AI & Soc (1990) 4: 73–90

AI & SOCIETY

Connectionism, Classical Cognitive Science and Experimental Psychology

Mike Oaksford, Nick Chater and Keith Stenning
Centre for Cognitive Science, University of Edinburgh, UK

Abstract. Classical symbolic computational models of cognition are at variance with the empirical findings in the cognitive psychology of memory and inference. Standard symbolic computers are well suited to remembering arbitrary lists of symbols and performing logical inferences. In contrast, human performance on such tasks is extremely limited. Standard models do *not* easily capture content addressable memory or context sensitive defeasible inference, which are natural and effortless for people. We argue that Connectionism provides a more natural framework in which to model this behaviour. In addition to capturing the gross human performance profile, Connectionist systems seem well suited to accounting for the systematic patterns of errors observed in the human data. We take these arguments to counter Fodor and Pylyshyn's (1988) recent claim that Connectionism is, in principle, irrelevant to psychology.

Keywords: Connectionism; Classical cognitive science; Experimental psychology; Memory; Inference

Introduction

There has been an enduring tension in modern Cognitive Psychology between the computational models available and the experimental data obtained. Standard computational models have assumed the symbolic paradigm: that it is constitutive of cognitive processes that they are mediated by the manipulation of symbolic structures. Such schemes easily handle formal inferences, and memory for arbitrary symbolic material. However, context sensitive defeasible inference and content addressable memory retrieval have remained problematic. By contrast, in the empirical data on human memory and inference, the opposite profile is observed. Everyday mundane reasoning is both context dependent and defeasible, and yet is performed easily and naturally, whereas subjects are typically unable to perform the simplest formal reasoning task (Wason and Johnson-Laird, 1972; Evans 1982). In memory, content addressable access in knowledge rich domains seems natural and unproblematic for human subjects, whereas people can retain only very small quantities of arbitrary material. Despite this tension between experiment and theory, Fodor and Pylyshyn (1988) have recently reaffirmed what they term the "Classical symbolic paradigm". That is, they argue that symbolic cognitive processes are autonomous from their implementation. Thus they question the relevance of Connectionist theorising for psychology, and suggest that Connectionism should be viewed as a theory of implementation for autonomous Classical architectures. We have argued (Chater and Oaksford, in press) that cognitive processes are not autonomous from their implementation, and that interaction between Classical and Connectionist models is required to explain much of cognition, on *computational* grounds. In this article, we review some *empirical* grounds for non-autonomous accounts of cognition. We raise problems for the Classical (autonomous) approach, and suggest that Connectionist (non-autonomous) models may have more explanatory power. Let us briefly summarise the main points of the debate between Classicist and Connectionist.

Classicism versus Connectionism

Fodor and Pylyshyn (1988) claim that Connectionism eschews structured representations and thus amounts to an unwitting return to associationism. They raise two problems which must be resolved by any cognitive theory. First, *constituency*: how do people keep track of symbols so that they can play the same role in different representations? Second, *inference*: how do mental states systematically track semantic relations? However, a large amount of Connectionism theorising explicitly addresses the issue of implementing structured representations such as semantic networks (Hinton, 1981), frames and mental models (Rumelhart, *et al.*, 1986) in PDP systems. So the issue of structured representation does not separate Connectionist and Classicist. Chater and Oaksford (1989) have argued that the real locus of the dispute concerns whether cognitive architecture is computationally and explanatorily independent of its implementation.

For Fodor and Pylyshyn (1988) psychology is the study of the symbolic structures and processes which define the cognitive level. They claim that since Connectionist architectures are *non-symbolic*, Connectionism is properly viewed as a putative implementational theory for a standard symbolic architecture. Since they take the cognitive architecture to be independent of its implementation, they conclude that Connectionism is irrelevant to psychology. It is this *autonomy* assumption which separates Classicist and Connectionist (Chater and Oaksford, in press). The Connectionist believes that (i) the primitives of the cognitive level architecture depend on the implementational substrate (i.e. a neural network), and (ii) the cognitive processes which psychologists can postulate crucially depend on these primitives. So a psychological account of memory and inference should pay close attention to the properties of neural implementation. And PDP is a preliminary attempt to characterise these properties. Given any particular *psychological* phenomenon the Classicist must give an explanation at the cognitive level, i.e. in terms of purely symbolic processes. On the other hand, the

Connectionist may freely advert to properties of both the cognitive (symbolic) level and the implementational (Connectionist) level.

The autonomous symbol processing paradigm has an implicit theory of memory and inference. A memory for an individual is simply a stored atomic symbol denoting the individual, a property of that individual is simply another symbol or symbolic structure. That is, a memory is a proposition encoded in some logical language. Memory is just a repository for formulae of this language. Inference consists of formal operations defined over these formulae. On this view the psychology of memory delimits the capacity and retrieval operations of the data base; the psychology of reasoning specifies the nature of the theorem prover. Both may set constraints on the nature of the data structures implemented in human memory (e.g. Collins and Quillian, 1969; Johnson-Laird and Steedman, 1978).

Prima facie this Classicist model of cognition makes strong predictions for human memory and inference. Memory for individuals and their properties should be trivial; if the VAX can store and retrieve arbitrary lists of symbols, then humans should too. Inference should be trivial; any simple theorem prover can solve two-line propositional logic problems, so people should find them easy too. In this article, we will show that these predictions of standard computer models are in tension with the psychological data on memory and inference. We will argue that non-autonomous Connectionist explanations, which advert to both implementational and cognitive levels, may be better able to model human peformance. We now examine the psychological data on memory and inference.

