets, especially those with a relatively small number of hidden nodes, do not display such characteristics, and in

these cases, to say that the system as a whole represents the features it recognizes, or that it somehow constitutes a 'distributed representation' is to gloss over the fact that the initial notion of a representation, a copy with a determinate location, has been jettisoned. This development is a striking fulfilment of the following prediction of Wittgenstein's:

> [N]othing seems more possible to me than that people some day will come to the definite opinion that there is no copy in either the physiological or nervous systems which corresponds to a *particular* thought, or a *particular* idea, or memory. (1982, p. 504)

While the connectionist research programme is still in its early stages, it has already proven itself a promising alternative to conventional models of memory. As well as dispensing with the need for postulating a language of thought, it offers the prospect of a resolution of some of the other outstanding difficulties which have dogged cognitivist accounts of the mind. It does away with the need to explain why memory is deeply context dependent, why remembering is so unlike the purely factual recall one would expect if memory were simply a matter of bringing up a stored memory trace. It implies that remembering is not something we passively undergo but something we do, an activity. Because it regards memory as arising out of a system of distributed resonances, there is no need to add in the ability to see similarities and generalize in order to collate disparate memories; instead, similarities may emerge as related items interact, reinforcing their common aspects. Finally, it holds out the hope of a naturalistic account of human intelligence which draws on our knowledge of the brain. This list only hints at the potential advantages which a connectionist approach has to offer. Connectionism itself is primarily a programme within artificial intelligence and workers in the field often make the claim of 'neural plausibility' without giving it much concrete backing. But Edelman's recent work on the development of brain structure indicates that some of the same ideas are also at work in biology and neurophysiology (see Edelman (1985); Rosenfeld (1986, 9 October); Young *et al.* (1987, 12 March)). Of course, there is no guarantee that connectionism's promise will be borne out in practice, and it is entirely possible that it will soon run into as many problems as its competitors. But connectionism can be of value in opening our eyes to alternatives to cognitivism, regardless of whether or not it proves to be a successful research programme.

In the *Remarks on the Philosophy of Psychology*, written shortly after the second world war, Wittgenstein wrote:

> No supposition seems to me more natural than that there is no process in the brain correlated with associating or with thinking; so that it would be impossible to read off thought-processes from brain-processes. I mean this: if I talk or write there is, I assume, a system of impulses going out from my brain and correlated with my spoken or written thoughts. But why should the *system* continue further in the direction of the centre? Why should this order not proceed, so to speak, out of chaos? (1980a, p. 903)

Certainly, when one acts, impulses travel along the nerves connecting the brain to

the parts of the body which act. However, Wittgenstein thinks there is no reason to believe that there is a corresponding pattern in the brain, the 'centre' of the nervous system. In other words, he questions the assumption that we will be able to identify any process in the brain which initiated the action, let alone a formula of the kind Fodor demands. He illustrates this question with a parable:

> The case would be like the following—certain kinds of plants multiply by seed, so that a seed always produces a plant of the same kind as that from which it was produced—but *nothing* in the seed corresponds to the plant which comes from it; so that it is impossible to infer the properties or structure of the plant from those of the seed that it comes out of—this can only be done from the *history* of the seed. So an organism might come into being even out of something quite amorphous, as it were causelessly; and there is no reason why this should not really hold for our thoughts, and hence for our talking and writing. (1980a, p. 903)

Wittgenstein is not simply saying that we might be unable to distinguish seeds from different plants on the basis of their present chemical composition and structure. He is also asking us to suppose that it would be impossible to do so, because "*nothing* in the seed corresponds to the plant which comes from it". In that case, the only way of identifying the seed would be on the basis of what he calls its "history": which plant it came from, or which plant it gives rise to.

In his next remark, Wittgenstein considers the wider implications of the idea that our brain states could be as amorphous as the seeds in the parable:

> It is thus perfectly possible that certain psychological phenomena *cannot* be investigated physiologically, because physiologically nothing corresponds to them.
>
> I saw this man years ago: now I have seen him again, I recognize him, I remember his name. And why does there have to be a cause of this remembering in my nervous system? Why must something or other, whatever it may be, be stored-up there *in any form*? Why *must* a trace have been left behind? Why should there not be a psychological regularity to which *no* physiological regularity corresponds? If this upsets our concepts of causality then it is high time they were upset. (1980a, pp. 904–905)

