THEORIA ET HISTORIA SCIENTIARUM, VOL. VI, N° 2
Ed. Nicolas Copernicus University 2002

Rom Harre

The Memory Machine

Abstract

The foundations of an adequate cognitive science that binds the cognitive activities of human beings into a coherent conceptual system with the neurological basis of these activities has been slow to develop. The problem is partly due to the complexity of the relationships that must be set up between a naturalistic analysis of the discursive practices such as remembering, and the brain mechanisms by which they are accomplished. There are plenty of intermediate models of cognitive 'mechanisms' but they have been developed for the most part with little attention to the ontological constraints on model building that are central to the physical sciences. The transition from human activity to neurological hypothesis can be accomplished by a second step of modeling in which standard cognitive models are revised by the use of connectionist architectures, to provide a foundation for plausible neurological hypotheses. The argument is set out in the context of the psychology of remembering.

1. Introduction

It is well to bear in mind that remembering as an every day practice is a social activity. Quite often a successful attempt to recover the past requires a dialogue in which more than one person is involved. Just what 'successful' means in the context of remembering is a difficult problem. In most cases in real life, there are no records or traces of what occurred, against which to test the verisimilitude of what we claim to have happened. Not only that, but we have to learn what it is to remember something, as opposed, for example, to imagining it.

The lexicon of words for referring to this activity are nearly all 're', that is 'again' words. We have re-member, re-call, re-collect, re-minisce, re-live, re-cover and more. To remember somethings, in all sorts of possible ways, to experience it

again. But it is in the nature of time that we cannot experience *it* again. All *acts* of remembering are, at least in principle, attempts to recover some aspect of the past by conceiving a representation or description of it. Such acts differ by the degree to which the opinions of other people influence our accepting of this or that thought or utterance or drawing as a reasonably correct representation of the past. We also use the word 'remember' for the permanent possession of an ability or skill, even though we do not remember the occasion on which we learned it.

The main uses of the word 'know' fit nicely with these two major uses of 'remember'. We show that we 'know that' something is the case or happened, and so on, by recalling it. We could equally well have said that we 'remembered that' such and such is the case, happened and so on. In exercising a skill we show that we 'know how to do something' or we could say that we 'remember how to do something'.

At first sight it would seem as if identifying instances of these kinds of remembering and noting what was happening in the brain and nervous system when someone performed one of them would be enough to allow us to pick out the neural toolkit with which we carry out everyday remembering tasks. There are many other structures and processes that occur in the human brain and nervous system, but only certain regions are active when someone is exercising their powers of recall. The tools for remembering must be among those we are endowed with by Darwinian selection.

However, the seemingly simple step, from an analysis of remembering practices to a cognitive science of remembering involving the mechanisms by which remembering is achieved, is much more difficult than one might suppose at first sight. In this chapter we explore some important suggestions as to the ways that it might be accomplished, and assess the progress that has been made so far in this work. We will also identify certain blind alleys that once looked promising but have turned out to be dead ends.

We will discover that two major technical specialities are required to achieve our goals. There is the neurophysiology and neuroanatomy of those organs in the brain that we use for remembering. However, to be able to understand how they work, we need to draw on knowledge engineering, in particular the AI techniques of connectionist modeling. We will see how by bringing these two technical specialities together the transition from common understanding through cognitive psychology to models of remembering machines and finally to an understanding of certain neural systems can be achieved. We must also include studies of various prostheses, like agendas, electronic organizers, and knots in handkerchiefs, that serve as ancillary devices for performing some of the same range of tasks. Not only are they of interest in their own right, but it is also possible, indeed likely, that how they work can throw light on how natural, organic memory machines work.

Learning and remembering are a complementary pair of concepts. In learning we acquire certain skills and abilities, certain bodies of knowledge, that are relatively permanent. In remembering we make use of those skills and find public and sometimes private expression for what we have learned. There seems to be a very strong tendency to use the vocabulary which has its home in the public world of exercises of skills and displays of knowledge for the states of the human person that are the basis of the permanent possession of skills and bodies of knowledge. This is one of temptations that we must learn to resist as we build our psychology of remembering.

Glenberg is quoted in Gamham (1997) as answering the question 'What is memory for?" with the empty cliche 'in the service of perception and action in a three-dimensional environment'. It is much better to ask' What do people use their biologically inbuilt or shop-bought memory-machines for?'. This question has indefinitely diverse answers: 'For keeping appointments, for finding the way home, for remembering one's mother's birthday, for ordering lunch in France, for taking part in a quiz show and so on and so on, including for taking part in a psychological experiment'. Having identified the various tasks comprehended under the umbrella concept of 'remembering', including its metaphorical extensions in knowledge engineering contexts, the next step might seem to be to simply pick out the neural tools that are being used in carrying out these tasks. Unfortunately matters are not so simple.

Between the phenomenon and the neural tools we must insert an AI model by means of which the diverse structures of task and tool are brought into coherence, made to match. Attempts to go from the way acts of remembering are performed to the way neural tools work have not been successful except in so far as they have involved the construction of informal AI models. We will find very instructive examples of these in the current cognitive psychology of 'memory'. Such an intermediary step is necessitated by the simple fact that remembering is a *normative* discursive practice, while neural activity is causal and material. To remember is to recollect correctly. But how do we know that that has been accomplished? The present recollection must be a good representation of the past state of affairs. However, the past has no existence in the present. Most real happenings leave no trace. This fact is often overlooked in Ebbinghaus-type experimentation[1] in that the question of the authenticity of the stimulus materials is never raised, and indeed is just taken for granted. However, in real remembering it is usually problematic and can always be challenged.