The Psychological Data

The most obvious experimental approach is simply to have subjects learn arbitrary lists of properties and perform formal reasoning tasks. On the Classicist position it seems that human performance on simple memory and reasoning tasks should be close to perfect. In the following sections we adduce evidence which appears to violate these predictions.

Memory

One principal empirical motivation for the development of Connectionist systems has been the tension between the gross operating characteristics of human memory and those of standard symbolic devices. For example, human memory appears to be content addressable, and display graceful degradation – human memory appears tolerant (within bounds) to degraded inputs and damage (Bobrow and Norman, 1975; Norman and Bobrow, 1975, 1976, 1979). In contrast, standard symbolic devices are not content addressable and they will fail to retrieve a memory trace in the face of damage or degraded input. In this section we amplify on these criticisms of the Classicist view of memory, by considering (i) the abundant evidence on the limitations of the human memory system, and (ii) the specific way in which people bind together the properties of individuals in memory.

It has been a fundamental observation in the psychology of human memory that recall for arbitrary material is severely limited. Short-term memory tasks for arbitrary material reveal serious memory capacity limitations. To a first approximation, subjects can only recall 7±2 "items" (Miller, 1956). Human memory performance for such material is also severely limited in long-term memory tasks (Baddeley, 1976; Lindsay and Norman, 1977). Typically the only way of improving performance is by imposing a high level organisation using mnemonic strategies (Lorayne and Lucas, 1974). Similarly, short-term memory performance is bounded by severe resource limitations which may, in part, be overcome by the use of high level recoding strategies or "chunking" (Miller, 1956).

Performance is superior with normal structured material because subjects need only recognise a pre-existing organisation rather than impose their own (Bartlett, 1932). People seem to be remarkably good at exploiting redundancy in memory for large amounts of complex, structured material. For example, people have a reliable recognition memory for very large numbers of briefly presented photographed scenes (Shepard, 1967). In contrast, it is extremely hard to recall the layout of an impoverished visual stimulus, such as the fixed stars. However, people find it much easier to recall a much richer stimulus, such as a face. Faces are of course highly redundant, whereas the position of the stars is not. People attempt to remember such arbitrary material by imposing meaningful interpretations upon them. In the case of the stars, the naming of the constellations attests to the utility of such a strategy. In general, the more coherent the subjects find the material to be, the better it is remembered (Craik and Lockhart, 1972; Bransford and Johnson, 1973).

This data reveals a mismatch in the gross organisation of human memory and Classical computational models. We now turn to more recent experiments in which the fine detail of memory performance is examined. The MIT task (Stenning and Levy, 1988; Stenning *et al.*, 1988) requires subjects to recall the appropriate *bindings* of properties to individuals. It thus addresses the question of how people keep track of structured knowledge about individuals. Two individuals are described in a short paragraph of text and on recall subjects select the properties which attach to each individual. If each individual has one of two professions, nationalities, temperaments and statures there are 136 possible pairs of individuals. The task is to remember which of these pairs was presented.

Analysis of the pattern of errors reveals the organisation of the underlying representations. The average error rate is just half a property wrong per paragraph. Errors are clearly interdependent: there are more multiple property errors than would be expected if errors on each dimension were independent. A frequent error is assigning two different values (say tall and short) to the wrong individuals. What is difficult is *binding*, i.e. remembering whether it was a Polish bishop and a Swiss dentist or a Polish dentist and a Swiss bishop. It is easy to remember whether they were both Swiss or both Polish. Further, multiple errors are more common than would be predicted if each of the eight properties were represented independently. If the properties for each individual were stored in separate logical formuale such statistical dependence would not be expected. On a Classical architecture binding is trivial and so it is unclear why there should be any interference between separate memory traces at all.

The patterns of performance revealed here seem to be wholly unexpected, on the Classicist view. If the very basis of memory is the storing of formal symbolic structures, then encoding arbitrary lists should be trivial compared wth storing complex material, where it is opaque how the stored memory (a face, or the gist of a passage) can even be represented in propositional form. If such material can be represented propositionally, then the structures required will surely be extremely elaborate. On a Classical architecture, the property list should be easy to remember and the complex material hard to remember. Yet the data on human performance show precisely the reverse.

Adversion to the competence-performance distinction is of no avail here. If this distinction applies to the study of memory, then (i) content addressability, (ii) graceful degradation, and (iii) capacity limitations provide the data which characterise the competence state requiring explanation. Similarly, grammaticality judgments in linguistics characterise the competence state captured by the competence grammar. Competence theories, moreover, must account for the majority of the evidence. Otherwise they cease to be falsifiable empirical theories. The classical model is at variance with findings (i) and (ii), can only account for (iii) via arbitrarily imposed *ad hoc* limitations, and moreover fails to explain subjects' *systematic* binding errors. Hence it could only be preserved as a model of human memory on the assumption that it is impervious to falsification. In this case the Classical account is not a cognitive model but rather an *a priori* assumption concerning the nature of computation in the brain which could only be legitimised in the absence of a competitor. Connectionism may provide the competition which promotes the mismatch between the Classical model and the empirical data to falsificatory status.