Clearly, this is an alternative to Fodor's assumption that an organism remembers a proposition if and only if it stores a representation. But the price may look unacceptably high. For in "upsetting our concepts of causality," it appears to fly in the face of standard scientific accounts of many everyday phenomena. How could there be seeds in which "*nothing* in the seed corresponds to the plant which comes from it"? Don't we know that plants and seeds have a common genetic structure, and that this is the cause of the seed's turning into the plant that it does? How could there be brains in which "*nothing* in the brain corresponds to the thought which comes from it"? Don't our brains have a neurological structure which causes us to

think as we do? These difficulties have led McGinn to read Wittgenstein as asserting that:

> there is no good *a priori* reason to believe that for every mental state there is a corresponding physical state, and states of the nervous system could play no role in constraining our application of psychological concepts to people—such states play no part in the language-game of describing people psychologically. He even suggests that for all we know our behaviour could proceed from internal physical chaos, and that psychological and behavioural differences between people need not correlate with any underlying physical differences: in short, psychological phenomena may turn out to have no physical explanation. (1984, p. 112)

As a result, McGinn concludes that Wittgenstein is committed to denying that mental states are causally supervenient on physical states. For on this account, two people might hear a sound which causes each of them to be in the very same state of neuro-chaos, while only one of them understands what's been said. This objection is, however, open to the response that the situation McGinn describes can easily be made sense of: on one occasion, the sound occurs as part of a conversation in English, and the hearer does understand it; on the other occasion, the second hearer hears the very same sounds uttered in the course of a discussion in a foreign language, and in responding to them in the same way as the first hearer, fails to understand what's been said.

In a footnote, McGinn points out that the example of the seeds shows that Wittgenstein's "claim about psycho-physiological correspondence reflects a more general thesis about causation and explanation, and does not derive from the specific character of psychological concepts" (1984, p. 114, n. 28). In other words, McGinn reads Wittgenstein as committed to an unacceptable "general thesis about causation and explanation" on which

> some physical events have no physical explanation, *viz.* those that issue from psychological differences which correspond to no physiological differences. As some sort of physicalist I find this consequence intolerable (though I confess I would find it hard to *prove* to someone that it is). Those of my readers who share this degree of physicalism will accordingly feel bound to dissent from Wittgenstein's general stance on the mind–body relation, and will want to say that there *is* some state of the brain corresponding to each state of understanding which is the causal source of linguistic behaviour. . . . we will want to claim that each state of understanding has *some* physical realization, and that the causal powers of the former are in some way grounded in those of the latter. (1984, p. 113)

In short, McGinn takes Wittgenstein to hold that thoughts need have no physical basis whatsoever. The problem which worries McGinn is encapsulated in Wittgenstein's claim that "it is high time that our concepts of causality were upset." McGinn does suggest a motivation for Wittgenstein's claim, namely the idea that

behaviour is the "ultimate criterion for understanding," for "our certification of someone as understanding signs is *independent* of what happens physically inside him" (1984, p. 114). In other words, if a person acts appropriately in response to our words, then that is enough for us to say that the person understands; the condition of his or her internal organs is simply irrelevant. Evidence for this interpretation can be found in Wittgenstein's writings. The following passage is a good example:

> Thinking in terms of physiological processes is extremely dangerous in connection with the clarification of conceptual problems in psychology. Thinking in physiological hypotheses deludes us sometimes with false difficulties, sometimes with false solutions. The best prophylactic against this is the thought that I don't know at all whether the humans I am acquainted with actually have a nervous system. (1980a, p. 1063)

Taken out of context, this sounds like an extreme and unmotivated scepticism about physiology, but what motivates this remark is his conviction that conceptual problems will not be solved by science. This motive emerges very clearly in a passage where Wittgenstein approaches scepticism about other minds by asking us to imagine what would happen if we had a neuroscanner which permitted us to watch the functioning of other people's nervous systems (1980b, p. 702; 1981, p. 557). What further evidence could one ask for? Of course, some people might still ask whether the person under observation really feels pain when the neuroscanner says they do. But it is easy to imagine that what those people saw on the neuroscanner would determine their reaction without their having any qualms about it. And this is the point of this thought experiment for Wittgenstein: for the same point can be made about 'outer behaviour', what we actually see other people do: "This observation fully determines their attitude to others and doubt does not occur" (Wittgenstein, 1981, p. 557).