[1] Though Ebbinghaus used single participants, usually himself in his experiments, he pioneered a type of experiment that is still performed. Meaningless signs are studied and their rate and time of recall are treated as independent variables. Correlations are established between aspects of the material and process of learning and the material as remembered.

This point is very important for the general theory of remembering. Just to take an experiment at random: here we have a description of an experiment of Postman and Phillips (1965).

> ... subjects are presented with a list of unrelated words and asked to recall as many as possible in any order they wish. ...when recall is immediate, there is a tendency for the last few items to be very well recalled, the so-called recency effect. After a brief filled delay, however, the recency effect disappears' (as reported in Baddeley, 1998: 38b).

Notice that it is taken for granted that the material has survived unchanged from the moment at which it was presented to the participants to the moment at which they recall 'it'. Try doing that with your dinner of two weeks ago! Or almost anything else that one might be required to remember. In real life we almost always assess the accuracy of recall by comparing the content of one act of remembering with another. Reliable material records are very rare. It is also worth noticing that the remembering powers of the experimenter are actually given privileged status in the situation, since no one casts doubt on the authenticity of experimental material as presented by the experimenter. He/she has memorial power.

A cognitive science in full

There are five stages to a complete treatment of a phenomenon like remembering. In the first stage we conduct an analysis of the practice as it is carried out in everyday life, including studies of analogous phenomena in other cultures than our own. With this material in hand (often unwisely simply assumed or taken for granted by psychologists) we can begin to identify different kinds of remembering, some by content, some by function and some by other criteria that emerge in the effort to construct a taxonomy. At this point one invokes the overall Task/Tool metaphor, looking now towards the tools by which people conduct memorial tasks. The tools that have interested most psychologists are 'natural', the organs of thought, located in the brain and nervous system. Models of these organs are constructed loosely indebted to concepts drawn from knowledge engineering. Before the step to neural investigations can be made these models should be refined and developed in detail. Only then can there be a confluence between neural studies and the study of cognitive phenomena in question. Barnes and Hampson (1997: 496) put the matter very well: 'Fortunately, the development of connectionist science has provided the behavior-analytic community with an opportunity to forge those all-important links with those involved in the study of neurophysiology [and anatomy]'.

After constructing an AI simulation of the information processing structure of various varieties of remembering the next task is to provide a model for neurophysiological testing. Such a model must fit the known features of the discursive performances involved in remembering on the one side, so to say, and the possibilities of neural functioning on the other.

It turns out there are two very different ways of implementing the AI transition. It really turns on just how the AI transition is to be made.

> Marr's ideas [on perception] are made more plausible by the way that Marr explained the short-comings and failings of approaches that focused on the detailed computational mechanisms(artificial intelligence [old style]) or neural mechanisms (neurophysiology) without considering the more abstract computational theory (Garnham, 1997).

The way Garnham and others have suggested we proceed is through the idea of mental models. This is such an important theoretical advance, whether or not it proves viable in the long run, that I must explain it with some care. It runs strongly counter to the dominant 'storing' and 'coding' metaphor. In that scheme, items of information are expressed in some coded form and stored as such. But in mental modeling there is no store, but rather a representation (literally) that is a model of the world or some aspect of the world, to which every source contributes, enriching, revising and correcting it. It makes no difference whether the material to be incorporated in the model comes from what is said, or what is perceived, or what is acted out. To remember is to do something like perceiving, that is scanning a landscape for what might be important. Here we have a use of the term 'model' (Miller and Johnson-Laird, 1976) that is more or less the same as the use made of the expression by physicists and chemists. However, though Garnham and others still use the old vocabulary of 'representation' and 'information' the meanings of these expressions is not metaphorical. A mental model is a genuine representation of some aspect of the world, and 'information' is what is retrieved from the model, and presented in some public form, such as statement or a map or a set of instructions. 'Information' is not what the model consists of. In some cases, as we will see, the model is made of neural tissue.

Neisser's Paradox

Modeling, the heart of cognitive science as it is of any scientific enterprise, is constrained by two external relations, by what we know about public conduct and procedures and skilled performances on the one hand, and by neurological possibilities on the other. This point was first brought to prominence by Ulrich Neisser (1978: 2). He presented a paradox: 'if X is an important or interesting

feature of human behavior, then X has rarely been studied by psychologists'. The led Neisser to the concept of 'ecological validity', that is that the results of laboratory research should be generalisable to the patterns of ordinary life. We have already seen that laboratory work in the study of remembering, using Ebbinghausian methods, is hopeless compromised from the start, *if it purports to be an investigation of people remembering things.* But conceived within the project of building AI models of the tools people use for carrying out tasks of remembering, it might appear as a quite different undertaking, and of great importance and interest.

The solution is neither to reject Ebbinghausian experiments out of hand, and demand ecological validity, nor to insist on bringing all psychological phenomena into the 'laboratory'. The upshot would be triviality on the one hand and an intolerable complexity on the other. The solution is to find a conceptual system that will comprehend both methods of enquiry, each assigned its proper role. The Task/Tool way of looking at cognitive psychology provides just such a comprehensive conceptual system.

The ties between what is revealed to a naturalistic analysis and conceptions of tools constructed by mental modeling are strong but non-reductive. Neural mechanisms are the 'natural' tools for certain cognitive tasks. In addition a full psychology of remembering must include various non-natural devices such as the prosthetic devices from knowledge engineering, such as electronic organizers.