Inference

The Classicist has an implicit theory of memory and inference. The data we have surveyed on memory (Stenning and Levy, 1988) appear to violate the expectation that a declarative memory should easily keep track of simple property lists. In this section, we address similar difficulties with the Classicist predictions for inference. If symbol manipulation forms the very basis of cognition, then trivial deductive inferences should be trivial for human subjects. However, data on deductive reasoning appears to violate the Classicist's predictions.

The conditional *if . . . then* construction is central to formal attempts to characterise inferential processes in logic. However, relative to normative logical theories the human data presents a problem. Human conditional reasoning appears beset by various non-logical biases apparently reflecting the influence of content (Wason and Johnson-Laird, 1972; Evans, 1982), memory (Griggs and Cox, 1982), and prior beliefs (Pollard and Evans, 1981). Thus, normative conceptions of rationality appear radically at odds with people's observed facility for logical thought. The most striking demonstrations of apparently non-logical performance occur in the Wason Selection task (1966).

Wason's task concerns how people assess evidence relevant to the truth or falsity of a rule expressed by means of a conditional sentence, normally using the *if . . . then* construction. Subjects were presented with four cards, each having a number on one side and a letter on the other. On being presented with a rule such as, "if there is a vowel on one side, then there is an even number on the other" and four cards, showing, e.g. an "A", a "K", a "2" and a "7", they would have to select those cards they *must* turn over to determine whether the rule was true or false.

The Classicist expectation was that performance would depend on the *logical form* of the construction used to express the rule. Responses should be predictable from the truth tables which supply the meanings of the various logical terms. A conditional, $p \rightarrow q$ is true just in case either p is false or q is true; conversely it is false just in case p is true and q is false. So, the rule is true if and only if each card has a consonant on one side or an even number on one side. Conversely, it is false if any card has a vowel on one side and an odd number on the other. Hence, only the "A" card and the "7" card must be turned over. If a subject is exhaustively trying to make the rule true then these are the only undecided cases. If a subject is trying to make the rule false then these are the only cases with the potential to do so.

Typical results were: p and q card only (46%), p card only selected (33%), p, q and $-q$ card selected (7%), p and $-q$ card selected (4%), others (10%) (Johnson-Laird and Wason, 1970). The logical response was observed in only 4% of subjects responses. Even more puzzling, on a proof theoretic account, are the various manipulations which facilitate performance on this task.

The most significant manipulation to produce a marked facilitation of the logical response was in the thematic variants of the task where realistic or contentful materials were used. Wason and Shapiro (1971) employed the following rule: "Every time I go to Manchester I travel by car". Presented with four cards indicating travel destinations on one side and modes of transport on the other, 63% of subjects made the logical response. It appears that the more sense subjects can make of the materials the more likely the are to perform logically. Yet on the Classicist view this is puzzling, since logical reasoning depends only on the *form* not on the *content* of the materials.

While subjects fail on deductive reasoning tasks which are trivial in symbolic architectures, human common-sense inference is far more sophisticated than that of any AI system. On the Classical view, common-sense inference should involve very elaborate logical reasoning, defined over large sets of premises encoding the relevant world knowledge. According to this view, common-sense reasoning should be more difficult than simple deductive reasoning. Yet the data on human performance show precisely the reverse. Minsky's (1975) frame theory was an attempt to capture the structure of the routinely performed defeasible inferences involved in the comprehension of even the simplest texts. It is unclear that such mundane reasoning can be constructed successfully in standard symbolic architectures (McDermott, 1986; Chater and Oaksford, 1989).

This data presents a severe challenge to the Classicist view of cognition as autonomous symbol manipulation. Reasoning appears to be crucially dependent on assimilating the task to world knowledge rather than extracting the formal structure of the task description. Subjects appear to be using *common-sense* knowledge–rich reasoning strategies, even when confronted by a formal task. Species of common-sense non-demonstrative reasoning, like induction, abduction and default inference are unlikely to be formalisable, *in logic* (from an AI perspective: Israel, 1980; McDermott, 1986; from the philosophy of logic: Harman, 1986; from the philosophy of science: Kuhn, 1970; Masterman, 1970; from cognitive science: Chater and Oaksford, in press). The Classicist attempts to explain common-sense reasoning in terms of logic, which may be impossible in principle. So, proof theory can not be the basis of cognition. Rather, it seems that whatever (non-logical) processes underpin common-sense reasoning may also underpin performance on logical reasoning tasks. (For an account of this data which relies on common-sense principles of reasoning, see Oaksford, 1988; Oaksford and Stenning, 1988.) As with memory, adversion to the competence-performance distinction is of no avail considering that up to 96% of the observed behaviour on Wason's task is beyond the scope of the competence model.

Theory

Prima facie in both memory and inference tasks, the richer the informational content of the materials used, the better performance seems to be. This seems mysterious on a Classical account. To encode a richer stimulus should require more formulae to be stored in the declarative data-base, and so both storage and retrieval should become harder. On the Classical view, resource limitations should become more acute with richer stimuli. However, the psychological evidence seems to indicate that in human memory, performance is not impaired but *enhanced* with richer stimuli. Turning to inference, on the Classicist account of cognition, common-sense reasoning involves the construction of elaborate logical derivations. Common-sense reasoning should be very much harder for human subjects than proof-theoretically trivial logical deductions. Precisely the opposite performance profile is observed in the psychological data.