McGinn's reply to the idea that understanding might proceed from neural chaos is that we are being asked to entertain an impossible supposition: such a state of affairs is logically possible, it can be imagined, but as a matter of fact, it is scientifically impossible. Certainly, McGinn's reading is plausible, and I am just enough of a physicalist to agree with him this far: it would be astounding if it turned out that thoughts have no physical basis whatsoever, if that is to mean that there is no connection between the state of one's brain and the state of one's mind. McGinn's problem raises a larger and more general difficulty in the philosophy of mind. That difficulty is a matter of reconciling two theses which can seem both undeniable and incompatible. The first, 'ontological monism', the thesis that mental events are causally supervenient on physical events, is McGinn's minimal physicalism. The second, 'conceptual dualism', the thesis that mental language and physiological theories are incommensurable, amounts to a commitment to the duality of folk psychology and scientific theory. The first thesis states that we live in one world; the second that we must think of it in two radically different ways. The two are not contradictory, but there is a tension between them. McGinn reads Wittgenstein as denying ontological monism in order to protect conceptual dualism, and thus folk psychology, from reductionism.

Roughly, Wittgenstein's response to this tension is that it arises out of a misunderstanding of our language, which first leads us to divide ourselves into inner states and outer behaviour and then worry about how to put the two back together again. But this is probably too brief to be much help. (For some extended exposition of this train of thought, see Cook (1969) and Stoutland (1988).) Let us start by looking at a passage from the *Philosophical Investigations* which initially appears to provide strong support for McGinn's reading. In p. 157, Wittgenstein discusses the case of a person who is trained as a "reading machine"—taught to read aloud, to produce the appropriate sounds, by following a sequence of written characters. He compares this with a machine such as a pianola, where we would be entitled to say the pianola started to read once the internal mechanism was properly set up, while in the case of a human "reading-machine", the question whether the person is reading is determined by behavioural criteria, such as the number of mistakes. Wittgenstein's interlocutor replies that this is only because we know too little about what goes on in the brain and nervous system:

> If we had a more accurate knowledge of these things we should see what connections were established by the training, and then we should be able to say when we looked into his brain: 'Now he has *read* this word, now the reading connection has been set up.'—And it presumably *must* be like that—for otherwise how could we be so sure that there was such a connection? (1967, p. 158)

In response, Wittgenstein tries to show his interlocutor that he has taken a methodological conviction for an *a priori* truth:

> That it is so is presumably *a priori*—or is it only probable? And how probable is it? Now, ask yourself: what do you *know* about these things? —But if it is *a priori*, that means it is a form of account which is very convincing to us. (1967, p. 158)

The interlocutor is making an assumption and presenting it as a result. In discussing the closely related idea that individual experiences are stored as 'traces' in the brain, Wittgenstein wrote that this "might be held to be artificial or far-fetched; but it is important that it is *possible*" (1980a, p. 157; cf. 1969, p. 7 ff.). But he followed this recognition with a warning of the dangers of speaking this way:

> If there is such a thing as a temptation to regard the differential calculus as a calculus with infinitely small magnitudes, it's conceivable that in another case there may be an analogous temptation, a still more powerful one—when, that is, it gets nourishment on every side from the forms of language; and one can imagine it becoming irresistible. (1980a, p. 158)

What Wittgenstein is suggesting is that the language of 'states of mind' and 'traces in the brain' makes it only too natural for us to assume that there must be something stored in the brain which corresponds to our past experiences. But although many people have been attracted to such theories, we don't actually have any such account of what goes on in the brain when one has a given thought, let alone how that

pattern of activity might be stored and retrieved. Compare this with the account we do have of the relationship between what goes on inside a computer and what one sees on the screen.

We need to be much more suspicious about the motivations which make such theories so attractive. Why did Penfield's optimistic claims that he had solid evidence for the theory that every moment of experience is recorded in the brain receive so much credulous publicity and why are they still taken for granted by many? Why have scientists put so much effort into searching for memory molecules or 'grandmother neurons' (the neuron which remembers Grandma)? Experiments of this genre can only produce results if one makes a large number of unjustified assumptions. For instance, we are told that scientists have identified cells in a monkey's visual cortex which are supposedly detectors for corners or angles. But such cells are only identified in experiments which require an anaesthetised and immobile animal in an artificially simplified environment; as soon as it can move freely in an ordinary environment, it is no longer possible to distinguish cell function in this way [3] (Hubel & Wiesel, 1977; Livingstone & Hubel, 1988).