There is another dubious assumption into which cognitive psychology of remembering sometimes seems to slip. We start with the correct principle that whatever characterises the diversity of type of material to be remembered, be it content, form or sensory modality, must also characterise what is recalled. It does not follow that whatever characterises what is remembered and what is recalled must also characterise 'memories' as they are maintained in the remembering system. Once again the idea of modeling enables us to avoid slipping in to this assumption.

2. The Cognitive Psychology of Remembering

2.1. Introduction

In discussing the current scene I will generally prefer the term 'remembering' with its connotations of process and activity over 'memory' which its substantival implications. Most of the authors to be discussed tend to use the noun, even though they usually mean to refer a cognitive *process.*

In all the sciences knowledge is presented in a complicated pattern of metaphors and analogies. This has the enormous advantage of allowing creative

thinking to be achieved, while it has the disadvantage that often times aspects of the source of metaphor or analogy creep into the application of the metaphorical, usage with unfortunate results. Physicist are more alert to this problem than psychologists. One of the skills we must acquire in developing our understanding of cognitive psychology is a sensitivity to the limits of the working metaphors, which are both necessary to the scientific standing of the field and remain problematic. They are both growing points and dangers.

The metaphors and analogies that do the work in cognitive psychology of remembering are, as I shall show, the beginnings of the building of an AI model or models for different aspects of remembering. We could call this an informal AI treatment or perhaps better, the setting up of a proto-AI. We need to track down the metaphors and to sharpen the AI hypotheses which they make possible.

We begin our survey of the cognitive psychology of remembering with some recent proposals for schemes for classifying of types of remembering. It will quickly become apparent that there is a wide variety of considerations and criteria in use in distinguishing kinds of remembering, each of which has value in particular contexts. Many acts of remembering could be classified under more than one heading.

2.2. Some Types of Remembering

Introduction

It would seem to be obvious that the ways we remember visual, auditory and tactile experiences must be different from the ways that we remember meanings, stories, recipes and so on, and different again from the way we remember the way home, knowing which way to turn at each junction when we reach it, though we could neither visualise the route nor give adequate instructions to a visitor beforehand. While there is something in these commonsense distinctions one hundred years of studies of how and what people remember has led to a variety of classificatory categories, each having a certain utility in an appropriate context. I believe we can say definitively that some proposals are certainly mistaken and others still in need of clarification, while others look like being here to stay.

A Pervasive Metaphor

Two preliminary observations are in order to understand the significance of the various taxonomies. First of all the entities and structures referred in these classification systems are *abstract* entities and processes. The question of whether

they have real world analogues is left open. This is one of the most important reasons for saying that cognitive psychology of remembering should be treated as a proto-AI. Secondly, there is the pervasive metaphor of remembering as storing, and of memory as a store or stores. Along with that image goes the metaphor of memories as discrete items of knowledge. Each of the three major classifications to be discussed has a different basis. One has to do with the structure of the remembering system, another with the content of what is remembered, and a third with attributes of the remembered items, as they appear in recall. In each the root metaphors are involved but modified in various ways. It will emerge that in using neural nets as the basis for an AI interpretation of some cognitive psychology models, the root metaphors come under some strain.

Short term 'memory'

The distinction between short-term and long-term memory was once fundamental to the psychology of memory. Though cognitive psychologists recently have more or less dropped the distinction in its simple form, the concept of a distinctive short-term memory is the basis of some very important research and consequential model building. It has been important for verbal learning, especially when the learning of verbal material has been divorced from discursive contexts[2]. It is very closely tied to the 'store' metaphor, in that it pictures representations of current happenings as *stored* in short term memory. Some are transferred elsewhere, but it is now thought to be a mistake to imagine that there is a long-tern memory machine, which functions like the short-term one. The high point of this 'short/long' metaphor was many decades ago. One sees it pervading the work reported in Atkinson and Shiffrin (1968) for example. It is implicit in George Miller's famous formula for the limits of short term remembering: that the item should consist of seven plus or minus two elements, whatever they were.

Along with the idea of remembering as storage goes another important principle, that is also clearly a step toward a certain kind of proto-AI interpretation. Anderson and Shiffrin proposed that each perceptual mode, sight, hearing, touch and so on had a independent memory 'store'. In so far as this model has come under criticism, notably by Engelkamp and Zimmer (1996), that route to an AI interpretation has been seen to be a dead end.

Included in this 'package' is another important idea: that all remembering, what ever its origin, that is whatever perceptual mode it is acquired in, and whatever

[2] There are some reservations needed in taking 'Ebbinghausian' methodologies too seriously, especially in linking laboratory studies with the problems of remembering and forgetting in real settings.

its content when recalled, it is stored in the same form. Usually this was presumed to be propositional, and that went along with the idea that what was stored was knowledge. After all, what was recollected was putative knowledge even if the recollection was incorrect. Once stored all subsequent cognitive processing makes use of the same abstract devices, such as hierarchical classification.