It seems that as the stimulus is made richer, subject's performance improves. Yet the stimulus must be enriched in an appropriate way. For example, a property list may be "enriched" by making it longer, making the properties more complex and so on. An inference task may be "enriched" by requiring that the subjects must use more premises, or must construct a more elaborate chain of deduction. Plainly such "enrichment" of the task will lead to worse rather than better performance. What sorts of "enrichment" lead to enhanced performance? We shall discuss this question in memory and inference in turn.

Memory

One simple suggestion is that the task will only become easier if the stimuli are made redundant. That is, each portion of the stimulus is not independent of the rest. Rather, there is organisation within the stimulus which allows each part (at least to some extent) to be predicted from the rest. Let us consider an example inspired by Miller and Selfridge (1950). It is very hard to remember a list of letters such as:

1. D P H O J U J P O

However, it is far easier to remember a list of letters which are organised into some meaningful configuration.

2. C O G N I T I O N

Remembering this sequence does not feel like remembering a sequence at all but rather is like remembering a single item, i.e. a word. In the case of (2) our lexical and morphological knowledge allows us to exploit the redundancy within the stimulus. This redundancy allows us to reconstruct the whole item from only a partial cue. Very simple-mindedly, if you recall that (2) began COG . . . , then you are highly likely to recall the whole word. On the other hand, recalling that (1) starts DPH . . . seems to help not at all in reconstructing the rest of the list. However, apparently unpredictable sequences such as (1) can be seen as redundant *if* they can be assimilated to appropriate knowledge. And indeed a simple transformation (shifting each letter along one place in the alphabet) maps (1) on to (2). So once this is realised sequence (1) becomes as predictable as (2).

Quite generally, recall has been found to depend on the degree to which subjects impose an organisation on, or can make sense of the materials to be learnt. For example, in text comprehension, Bransford and Johnson (1973) had subjects read a passage with or without the title *Watching a peace march from the fortieth floor*:

> "'The view was breathtaking. From the window one could see the crowd below. Everything looked extremely small from such a distance but the colourful costumes could still be seen. Every one seemed to be moving in one direction in an orderly fashion and there seem to be little children as well as adults. The landing was gentle and the atmosphere was such that no special suits had to be worn. At first there was a great deal of activity. Later, when the speeches started, the crowd quieted down. The man with the television camera took many shots of the setting and the crowd. Everyone was very friendly and seemed to be glad when the music started.'
>
> After hearing the passage *Ss* were asked to recall it. Most sentences were recalled well except for the one about 'the landing'. There was extremely low recall for this sentence, and *Ss* noted that there was one sentence (i.e. about the landing) that they could not understand. Even when presented with a 'cue outline' (e.g. Luckily the landing – and the atmosphere –), *Ss* exhibited very low ability to remember what this sentence was about.
>
> A second group of *Ss* heard the identical passage but with a different title: *A space trip to an inhabited planet*. These *Ss* showed much better free recall of 'the landing' sentence than did the first group, as well as greater ability to fill in the gaps in the cue outline presented above (see Bransford and Johnson, 1973). These results suggest that a sentence that would be comprehensible in isolation (i.e. 'the landing' sentence) can become incomprehensible when viewed from an inappropriate context, and that such incomprehensibility has a marked effect on ability to recall." (Bransford and McCarrell, 1975)

In the present discussion the moral is as follows. If a particular item to be recalled fits with the context in which it occurs, then it is easier to recall. Given that the subjects know what the story is about, the contents of the sentences become predictable (i.e. redundant). The pattern of errors indicates that the principles that underlie normal successful recall relate to knowledge of the world. Given that subjects know the passage is about a space trip to another planet, the sentence about the landing is highly predictable. However, when the sentence violates the expectations of the subjects the sentence is no longer predictable and so a straightforward redundant coding is not possible. So it seems that recall is a matter of finding the item which *best fits* the residual memory trace given world knowledge. World knowledge provides a vast interlocking array of defeasible constraints encoding the predictable organisation of the world. Best-fit matching involves maximising the degree to which the vast number of contextually relevant constraints are satisfied. A memory which exploits these constraints will easily assimilate predictable material since, inasmuch as the novel stimuli cohere with existing constraints, only minimal adjustments will be required to lay down the memory trace. Much information may be left implicit, since it follows from world knowledge. If the material is unpredictable, world knowledge cannot be exploited, and so the information must be encoded explicitly.

Any such best fit memory system will possess a variety of properties of human memories (Bobrow and Norman, 1975; Norman and Bobrow, 1975, 1976, 1979). Generally each part of a memory can function as the *context* of recall for the remainder of the memory. A best-fit memory will use predictability to reconstitute a whole memory from a sufficiently large arbitrary fragment (i.e. memory is *content addressable*). Such a memory should be resistant to corruption, as a sufficiently large veridical fragment should be able to reconstitute the original trace (i.e. memory is *noise resistant*).

Unorganised material, like arbitrary property lists, is hard to remember because there are no constraints between the items in the list. Memory relies on the ability to impose organisation on the stimulus. So, the more deeply that a stimulus is processed (Craik and Lockhart, 1972) and the better it is understood, the better memory performance will be. Mnemonic strategies (Lorayne and Lucas, 1974; Young and Gibson, 1962; Bower, 1970) can be seen as enriching the stimulus so as to impose an organisation upon it. So, paradoxically, remembering more can allow you to remember better (cf. Stenning and Levy, 1988, on the "Mnemonic Paradox"). It is unclear whether Classical accounts are capable of displaying these properties (Chater and Oaksford, in press).