Fodor's response to such criticism is that cognitivism depends on empirical assumptions, not *a priori* truths. What, then, are the alternatives to the cognitivist's assumtions? In order to answer this question, I will return to the passage which puzzles McGinn and argue that it does contain the kernel of a direct response to cognitivism: an alternative conception of the relation between thought and the brain.

While McGinn only examines pp. 903–905, Wittgenstein's discussion is not limited to those remarks. In p. 908 he asks us to imagine that:

> I want someone to take note of a text that I recite to him, so that he can repeat it to me later, I have to give him paper and pencil, while I am speaking he makes lines, marks, on the paper; if he has to reproduce the text later he follows those marks with his eyes and recites the text. But I assume that what he has jotted down is not *writing*, is not connected by rules with the words of the text; yet without these jottings he is unable to reproduce the text; and if anything in it is altered, if part of it is destroyed, he gets stuck in his 'reading' or recites the text uncertainly or carelessly, or cannot find the words at all—This *can* be imagined!—What I called jottings would not be a *rendering* of the text, not a translation, so to speak, in another symbolism. The text would not be *stored up* [*niedergelegt*] in the jottings. And why should it be stored up in our nervous system? (1980a, p. 908)

The conclusion is even more striking in the German: "stored up" translates "*niedergelegt*", which also means to lay down, or to write down. The "jottings", as Wittgenstein calls them, are radically different from anything we would normally think of as writing. Although, like writing, one could not remember what was said without them—if there were no jottings there would be no recall—they cannot be transformed into a written text by means of rules of translation. While it is true that the jottings are a necessary condition for the jotter's recalling what was said, we must beware of the danger of reifying what was said, of assuming that there must be

a written record of the information in question. One possibility is that the jottings are, like a knotted handkerchief, something that jogs the memory. Malcolm interprets the passage this way and gives the following illustration:

> Often when listening to a lecture, I have jotted down a single word here and there, or drawn lines between two such words, or put down question marks, exclamation marks, and so on. Anyone else who looked at those jottings would find them completely unintelligible. But . . . they enable me to remember the gist of the lecture. (1986, p. 196)

On Malcolm's reading, nothing is hidden, and there is nothing more to be said. Malcolm takes Wittgenstein's arguments to rule out any scientific theory of the relationship between mind and brain (1986, p. 193 ff.).

Another possibility, the one I advocate, is to regard the jottings and the memory they prompt as analogous to the initial state of a connectionist network and the output it leads to. While the memory has a linguistic form, the jottings do not; they are not strings of symbols with an articulated structure, and do not store information. On this account, the jottings and the memories, like the physical and intentional levels of explanation, are only connected by the fact that we have a use for both of them; otherwise, they remain entirely incommensurable.

This interpretation can also be applied to Wittgenstein's other parable, the story of the seed where nothing in the seed corresponds to the plant that comes from it. For DNA, despite the fact that is necessary for the transmission of genetic information and is often spoken of as a 'genetic code', is no more a type of writing than the jottings of the previous example. DNA only has the significance that it does within the context of the appropriate cell. Without a transcription system and a cellular environment, the DNA would be a lifeless string of nucleotides. Thus dinosaur DNA, if we came across it, would tell us nothing about the appearance and character of a dinosaur. Obviously the DNA is necessary for any living creature to reproduce, but we should not assume that its role in storing genetic information is a matter of storing a sequence of discrete items of information.

While only p. 908 seems to express a commitment to causal supervenience, none of the other sections which I have examined are actually incompatible with that thesis. Certainly, they can be read as McGinn suggests, but if one looks more carefully, one will see that Wittgenstein's repudiation of psycho-physical parallelism is directed at the idea that there is a *correspondence* between mental and physical processes. Now that we can see how there might be a looser way of preserving the connection between the two, we can see that rejecting the idea that there must be a correspondence between mental propositions and physical formulae need not commit us to the dualism McGinn rejects, on which "some physical events have no physical explanation" (1984, p. 113, quoted at greater length on p. 210). Thus, Wittgenstein does not say in p. 903 that there is nothing whatsoever in the brain connected with thinking, only that there is no brain process which is *correlated* with thinking. Similarly, in p. 904 it is a *correspondence* between the physiological and the psychological which is denied.