Though the concept of 'short term remembering' has survived into the present era of memory research the way this notion is interpreted and what is involved in the processes which the phrase comprehends have been greatly enlarged in detail (Gathercole, 1997). Most of the research has been concerned with linguistic or more generally symbolic material. The main development of the concept has been through the introduction of the idea of 'working memory' (Baddeley, 1998)

In furthering our project of looking for AI metaphors in proto-AI treatments of cognitive functioning, Baddeley's working memory theory offers an ideal example. First proposed twenty five years ago, though it continues to be refined in various ways, it remains a paradigm-defining idea.. The theory describes a hypothetical mechanism that consists of the postulation of a model, in the physical science sense, that is an imaginary mechanism which would perform similarly to whatever it is in the real human being that is used to perform a memory task. For example the model must produce an analogue of the phenomenon that shorter words are recalled more readily than longer words, and that word-like entities are recalled more readily than pseudo-signs which are not word- like.

In coming to understand Baddeley's model and to appreciate its significance for cognitive psychology it is of the greatest importance that its logical character should be clearly appreciated. To that end we will pause to study it quite closely.

There are three modules in the model, a central executive, a phonological loop and a visual-spatial sketch pad.. The loop is defined by Baddeley as follows:

> [it] is assumed to comprise two components, a phonological store that is capable of holding speech-based information, and an articulatory control process based on inner speech. Memory traces within the phonological store are assumed to fade and become unretrievable after about one- and-a-half to two seconds. The memory trace can however be refreshed by a process of reading off the trace into the articulatory control process, which then feeds it back into the store, the process underlying subvocal rehearsal. (Baddeley, 1998: 53-3).

Already in this schematic description we see a cluster of diverse metaphors, one for each component of the model. It is also worth remarking that this model is very much a child of the time of its conception during the dominance of system theory in control system engineering in the nineteen seventies. The model is fleshed out with two familiar metaphors, the 'store' and the 'rehearsal'. Since these are metaphors the qualification that for verbal remembering the rehearsal is sub-vocal is not strictly speaking necessary since these components

are abstract entities, not actual material things. This very soon turns out to be a model that is not subject to the constraint that it should yield a description of a real human memory tool system, for example some anatomical structure or structures in the brain with their neurochemical processes. Finally the processes that are suggested for the workings of the model involve yet another familiar metaphor, 'coding', which is part of a cluster of metaphors around the core trope of 'representation'.

At this point it will be illuminating to remind ourselves of the procedures for building and assessment of models in the creation of theories in physics. We can then look at the similarities and differences between model-making in each context, physics and chemistry ori the one hand, and cognitive psychology on the other. In both contexts models are constructed relative to an existing body of data which they must, in some way, account for. In physics it is not presumed that the conceptual system which is used to interpret phenomena as *data* is independent of the theory in process of creation. For example Newtonian mechanics provides the laws for molecular motion as kinetic theory, and the concepts for interpreting observations as data in experimenting on gases. The assumption that data are created by one set of interpretative concepts and procedures and models and theories by another is pervasive in psychology. Baddeley expresses the matter thus:

> attempting to constrain possible models by using a rich and robust pattern of results, any one of which is capable of being explained in several different ways, but which together place major constraints on possible explanations. (Baddeley, 1998: 52b).

Of course logic tells us that no matter how rich and robust the body of data it cannot constrain the possible models in the slightest degree. Even distinguished scientists slip into that error. But the important point is that Baddeley assumes that data could be collected and interpreted without having some model of the source of the data in mind.

However, the greatest difference between physics and psychology is that physicists generally presume that the models they are trying to build must be representations of possible real world entities, structures and processes. Psychologists have not always seemed to have felt constrained in this way. There is no suggestion that Baddeley's phonological loop must be a representation of something loop-like in the real world. We can see this in the general assumption that models are tested hypothetico-deductively, the core assumption in positivism, that is not in terms of their plausibility as representations of real world entities, but by the extent to which deductions from them can be matched to data. How are we going to achieve an advance from Baddeley's positivism to realist scientific theorizing in cognitive psychology? As we shall see recent proposals for AI modeling address this very point.

It would be a great mistake to take this disparity between physics and psychology as a criticism of psychological model-making. Fortunately psychology is a multilayered structure. Baddeley's phonological loop and his visual scratch pad are abstract entities tied conceptually to the phenomena they 'account for'. The development of AI, and particularly the computational or neural net version of it, has provided us with a third layer sandwiched between the abstract and ideal entities of most cognitive models and the real structures of body and brain. The making of abstract cognitive models, drawing on all sorts of metaphors, is an *absolutely essential* step in developing cognitive psychology. The point of this discussion is to show how it is to be further developed along the lines we are familiar with in the natural sciences.

To transform Baddeley's model of short-term remembering from an abstract model to an AI simulation, two steps would be needed. The components of the hypothetical mechanisms must be able to be interpreted as processing modules, and the processes that are imagined to take place within them must be interpreted as computations on the binary input to a Turing machine, or, in more recent AI., as the input/output pattern of a trained neural net.

In this section we are not concerned with the final step, from abstract representation of the cognitive tool as an AI simulation to its physical realisation in the structure and processes of a brain and CNS. It should be clear that the 'working memory model' falls short of a representation of anything that could be found in the neurology of a person. Yet, without it, the next step would be extraordinarily difficult.

Why has the concept of Tong term memory' seemed to have dropped out? This is partly due to the realisation that there may be several memory systems, of different types and working in different ways. Remembering for a long time may not be the result of the transfer of material form one 'store' to another essentially similar 'place'.