Inference

In inference it has been shown that facilitation was not simply determined by the use of contentful material (Griggs and Cox, 1982). Rather, subjects needed to be *familiar* with the rule. Prior experience with a contentful rule was shown to facilitate reasoning, whereas contentful rules with which subjects had no prior experience where shown to be non-facilitatory. Specific experience of a rule means that it need not be assimilated afresh.

Yet in reasoning this can not be the whole story. To suggest that successful reasoning with conditionals requires having reasoned with them before is altogether too restricting. It precludes people from generalising experience to novel domains, which is surely the *sine qua non* of human rationality and reasoning abilities.

More recent work (Cheng and Holyoak, 1985; Cheng *et al.*, 1986) serves to indicate that what needs to be right about contentful material also involves the *relations* encoded by particular conditional sentences. When subjects were provided with a *rationale* explicitly characterising a permission relation, a significant facilitation was observed over their performances for both a thematic variant (no prior experience) and an abstract variant, neither of which was provided with a rationale. The facilitation to the logical response was explained in terms of the production rule set which constitutes a *permission schema*. This left open the possibility that other *pragmatic reasoning schemas* may drive inference when other relations, e.g. causation, are encoded producing different patterns of inference. Assimilation of the task to *appropriate* world knowledge will facilitate response profiles which mirror logical performance, without invoking logical competence. In an abstract task, in which the materials cannot be assimilated to world knowledge, subjects cannot rely on pragmatic reasoning schemata, and may have to default to more general non-logical strategies. Since the rules by which we understand the world are typically defeasible, as we now show, it is not surprising that logical inference is so unnatural.

The data reviewed above suggest that the processes of inference are determined by world knowledge, however such processes may be implemented. Only by understanding the organisation of the world do you know that one thing follows from another. For example, you can not infer Fred is going to buy his morning paper from his passing your window at 9.00 a.m. if you don't know Fred's habits. On a proof theoretic account this knowledge might be encoded in first order predicate logic as a statement in some declarative data base, quantifying over events (e) and times (t):

1. $\forall e \forall t[(\text{Fred\_passes\_your\_window}(e,t)\ \&\ t = 9.00\text{am.}) \rightarrow$
   $\exists e'\ \exists t'(\text{Fred\_buys\_his\_paper}(e',t')\ \&\ \text{just\_precedes}(t,t'))]$

However, as we argued above, such a strategy can not work in general, since all these common-sense generalisations about the organisation of the world succumb to indefinitely many counter-examples.

If I see Fred going past my window at 9.00 a.m., I know he's about to buy his morning paper. But not if it's Christmas day, since there are no papers; but not if he's being mugged; but not if he's already reading *The Times*; but not if he falls over and breaks a leg; but not if his ferocious wife is with him wielding a shopping bag . . . etc. All of these possibilities override our generalisation that Fred buys a paper just after passing my window every morning. On a proof theoretic account the various conditions that might override our rules must be explicitly encoded, and a check made that none of them apply in any specific case. This is the standard approach to defeasible common-sense inference which has been pursued in the logical/AI literature (Reiter, 1985; de Kleer, 1986; McCarthy, 1980). However, these approaches all involve consistency checking, which is an NP hard problem. This means that such symbolic methods are computationally intractable.

Although this technical problem need not decide against a proof theoretic account of inference, there are more general difficulties for these symbolic methods. There are indefinitely many conditions which may override my inference about Fred's buying his morning paper. Indeed we can invent them at will. These counter-examples are not merely rhetorical devices – they would in the right circumstances block the inference rule. For the Classicist, it seems that each of these possibilities must be explicitly encoded in the appropriate rule. To avoid an infinite list of default clauses we must appeal to a finite taxonomy which captures the infinitude of specific cases. Perhaps Fred will not get his morning paper in distracting situations, dangerous situations and so on. However, what counts as, say, a distracting situation is relative to what rule we are considering. A road accident might count as a distracting situation *vis-à-vis* Fred's getting his paper, but not *vis-à-vis* his getting to work. So it seems that categories in our taxonomy must be rule relative – i.e. the precise sense of "distracting situation" must be spelt out in detail *in each rule*. It is a considerable act of faith to believe that such specifications will be forthcoming. Default rules infect lexical inference as well as structural inference.

Common-sense defeasible inferences, if derivable *at all* on a proof-theoretic account, must be extremely complex. For people, these inferences are just *common sense*. This mode of inference underlies comprehension, categorisation, perception and action (Bransford and Johnson, 1972, 1973; Bransford *et al.*, 1972; Bransford and McCarrell, 1975; Clark and Haviland, 1977; Minsky, 1975; Murphy and Medin, 1985). It is the basis of all cognitive performance. People find this mode of inference easy. So, the computational architecture of the mind must be so constituted as to facilitate such inferences. But the Classical proposal of an architecture based on symbol manipulation seems unpromising. So, perhaps the non-logical performance on tasks such as the selection task is not so surprising in view of these considerations. The strategies which subjects adopt when assimilation to world knowledge is impossible are none the less likely to reflect assumptions about the defeasible nature of the rules which we use to understand the world (Oaksford, 1988).