Neither does Wittgenstein say that seeds might grow into certain types of

plants without *any* cause, only that this might occur "as it were causelessly". In p. 905 and p. 908 he rejects the idea that there is a trace 'stored up' or 'written down'; he does not assert that the brain has no causal role in thought. The translation of p. 918 is misleading in this respect, for it has Wittgenstein ask "Why don't we just leave explaining alone?" Actually, "explaining" translates a pronoun which refers back to the phrase "physiological explaining". Even so, this is close to a flat repudiation of the ontological monist's goal of a causal, scientific account of the relationship between mental processes and brain physiology. And elsewhere, Wittgenstein does clearly reject the use of psychological and physiological explanations in this context:

> *Not* to explain, but to *accept* the psychological phenomenon—that is what is difficult. (1980a, p. 509)
>
> Let us represent seeing to ourselves as something enigmatic!—without introducing any kind of physiological explanation. (1980a, p. 963; see also p. 918, p. 1012, p. 1024)

Certainly, Wittgenstein thought that the myth of mental processes distorts our self-understanding, and wanted to describe simply what goes on without resorting to any kind of theory.

> Thinking is an enigmatic process, and we are a long way off from fully understanding it. And now one starts experimenting. Evidently without realizing *what* it is that makes thinking enigmatic to us.
>
> The experimental method does *something;* its failure to solve the problem is blamed on its still being in its beginnings. It is as if one were to try and determine what matter and spirit are by chemical experiments. (1980a, p. 1093; see also 1967, p. 89 ff., p. 305 ff.)

I find it hard to say whether Wittgenstein was only opposed to the misuse of physiological and psychological explanations in philosophical psychology, or whether, like Malcolm, he dismissed work on such matters as the physiological basis of memory as misguided. For instance, Wittgenstein's discussion of noticing an aspect in the *Philosophical Investigations* makes it clear that he considers that any physiological explanation of the phenomenon, say an experimental investigation of the physiology of the eye movements involved, would be irrelevant to the problems that concern him. But there is no suggestion that there is anything wrong with all scientific research into the physiology of the eye (see Wittgenstein, 1967, pp. 193, 203, 212).

Nevertheless, some of Wittgenstein's ideas, summed up in the remark that "The brain looks like a writing, inviting us to read it, and yet it isn't a writing" (1982, p. 806) are a striking anticipation of recent scientific developments. Furthermore, much of Wittgenstein's criticism is directed at the effect of certain misleading models and a conception of scientific explanation which requires uniform laws. Thus, the suggestion that "certain psychological phenomena *cannot* be investigated physiologically, because physiologically nothing corresponds to them" (1980a, p. 904) is motivated by the thought that it may be impossible to express the

relationship between physiological phenomena and psychological phenomena by a covering law. I believe we can accept this critique of brain writing while still leaving open the possibility of a better scientific account of the relationship between thought and the brain, one which does not limit itself to the covering law analysis and which does not postulate a language of thought. It is quite possible that scientists will one day find a holistic theory on which nothing in the brain corresponds to our thoughts, yet still explains how the jottings in Wittgenstein's parable record what was said or how a chaotic brain enables us to have the thoughts we do. On the other hand, it is also possible that these new ideas about the brain will not live up to their initial promise. The aim of this paper has been to bring out some of the promising philosophical implications of connectionism and to contrast it with the dualist assumptions made by cognitivists. I have made extensive use of Wittgenstein's writings on mind and brain because he was keenly aware of the difficulty of making a clean break with dualism. As a result, he saw clearly that there is no *a priori* reason why there must be a language of thought, that it might be impossible to translate brain processes into thoughts because there might not be anything in the brain which translation can latch on to. I have argued that even if this is so, it does not require us to give up the quest for a scientific understanding of the brain. But we will have to give up the idea that physical science can also explain our mental lives.

Acknowledgements

I want to thank Christian Burman, Panayot Butchvarov, Hubert Dreyfus, John Heil, Lars Hertzberg, Curt Huber, Nancy Mullenax, Peter Pfeiffer, Jay Rosenberg, Roger Shiner and Hans Sluga for their comments on previous drafts of this paper. Earlier versions of this paper were presented at the North Western Conference on Philosophy, Tacoma, Washington; Abo Academy, Turku, Finland; the Canadian Philosophical Association, Windsor, Ontario; Bristol University and the Central Division of the American Philosophical Association, Chicago.