Procedural and Episodic remembering

This distinction as used by cognitive psychologists, for example Baddeley (1998: 149b), is very close to the distinction proposed by analytical philosophers between 'knowing how' and 'knowing that'. The analytical distinction was drawn from analysis of the ways that remembering concepts were used in everyday life, and it plays an important part of in the discursive psychology of remembering. The distinction as used by cognitive psychologists then has a very interesting character. It has a foot in discursive, naturalistic psychology and a foot in proto- AI. It is based on a fundamentally different metaphor from the short and long term distinction since it is a classification by *content.* The distinction in content

makes itself felt in a distinction between what is recollected, namely a procedure in the one case, for example a performance skill, and a putative matter of fact on the other.

Here is how Engelkamp and Zimmer (1994: 1) define episodic remembering:

> ...memory for objects or events we have seen, ...speech that we have heard or texts that we have read ...actions that we have performed.

The terminology becomes a little complicated because the term 'declarative memory' can be used in contrast to procedural memory, taking in episodic and semantic memory. However, it remains true that episodic and semantic memory are distinct in one important respect. Generally the context is preserved as either an explicit or implicit part of what is remembered of an autobiographical episode, whereas in learning and remembering meanings, the episodic context is not, in general, remembered. Procedural memory then comprehends skill and habit, when what is remembered is a procedure, rather than something that would described propositionally. The distinction is not entirely satisfactory in that one could make a case for treating semantic remembering as a skill, or even a habit.

At this point we need to attend to another distinction, that between single mode and multi-modal remembering. The idea is very simple. Is remembered visual material 'stored' in one system, independently of the auditory material that is stored in another, and of the verbal material that is in yet another and so on? Or are these simply sub-systems of a larger integrated memory 'machine', so that there is low-level cross-modal influence in both learning, remembering and forgetting? In a great deal of 'classical' cognitive psychology of remembering there seems to be a 'modal-eliding' principle at work. If we find that items that have a linguistic origin and those that have a perceptual origin are involved in remembering something, how is this achieved? One answer would be say that all remembering has the same form, for example, propositional, so the modality of its source is irrelevant. In the end, it is argued, by Pylyshyn (1973) for example, all information is of the same kind, namely propositional. Here we have some old-fashioned metaphors clashing with one another. Multi-modal theory is based in the idea that each modality has its own memory machine, a part of some larger device. Models multiply!

Among the main advocates of multi-modal remembering are Engelkamp and Zimmer (1994). Their enthusiasm for the multi-modal model is partly the result of some Ebbinghausian type experiments which show that if one is required to act out the content of the phrase one is being asked to remember one remembers it better. According to them, different subsystems contribute to episodic memory. These subsystems are abstract entities, created from sensory, kinesthetic and motor systems. The interaction of the subsystems produces episodic memory.

Engelkamp and Zimmer (1994: 464 - 5) derive their subsystems from a blend of functional and content analyses. As they remark, 'the units we chose were determined by our interests in memory for these entities in the real world'. That is they derived their subsystems as picture nodes, word nodes and motor programs.

Implicit and explicit remembering[3]

It seems to me that this distinction, first proposed explicitly by Schacter (1987), is very closely tied to that between procedural and episodic remembering. The distinction between implicit and explicit remembering is very clear. It is simply that when we observe someone doing something X, we know that some prior condition Y must have been satisfied before or together with the doing of X so that doing X is possible. Among the conditions is that some procedure or some item of information should be available for use, though not consciously recalled in the performing of the task. It naturally merits the description 'implicit'. Someone tells me that it is 'Five to twelve'. That person must have implicitly remembered that the long hand on '11' means '55 minutes of the hour have elapsed' and that the short hand on '12' means midday when the sun is shining.

When the implicit/explicit distinction is used in linguistic contexts it is tied with another distinction, that between semantic and lexical memory, that is between recollecting the meaning of a statement or word and recollecting its shape, grammatical form and even the language it was in.[4] Priming is the phenomenon by which some item of information, for example, given to someone earlier, and not explicitly recalled, can be seen to have influenced the perception or understanding of some item presented later. This is one of the contexts in which the distinction between implicit and explicit remembering is important. In studying the phenomenon of 'priming', it is easy to distinguish experimentally between the two kinds of implicit linguistic memory simply by using bilingual participants. For example in semantic priming the speed of recognition of a word is affected by the content of words presented a little earlier and not explicitly recalled when someone is trying to recognise the newly presented word. Recognition rates are also affected by the form of the words presented earlier. This is lexical priming. For a bilingual the semantic priming effect will be independent of the language in which the priming act is performed. Lexical

[3] To make this field yet more confusing there is another use for the term 'implicit' to describe the case in which someone recalls something without being able to remember learning it.

[4] I frequently work in Spanish speaking countries, often with people who are bilingual, and I am sometimes quite unsure in which language a certain piece of information was given to me, though I can recollect the information accurately in either English or Castillian.

priming wili be sensitive to choice of language. ‘Horse’ and ‘caballo’ are semantically equivalent but lexically different.[5]

Prospective and retrospective remembering

This distinction is clear in import, but its implementation in detailed cognitive models strikes me as rather tentative. Retrospective remembering is any exercise of the ‘memory machine’ to recall something which one has already done, intentionally. Prospective remembering is simply to remember to carry out some task according to an already formed intention. Anyone familiar with the literature on intentions will be aware that it is one of the most slippery and ambiguous concepts in the whole of cognitive psychology. Empirical studies have concentrated on studies of the relationship between prospective and retrospective remembering. An ability to remember things in the past does not seem to be all that is required to remember to do something one has planned or intended to do. This is a very odd result, since commonsense would suggest that all that would be required would be remembering that one had made up one’s mind, declared an intention and so on. Most of the empirical studies have been carried out with participants whose abilities in one or the other or both of these everyday remembering tasks has been impaired. The reasoning is familiar in cognitive neuroscience. If the disturbance of a cognitive function is correlated with damage to a part of the brain, then it is inferred that the intact part of the brain was the organ by which the function was implemented.