Modelling

The Classical account of mind may be incapable of satisfying several well established constraints on human cognition. Roughly, human memory should be able to implement large scale best-fit searches by exploiting the organisation of the world; and the inferences people draw should be determined by their knowledge of the way the world is organised (rather than by deductive rules of inference). Below, we first outline how Parallel Distributed Processing captures many of these desirable properties. We then turn to a specific model of Stenning and Levy's (1988) data, which accounts for the observed dependencies between stored properties as revealed in the analysis of the error data (see above). We also briefly outline how such a memory system will give rise to automatic generalisations which begins to suggest an alternative account of the data on human inference by suggesting a possible mechanism for default inference.

Memory

Let us first recap on the desirable properties of human memory:

1. Human memory is content addressable.
2. Human memory is noise resistant.

These are both direct consequences of the Parallel Distributed Processing approach to massively parallel soft constraint satisfaction or best-fit searches:

"One way of thinking about distributed memories is in terms of a very large set of plausible inference rules. Each active unit represents a 'microfeature' of an item, and the connection strengths stand for plausible 'microinferences' between microfeatures. Any particular pattern of activity of the units will satisfy some of the microinferences and violate others. A stable pattern of activity is one that violates the plausible microinferences less than any of the neighbouring patterns." (Hinton *et al.*, 1986, 80–81)

Given this intuitive picture it can be shown informally how the various properties emerge.

Content addressability. When an arbitrary sufficiently large fragment of the pattern is presented the microinferences act so as to produce the nearest possible completion (stable state). Hence, any sufficiently large part of the content of the memory will address the whole memory.

Noise resistance. If part of the presented fragment is wrong (the fragment is inconsistent with any stored memory), the microinferences will still find the best available fit. Hence a degree of noise can be tolerated, since the correct part of the fragment will ensure that the best fit pattern prevails.

In addition to being able to capture general properties of human memory, PDP models have also been used to model the fine structure of performance on particular memory tasks. For example, Stenning and Levy (1988) present a Parallel Distributed Processing model of the MIT task, discussed above. The binary features derived in a multiple linear regression analysis are used to represent the pair of stored individuals. At retrieval time, the first and second individuals and their properties must be reconstituted from the complex distributed representation in terms of matched and mismatched properties and logical combinations of properties (roughly, these express that *someone* is, say, tall and Polish; they make no explicit references to the first or second individual mentioned). A three-layer network was used, where the binary input units stand for the features derived from the regression model. These inputs units are completely connected to a layer of 16 hidden units which in turn are connected to two sets of four output units corresponding to the two individuals' properties. Just as the person can recall the individuals in either order, the network must be able to identify the individuals at either output location. The order of recall for the network is specified by an extra binary input.

The network learns the mapping from features to individuals by cycling through each logically possible input vector, each paired with its correct output (240 vectors in all). Learning reaches criterion after about 300 iterations of back-propagation. A good test of the adequacy of the representation postulated to underlie the human data is whether corrupting that representation generates similar error patterns to those found in the human data. The error patterns for the network were generated by coding the paragraphs human subjects saw into the input feature representations and then subjecting these representations to random noise. This involved flipping the values of the input features with a 3% probability. For comparison, the actual eight property descriptors were also directly subjected to a similar corruption with a 5% probability. This amounts to storing (and corrupting) each individual and their properties independently, as is most natural in a Classical framework. The match of the corrupted inputs to the observed human error data is good, although not as good as that of the original regression model. Stenning and Levy (1988) note various specific similarities and discuss possible explanations for these. The direct corruption, on the other hand, leads to quite different error patterns. So the character of a distributed Connectionist system may model the pattern of performance more closely than independent conventional symbolic architectures.

Inference

We have argued that an alternative must be found to Classical symbolic accounts of common-sense reasoning. Current symbolic schemes fail to characterise the most frequently occurring inference patterns of subjects in deductive reasoning tasks, such as Wason's Four Card Problem. Our diagnosis has been that *common-sense* inference underlies performance on *deductive* reasoning tasks and not the other way round (Oaksford, 1988; Oaksford and Stenning, 1988). Recent Connectionist theorising suggests a promising avenue for research in capturing the defeasibility of common-sense inference.

Hinton *et al.* (1986) discuss a novel implementation of semantic nets in PDP hardware (originally in Hinton, 1981):

"People are good at generalising newly acquired knowledge . . . If, for example, you learn that chimpanzees like onions you will probably raise your estimate of the probability that gorillas like onions. In a network that uses distributed representations, this kind of generalization is automatic. The new knowledge about chimpanzees is incorporated by modifying some of the connection strengths so as to alter the causal effects of the distributed pattern of activity that represents chimpanzees. The modifications automatically change the causal effects of all similar activity patterns. So if the representation of gorillas is a similar activity pattern over the same set of units, its causal effects will be changed in a similar way." (p. 82)

The similarity metric used in automatic generalisation is induced by pattern similarity. This similarity metric need not be specified by the programmer (who antecedently knows which generalisations he wants to go through), but may itself be learnt by the network (Hinton, 1987). Notice that gorilla is given "likes onions" as a default. Yet this default may be overridden by explicitly storing information to the contrary. Further the default rule may be overridden if "gorilla" has a similar pattern to "orangutan", whose representation does not include "likes onions". The similarity metric gives us plausible inference rules for free, and the autoassociative mechanism weighs up these various defeasible constraints to settle on the best-fit interpretation. Defeasible constraints just are default rules – and defeasible constraints are the very fabric of PDP implementations. Of course, this is merely a toy example. Yet it is suggestive of how implementation may unburden the cognitive architecture of the problems created by common-sense inference.