Notes

[1] While the original passage is about knowledge and deception, its moral is true of memory, too:

> Consciousness of lying is a *capacity*. It is no contradiction of this that there are characteristic feelings of lying.
>
> For knowledge is not *translated* into words when it is expressed. The words are not a translation of something else that was there before.

Wittgenstein, 1980a, pp. 735–6; Wittgenstein, 1981, pp. 190–1.

[2] For a representative selection of approaches to connectionism see McClelland, J. L., Rumelhart, D. E. & the PDP Research Group (1986); Rumelhart, McClelland & the PDP Research Group (1986); Horgan & Tienson (1987); Graubard (1988); Pinker & Mehler (1988); Smolensky (1988); Churchland (1990).

[3] The best response to those who think that such claims are firmly based on scientific evidence is still Duhem's classic treatment of the relationship between theory and observation (1977, ch. six). For a brief discussion of some examples of problems with particular experiments intended to localize specific aspects of mental functioning, see Malcolm (1977, p 260 ff.).

References

CHURCHLAND, P.M. (1990) Cognitive activity in artificial neural networks, in: D.N. OSHERSON & E.E. SMITH, *An Invitation to Cognitive Science, Volume 3: Thinking*, pp. 199–227 (Cambridge, MA, MIT Press [Bradford Books]).

COOK, J.W. (1969) Human beings, in: P. WINCH (Ed.) *Studies in the Philosophy of Wittgenstein*, pp. 117–151 (London, Routledge & Kegan Paul).

DESCARTES, R. (1986) *Meditations on First Philosophy*, translated from the Latin edition of 1641 by R. Rubin. Reprinted in J. PERRY & M. BRATMAN, *Introduction to Philosophy* (New York, Oxford University Press).

DREYFUS, H.L. & DREYFUS, S.E. (1988a) On the proper treatment of Smolensky, a response to Smolensky (1988), *Behavioural and Brain Sciences*, 11, pp. 31–32.

DREYFUS, H.L. & DREYFUS, S.E. (1988b) Making a mind versus modeling the brain: artificial intelligence back at a branch point, *Daedalus*, 117, pp. 15–43. [Reprinted in: GRAUBARD (1988).]

DUHEM, P. (1977) *The Aim and Structure of Physical Theory*, translated by P. P. Wiener (New York, Atheneum).

EDELMAN, G. (1985) Neural Darwinism: population thinking and higher brain function, in: M. SHAFTO (Ed.) *How We Know* (San Francisco, Harper and Row), pp. 1–30 .

FODOR, J.A. (1975) *The Language of Thought* (Cambridge, MA, Harvard University Press).

FODOR, J.A. (1987) *Psychosemantics* (Cambridge, MA, MIT Press [Bradford Books]).

FODOR, J.A. & PYLYSHYN, Z.W. (1988) Connectionism and cognitive architecture: a critical analysis, *Cognition*, 28, pp. 3–71. [Reprinted in: PINKER & MEHLER (1988).]

GRAUBARD, S. (Ed.) (1988) *The Artificial Intelligence Debate: False Starts, Real Foundations* (Cambridge, MA, MIT Press [Bradford Books]).

HEIL, J. (1981) Does cognitive psychology rest on a mistake? *Mind*, 90, pp. 321–342.

HORGAN, T. & TIENSON, J. (Eds) (1987) Connectionism and the philosophy of mind, *The Southern Journal of Philosophy*, 26, supplementary volume.

HUBEL, D.H. & WIESEL, T.N. (1977) Ferrier Lecture: Functional architecture of macaque visual cortex, *Proc. Royal Soc. London, B*, 198, pp. 1–59.

LIVINGSTONE, M.S. & HUBEL, D.H. (1988) Segregation of form, color, movement and depth: anatomy, physiology and perception, *Science*, 240, pp. 740–749.

MALCOLM, N. (1971) The myth of cognitive processes and structures, in: T. MISCHEL, (Ed.) *Cognitive Development and Epistemology*, pp. 385–392 (New York, Academic Press).

MALCOLM, N. (1977) *Memory and Mind* (Ithaca, NY, Cornell University Press).

MALCOLM, N. (1978) Thinking, in: E. LEINFELLNER, W. LEINFELLNER, H. BERGHEL & A. HÜBNER (Eds), *Wittgenstein and his Impact on Contemporary Thought* (Vienna, Hölder-Pichler-Tempsley), pp. 411–419.

MALCOLM, N. (1986) *Nothing is Hidden* (Oxford, Blackwell).