Some of the difficulty that one has in seeing how these different kinds of remembering are related to one another, if they are, arises from the use of one or more metaphors incompatible with each other and with the general trend of contemporary cognitive psychology to persist in metaphors which encapsulate points of view that the AI transition has superceded. Ellis (1996)

2.3. Some important metaphors

Four main metaphors need to be deconstructed in order for us to see clearly how the cognitive models, derived according to the principle by which modules are matched to distinctions of function, that was emphasised above. These are ‘representation’, ‘information’, ‘(en)coding’ and ‘processing’. It is necessary to

[5] Key choice may have a priming effect in musical apprehension, and there may be an analogue with semantic and lexical priming effects. Major and minor keys are generally held to be emotionally distinctive.

remind the reader that deconstructing a metaphor is not necessarily a criticism of the practice of using it. Science would be nothing without the metaphors by means of which theories are constructed, new concepts are built, models are conceived and their structures worked out. Nevertheless they are metaphors, and when we begin the business of transforming an abstract model, such as Baddeley's 'working memory system', into a plausible AI model we must pay close attention not only to the literal meaning of the metaphors through which the creative thinking that brought the new model to light was achieved, but also to the meaning the expressions have acquired in their metaphorical usage. These will not be the same. Yet there may be subtle influences from the original meanings which are misleading and need to be attended to when we use the expression as a well- established metaphor.

Interestingly each of the four has a long history, involving major but superceded theories of how remembering and other cognitive processes occur. 'Representation' implies that there is a kind of simulacrum of the item remembered to be found somewhere in the human person in some form or another. Perhaps it even takes the form of a picture for remembering a landscape, or a proposition for remembering what has been said, and so on. Literally 'representation' is to present whatever it is again. The naive sense of the expression reappeared in computer science in the days before neural nets, when it was thought that the compiler in rendering keyboard input into electrical impulses ordered in accordance with a binary system, created representations in the machine of what had been put in via the keyboard. When the idea of there being a one to one correspondence between input units and states of the computer as a material machine was abandoned, the notion of representation was stretched once again to describe how the whole structure of a neural net 'represented' something, for example a non-Linnaean classification. The term has been so leached of any meaning that at most its use suggests a weak relationship between what is input and what it is the consequential state of the computer, the brain, the nervous system, and so on. Only in mental model theory is something of its root sense restored.

'Information' has a shorter history, but has also become almost vacuously generalised. Originally 'information' meant the content of an 'informative' proposition, the fact that saying or writing something conveyed to someone who knew the language. In this sense a newspaper or a manual of instructions would contain information. When Shannon developed his general theory of transmission lines for such systems as telephones, he called it, non-metaphorically, 'information theory' It was concerned with the constraints on the transmission of information over transmission lines. But the mathematical treatment of the properties of such lines quickly changed the meaning of the expression 'information' into a metaphor. For example the 'information content' of a message 'b' is the logarithm of

the inverse ratio of the probability that the original message was 'a'. The metaphorical use of this term is widespread. Indeed one might want to say that the same vocable, 'information', is being used for two quite different concepts. For example in a fairly standard description of Rolls' account of the architecture of the hippocampus, thought to be the seat of certain aspects of the neurophysiology of the remembering machine, we have such expressions as 'during its course through the brain this information [namely the input from perceptual systems] ... is then communicated to the EC' and so on. Of course in one sense there is no *information* in the brain at all. There are only electrical pulses and synaptic chemistry. One might say that the*people using this metaphor are not misled by it. Only the lay-person unfamiliar with neuroscience would draw misleading conclusions. Students, with a foot in both camps need to be sensitised to both the necessity and the dangers of metaphors in science.

'Coding' and 'encoding' is the archaeological deposit of a thoroughly bad theory of interpersonal communication, sometimes ironically referred to as the 'conduit' theory. The theory is as ancient as it is wrong headed. It is based on a picture of seventeenth century origin of the process of interpersonal communication in which a thought in someone's mind is encoded in language, then recoded as spoken sounds, and in that form crosses the abyss to another person who decodes it, first from sounds into words, and then from words into thoughts. As a metaphor it supposedly links the matter perceived, for example, and the stored 'entity' that represents it in the code.

With the exception of the thoroughly misleading metaphor of 'coding', which it would be well to do without, the other major metaphors have earned their places.

3. Transforming a cognitive model into an AI simulation

Why is contemporary cognitive psychology of remembering proto-AI?

It is very easy to add the principle that abstract functional units are models of real morphological units and that abstract processes are models of real processes to the methodology of cognitive science. These are two distinct principles since there may be process models which do not require 'cognitive organs' other than the whole brain or indeed the whole nervous system.

We have seen how deeply current cognitive psychology of memory is influenced by AI metaphors. This may be quite subtle. For example the idea of a memory store can be found in Plato, where it is criticised by Socrates. But as used for example by Baddeley in the post computer age it is surely infected by its usage in computer engineering. And of course that goes for ordinary language as well. Metaphor is an interactive trope, so calling a part of a computer its

'memory' borrows from common usage, while common usage begins to be affected by the widespread use of this metaphor. Here are two examples from Engelkamp and Zimmer (1994: 98) : 'word node' meaning a microprocessor relating words to one another, and the phrase 'speaking and acting is program activated' meaning 'when people were asked to name pictures when looking at them they remembered them better than if just looking'.