Classical approaches to common-sense inference have proved to be computationally intractable due to their reliance on consistency checking. So, it may be the case that a PDP account is the only straw afloat in arriving at workable mechanisms for defeasible common-sense reasoning. Not only do symbolic approaches seem unable to capture common-sense inferences, the data on Wason's Four Card Problem reveal that human reasoning performance is radically at variance with Classical expectations. Treating performance on deductive reasoning tasks such as Wason's Four Card Problem as mediated by common-sense inference can account for the most frequently occurring response patterns (Oaksford, 1988; Oaksford and Stenning, 1988). The possibility of PDP implementations provides the promise of models which can predict the *complete* response profile (for discussion see Oaksford, 1988). However, on a Classical account, *all* error must be assigned to performance factors. This seems unsatisfactory when up to 96% of subjects reponses on a logically trivial task like the Wason Four Card Problem are beyond the scope of the competence theory.

Conclusions

Fodor and Pylyshyn (Pylyshyn, 1973; Fodor, 1975; Fodor *et al.*, 1974; Fodor 1980; Fodor, 1983; Pylyshyn, 1984; Fodor, 1987, Fodor and Pylyshyn, 1988) have, over a number of years, defined the notion of computation as used in Cognitive Science. Cognition is computation. Computation is mechanised proof theory. So,

> "It would not be unreasonable to describe Classical Cognitive Science as an extended attempt to apply the methods of proof theory to the modelling of thought." (Fodor and Pylyshyn, 1988, 29–30)

This view carries both theoretical and methodological significance. As we have seen the theoretical implications for memory and inference seem not to accord with the empirical data. The implicit methodology fares little better.

On a Classical symbolic model, errors in memory cannot be explained at the proof theoretic, cognitive level, but rather must be seen as functions of the implementation. Similarly, apparent deviations from logical performance in conditional reasoning tasks must also be assigned to performance or implementational factors. According to the Classicist account such error data can therefore only inform us about the character of the implementation of the cognitive architecture and not about the cognitive architecture itself. For the Classicist the study of psychology is the study of cognitive architecture independent of its implementation. In consequence, the vast bulk of the empirical considerations discussed in this article do *not* count as *psychological* considerations. On such a restricted view of the domain of psychology, there is little psychology left. If we exclude the study of errors, then almost all experimental memory, reasoning and perceptual research is not psychology. Error methodology is a ubiquitous and a powerful investigative tool in psychological practice.

For example, the pattern of errors in Stenning's MIT task gives insight into the normal function of human memory. The subjects appear to represent the individuals according to a complex task specific set of features. For subjects to derive this encoding scheme means that they have extracted the particular organisation of the task. For example, since the dimensions are binary, a match–mismatch strategy is appropriate. The feature set of the multiple regression gives a redundant encoding of the individuals and their properties. However *that* this encoding is redundant is dependent on the structure of the memory task. This illustrates the general point that subjects attempt to find organisation in the material to be remembered so that it may be encoded redundantly. The MIT task demonstrates the remarkable facility with which such encoding strategies are devised and implemented. Such redundant encodings may be directly tied to a PDP model of retrieval processes. The exploitation of systematic redundancy by PDP processes operating over distributed representations is responsible for many of the desirable computational properties of these systems (that is Fodor and Pylyshyn's (1988) "lures" of Connectionism; see Chater and Oaksford, in press, for discussion).

For Fodor and Pylyshyn, psychology is taken to be the study of an autonomous Cognitive level, to which implementational considerations are held to be irrelevant. According to this view, Psychologists should not be interested in Connectionism (at best, they claim, a theory of implementation), just as they should not be interested in error. However, we have argued that:

- Autonomous Classical theories are radically at variance with human data on memory and inference. The limitations, patterns of errors and interference seem mysterious on the Classical view.
- *Contra* Classicist expectations, human performance in memory and inference improves as the stimulus becomes richer and can engage the vast interlocking array of defeasible constraints which encodes the predictable organisation of the world.
- Connectionist systems are able to exploit such redundancy and thereby emulate the observed characteristics of human memory, such as content addressability and graceful degradation.
- Human performance on prima facie symbolic tasks is heavily influenced by the content of the materials, in the tasks we have discussed. So insofar as symbolic processes are implemented in the brain (and Fodor and Pylyshyn's arguments for structured representation are persuasive on this point) they crucially interact with the properties of the implementation. Content effects cannot be explained *proof theoretically*!
- This leads naturally to the conclusion that modern cognitive psychology has always been committed to the interaction of cognitive and implementational levels. So Fodor and Pylyshyn's advocacy of autonomy is a *revisionary* restriction of the appropriate domain of psychological concern. If psychologists adopt this novel stricture then they will not be interested in Connectionism. However, accepting this restriction, would deprive cognitive psychology of most of its subject-matter and one of its most potent explanatory distinctions. Thus, issues of the implementation of cognitive architecture will continue to be a major source of constraint in psychological theorising. Hence, Connectionism may naturally be seen as providing new and important metaphors for thought, rather than as a psychologically irrelevant implementational theory.