MCCLELLAND, J.L., RUMELHART, D.E. & THE PDP RESEARCH GROUP (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 2: Psychological and Biological Models* (Cambridge, MA, MIT Press [Bradford Books]).

MCGINN, C. (1984) *Wittgenstein on Meaning* (Oxford, Blackwell).

NEISSER, U. (1982) Snapshots or benchmarks?, in: U. NEISSER (Ed.) (1982) *Memory Observed: Remembering in Natural Contexts*, pp. 43–48 (San Francisco, W. H. Freeman).

NEISSER, U. & WINOGRAD, E. (Eds) (1988) *Remembering Reconsidered* (Cambridge, Cambridge University Press).

PINKER, S. & MEHLER, J. (Eds) (1988) *Connections and Symbols* (Cambridge, MA, MIT Press [Bradford Books]).

PYLYSHYN, Z.W. (1984) *Computation and Cognition: Toward a Foundation for Cognitive Science* (Cambridge, MA, MIT Press [Bradford Books]).

ROSENFELD, I. (1986, 9 October) Neural Darwinism: a new approach to memory and perception, *New York Review of Books*, pp. 21–27.

RUMELHART, D.E., MCCLELLAND, J.L. & THE PDP RESEARCH GROUP (1986) *Parallel Distributed*

Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations (Cambridge, MA, MIT Press [Bradford Books]).

RUMELHART, D.E. & NORMAN, D.A. (1981) A comparison of models, in: G. E. HINTON & J. A. ANDERSON (Eds) *Parallel Models of Associative Memory* (Hillsdale, NJ, Lawrence Erlbaum Associates).

SEARLE, J. (1980) Minds, brains and programs, position paper with peer commentary and author's response, *Behavioural and Brain Sciences*, pp. 417–457.

SMOLENSKY, P. (1988) On the proper treatment of connectionism, position paper with peer commentary and author's response, *Behavioural and Brain Sciences*, 11, pp. 1–74.

STOUTLAND, F. (1988) On not being a behaviourist, in: L. HERTZBERG & J. PIETARINEN (Eds) *Perspectives on Human Conduct*, pp. 48–60 (Leiden, E. J. Brill).

VON WRIGHT, G.H. (1986) The Wittgenstein papers, in: V.A. SHANKER & S.G. SHANKER (Eds) *Ludwig Wittgenstein: Critical Assessments, Volume 5: A Wittgenstein Bibliography*, pp. 1–21. Revised and updated version of paper first published in *Philosopical Review*, 78 (1969), pp. 483–503.

WITTGENSTEIN, L. (1935–6) Notes for the 'Philosophical Lecture', MS 166, *unpublished manuscript:* for further information, see VON WRIGHT (1986).

WITTGENSTEIN, L. (1967) *Philosophical Investigations*, G.E.M. ANSCOMBE & R. RHEES (Eds) translated by G. E. M. Anscombe (Oxford, Blackwell).

WITTGENSTEIN, L. (1969) *Preliminary Studies for the 'Philosophical Investigations' generally known as 'The Blue and Brown Books'* (Oxford, Blackwell).

WITTGENSTEIN, L. (1980a) *Remarks on the Philosophy of Psychology, Volume I*, G.E.M. ANSCOMBE & G.H. VON WRIGHT (Eds) translated by G. E. M. Anscombe (Chicago, University of Chicago Press).

WITTGENSTEIN, L. (1980b) *Remarks on the Philosophy of Psychology, Volume II*, G.H. VON WRIGHT & H. NYMAN (Eds) translated by C. G. Luckhardt & M. A. E. Aue (Chicago, University of Chicago Press).

WITTGENSTEIN, L. (1981) *Zettel*, G.E.M. ANSCOMBE & G.H. VON WRIGHT (Eds) translated by G. E. M. Anscombe (Oxford, Blackwell).

WITTGENSTEIN, L. (1982) *Last Writings on the Philosophy of Psychology, Volume I; Preliminary Studies for Part II of the 'Philosophical Investigations'*, G. H. VON WRIGHT & H. NYMAN (Eds) translated by C. G. Luckhardt & M. A. E. Aue (Chicago, University of Chicago Press).

YOUNG, J.Z., AYALA, F.J., SZENTAGOTHI, J., STELLAR, E. & ROSENFELD, I. (1987) Neural Darwinism: an exchange, *New York Review of Books*, 12 March 1987, pp. 44–45.