This should not be taken as in any way a defect in cognitive psychology. On the contrary it is the essential condition that will make it possible for it to grow into a full blown scientific account, by the standards of the physical sciences. Models are to be judged by their ability to account for empirical 'facts', *and by their ability to map on to a level of reality different from that in which those facts are generated.*. In this case we have discourses in the realm of observable matters of fact and brain structures and processes drawn from the deep level of reality, relative to public discursive activities. Using our general Task/Tool metaphor, which should control the whole of cognitive science, we can say that tasks are specified in the discursive realm and tools and their ways of working in the neurological world. The methodology which we are studying in this chapter should enable us to relate the one to the other. The pattern will be more or less the same as that which allows a physicist to relate the aurora borealis to ionised gas molecules.

Commenting on the recent history of AI modeling of cognitive procedures. Barnes and Hampson (1997) remark that 'much of the connectionist research conducted during the 1980s was of the demonstration variety. In effect, connectionist scientists were content to develop models that successfully simulated a certain type of behavior ... More recently, however, connectionist science has been interested in constraining its models with neurophysiological data. A connectionist model, iteis argued, should not only simulate a particular performance, but should also be designed and operate in accordance with what is known about neurophysiological structures and processes'[6]

Mental Models and Models of Mentation

The idea of 'mental models' offers a very different way of setting up a simulacrum of cognition from the model building by Baddeley. His model is a picture he has sketched of what is happening in the cognitive processes the outcome of which he can observe. It is a model of a possible architecture and processes that it could sustain. But 'mental models' are analogues of the structures formed in the

[6] This point has been often been made by theoreticians of AI (Harre, 1988), though until recently this was ignored by psychologists.

brain and nervous system. They are like or indeed some think they literally are maps. We can get a very good idea of this proposal from studies of remembering spatial matters of fact[7].

Radvansky and Zachs (1997) introduce a summary of Tversky's original studies by explaining what they mean by a model of a situation. They say that 'a situation model should have a structure that is analogous to the situation in a real or imagined world that it represents' *(op. cit.:* 181); that is it should be an iconic model of the sort with which one is so familiar in physics. The grounds for believing in a situation model, for them, are empirical adequacy, that is 'during retrieval of information [remembering] from a situation model, evidence of this spatial structure should be observed in the data'[8].

There have been a number of studies in the course of which the notion of mental model of a spatial situation has become more robust. There is an interesting and direct relationship between these studies and recent work in discursive psychology. The phenomenon of 'foregrounding', that is that objects closer to the story teller are more readily recalled, suggests that the spatial model is 'shaped' by using the story-teller as its geometrical origin. In her studies of Plains Indian sign language and its use in story-telling Famell (1995) has shown that a story- space is gesturally created by the story teller to create a framework for the location of objects and incidents in the tale to be told. The model is constructed in real space, but it is functionally equivalent to the mental model in 'inner space'.

It should be clear that this kind of modeling is quite different in spirit, in content and in the directions of research which open out from it, from Baddeley's 'models of mentation'. Mental models are already in the realm of realist AI as iconic models of possible brain architecture, whereas Baddeley's are two steps short of that position.

4. A Worked Example: From Discursive Analysis to Brain Architecture and Process

Introduction

Learning and memory are integrally interwoven as cognitive processes. We do not say, literally, that we remember something of genetic origin. There are

---

[7] The reader will notice that the 'encoding' metaphor is still being used in the quotations from a recent (1997) discussion of spatial modeling, even though the theory is based on a quite different method of representation.

[8] There is something «fatal' about psychology. This is methodological nonsense, since the model was built in the basis of just such data. It cannot be *evidence* for a spatial structure, but the display of it!

expression such as 'remembering how to smile' used perhaps of someone at last coming out from under the shadow of a doomed love affair, but the metaphorical character of such expression is obvious.

What is the role of the hippocampus as an organ in the whole gamut of memory machines? The usual way of answering such questions is logically incoherent but practically efficacious. If a function is no longer displayed by a whole organism when a part of it has been damaged then in the intact organism that part was the organ responsible for that function. Thus it is generally agreed that the loss of function in individuals with hippocampal damage is in declarative remembering, but only for recent incidents of the relevant type. By using the above principle we get the positive claim that the hippocampus is the organ of some aspect or stage of declarative remembering.

The principle upon which cognitive science rests is simply that between a psychological account of a remembering process and a neural account of the organ that people use to perform that process, there must intervene an AI model, in particular a connectionist model;. That is tied to neural architecture by the synapse/node relationship that is basic to connectionist AI.

To illustrate this we will look briefly at Roll's account of the hippocampus as a remembering organ.

There is a thoroughgoing mapping between synapses in a real neural system and nodes in a net. In the neuroanatomy of real neural organs the relationship can work in either direction, from real neural architecture to net structures, or from net structures to real neural architecture. In accordance with this analytical scheme there are taken to be three *sets* or *fields* of neurons in each of the left and right hippocampus. Each contains a million or more cells. The pattern of excitation passes from the dentate gyrus successively to two further fields of cells, the whole acting as if it were a sequence of trained nets. Thus the AI model becomes the source of anatomical and physiological *hypotheses* about the structure and processes of the hippocampus, and successively, the relevant parts of the cortex. The result of using an AI connectionist model to suggest how the hippocampus 'works' in performing one cognitive task relevant to remembering semantic facts is nicely set out in Barnes and Hampson (1997: 518-519). We have seen how a neural net can be trained to pick out the entities that fall under a certain category. Barnes and Hampson use the metaphor of 'extracting a common frame'. Of course we know that neural nets do nothing of the sort, though they may look as of they do, since they manage non-Linnaean classifications, that is classifications that do not depend on necessary and sufficient conditions being explicitly entered in to the computing machine.

A more detailed net model that exemplifies the methodology we are discussing is McCelland's (1995) account of some aspects of the autobiographical memory machine.

Rolls (1990) treats the hippocampus as if it were a three-fold system of neural nets, that is as if were indeed a connectionist device. This treatment illustrates very clearly how the study of memory in the Tool/Task framework is a three or four stage pattern of analogies. In many treatments, and indeed especially clearly in that by Rolls, the relationship between the three conceptual system that are juxtaposed to one another is obscured by the persistence of the metaphors of storage, information and representation. These are concepts appropriate to the kind of models which are constructed by cognitive psychologists on the basis of the public memory performances they observe. There are no mental entities, especially those which the metaphorical concepts of 'information', 'storage' and 'processing' seem to suggest. 'Information' is a metaphor which ties the ordinary everyday business of remembering into the research process. 'Storage' sets up a certain naive model of how 'information' is kept. Rolls actually says (Rolls, 1997: 102) '[there is] a general problem affecting storage (i.e. learning) ...'. So to 'store' something is to learn it. And that is *all* it is. 'Representation' carries the old picture further, a picture which finds no place at all in Rolls' actual work, since there is nowhere in a neural net at which there is a representing symbol for any particular item of information, though, again in Rolls, (1997: 103) he says '... each episodic memory must be separately represented in CA3 ensembles...'. To use the word for what a whole trained subnet ensemble does is potentially misleading and we will avoid it.

The point, of course, is not to criticise Roll's pioneering work in cognitive science. Nor should it be inferred that there is something wrong with the modelbuilding methodology that he and Treves have adopted. On the contrary, in following his work through from 1990 to 1997 we are following an exemplary piece of scientific thinking. But, we need to know how the work is done, what the cognitive devices he is employing to understand the cognitive devices that are involved in remembering, for instance certain spatial aspects of our material environment. I have been arguing that progress in this area of research could not have been made without the intermediate model building that inserts a dynamic structure between public performance and neural architecture and activity. At the same time warning signals need to be emitted to alert the reader of cognitive neuroscience to the shift of ontological basis in the transition across types of models from the purely imaginary to the concrete and researchable. For the most part this distinction is so obvious in physics and chemistry that few are misled. Unfortunately this is not true in psychology.

Commentaries, like the one attempted in this paper, are an essential part of the research process,. They alert us to the complex patterns of reasoning and the subtle ways that moves in scientific thinking delete and insert ontological assumptions. It is terms of these assumptions that 'next step' advances are made. By bringing them clearly to light we know where we are presupposing the

existence of something material and where we are spinning useful fictions out of the repertoire of metaphors that are indigenous to the science in question.

I am grateful to my colleague, Darlene Howard, for helpfull comments or an early draft of this chapter.

References

Atkinson, R. C. & Shiffrin, R. M. (1968) 'Human memory: a proposed system and its control processes' In K. W. Spence & J. T. Spence (Eds) *Psychology of Learning and Motivation* New York: Academic Press, Vol. 2.

Baddeley, A. (1998) *Human Memory: Theory and Practice* Boston & London: Allyn and Bacon.

Ellis, J. A. (1996) 'Prospective memory or the realisation of delayed intentions' In.M. A. Brandimonte, *Prospective Memory: Theory and Applications* Hillsdale, N. J.: Earlbaum, pp. 1-22.

Engelkamp, J. & Zimmer, H. D. (1994) *The Human Memory* Seattle: Hogrefe & Huber. Famell, B. (1995) *'Do you see what I mean ': Plains Indian Sign Talk and the embodiment of action* Austin, Texas: University of Texas Press, pp. 193-210.

Gamham, A. (1997) 'Representing information in mental models' In M. A. Conway, *Cognitive Models of Memory* Cambridge, Mass.: MIT Press, Ch. 6.

Gathercole, S. E. (1997) 'Models of verbal short term memory' In Conway *op. cit.* Ch. 2. Harre, R. (1988) 'Wittgenstein and Artificial Intelligence' *Philosophical Psychology* **I** 105-115.

McClelland, J. L. (1995) 'Constructive memory and memory distortions' In D. L. Schacter (Ed.) *Memory Distortions; how Minds, Brains and Societies Reconstruct the Past* Cambridge, Mass.: Harvard University Press, pp. 69-90.

Neisser, U. (1976) *Cognition and Reality* San Francisco: Freeman.

Radvansky G. A. & Zacks, R. T. (1997) 'Retrieval of situation-specific information' In Conway, *op. cit.* Ch. 7.

Rolls, E. C. (1989) 'The representation and storage of information in neural networks in the primate cerebral cortex and hippocampus' In R. Durbin, C. Miall & G. Mitchison *The Computing Neuron* Reading., Mass.: Addison-Wesley.

Rolls, E. T. (1997) 'Brain mechanisms of vision, memory and consciousness' In M. Ito, Y. Miyashita and E. T. Rolls (eds.) *Cognition, Computation and Consciousness* Oxford: oxford University Press, Ch. 6.