Much of the data we have reviewed predate the emergence of the proof-theoretic view as the dominant meta-theory for cognitive science. The apparent incompatibility of data and meta-theory has been largely ignored. This is wholly legitimate while there is no competing computational paradigm which can address this data. Puzzling data only becomes falsifying data when an alternative explanation becomes available (Kuhn, 1970; Lakatos, 1970; Putnam, 1974). We believe that such an alternative computational and theoretical paradigm is now emerging in which much of this old data can be recast. However, we agree with the thrust of Fodor and Pylyshyn's (1988) argument that psychology must be concerned with structured representations, and structure sensitive processes, which were the very stuff of the symbolic paradigm. We believe that PDP should attempt to capture structured representations while retaining the desirable computational properties we have outlined. In particular, we require that our models be able to deploy rich, context sensitive world knowledge in a flexible and tractable way. Further, these models should be consistent with the empirical data, especially error data. Whether or not such a goal can be attained will ultimately decide whether PDP really does constitute a rival to the standard proof theoretic orthodoxy. In any case, this reconsideration of the data provides a challenge to the proof theoretician; it is unclear that this challenge can be met.

Acknowledgements

Special thanks are due to Mike Malloch who introduced us all to PDP and continues to be a major influence. We also thank Jerry Seligman and Robin Cooper for much fruitful discussion. Thanks are also due to the members of the PDP workshop at the Centre for Cognitive Science, University of Edinburgh, especially Richard Rohwer and David Willshaw. Nick Chater was funded by ESRC research grant C00428622023. Mike Oaksford was funded under ESRC Research Contract R000231282. Requests for reprints should be sent to Mike Oaksford, now at University College of North Wales, Department of Psychology, Bangor, Gwynedd LL57 2DG, Wales, UK.

References

Baddeley, A. D. (1976). *The Psychology of Memory*, New York: Basic Books.

Bartlett, F. C. (1932). *Remembering: a Study in Experimental and Social Psychology*. Cambridge: Cambridge University Press.

Bobrow, D. G. and Norman, D. A. (1975). Some principles of memory schemata. In D. G. Bobrow and A. M. Collins (eds) *Representation and Understanding: Studies in Cognitive Science*. N. Y.: Academic Press.

Bower, G. H. (1970). Analysis of a mnemonic device. *American Scientist*, 58, 496–510.

Bransford, J. D., Barclay, J. R. and Franks, J. J. (1972). Sentence memory: a constructive versus interpretive approach. *Cognitive Psychology*, 3, 193–209.

Bransford, J. D. and Johnson, M. (1972). Contextual prerequisites for understanding: some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, 11, 717–726.

Bransford, J. D. and Johnson, M. K. (1973). Considerations of some problems of comprehension. In W. G. Chase (ed.) *Visual Information Processing*, pp. 389–392. New York: Academic Press.

Bransford, J. D. and McCarrell, N. S. (1975). A sketch of a cognitive approach to comprehension: some thoughts about what it means to comprehend. In W. B. Weimar and D. S. Palermo (eds) *Cognition and Symbolic Processes*, pp. 189–229. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Chater, N. and Oaksford, M. R. (in press). Autonomy, implementation and cognitive architecture: a reply to Fodor and Pylyshyn. *Cognition*.

Cheng, P. W. and Holyoak, K. J. (1985). Pragmatic reasoning schema. *Cognitive Psychology*, 17, 391–416.

Cheng, P. W., Holyoak, K. J., Nisbett, R. E. and Oliver, L. M. (1986). Pragmatic versus syntactic approaches in training deductive reasoning. *Cognitive Psychology*, 18, 293–328.

Clark, H. H. and Haviland, S. E. (1977). Comprehension and the given-new contract. In Freedle, R. O. (ed.) *Discourse Production and Comprehension*, Volume 1, pp. 1–40. Norwood, N.J.: Ablex.

Collins, A. M. and Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behaviour*, 8, 240–247.

Craik, F. I. M. and Lockhart, R. S. (1972). Levels of processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, 11, 671–684.

de Kleer, J. (1986). Extending the ATMS. *Artificial Intelligence*, 28, 163–196.

Evans, J. S. B. T. (1982). *The Psychology of Deductive Reasoning*. London: Routledge & Kegan Paul.

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

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

Fodor, J. A. (1980). Methodological solipsism considered as a research strategy in cognitive psychology. *Behavioural and Brain Sciences*, 3, 63–109.

Fodor, J. A. (1983). *Modularity of Mind*. Cambridge, Mass.: MIT Press.

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

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

Griggs, R. A. and Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology*, 73, 407–420.

Harman, G. (1986). *Change in View*. Cambridge, Mass.: MIT Press.

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. Chapter 6 in *Parallel Models of Associative Memory*. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Hinton, G. E. (1987). Learning distributed representations of concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, 1987.

Hinton, G. E., McClelland, J. L. and Rumelhart, D. E. (1986). Distributed Representations. In *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, Volume 1: *Foundations*. Cambridge, Mass.: MIT Press.

Israel, D. J. (1980). What's wrong with non-monotonic logic? In *Proceedings of AAAI–80*, 1980, pp. 99–101.

Johnson-Laird, P. N. and Wason, P. C. (1970). A theoretical analysis of insight into a reasoning task. *Cognitive Psychology*, 1, 134–148.

Johnson-Laird, P. N. and Steedman, M. J. (1978). The psychology of syllogisms. *Cognitive Psychology*, 10, 64–99.

Johnson-Laird, P. N. (1986). Reasoning without logic. Chapter 1 in T. Myers, K. Brown and B. McGonigle (eds) *Reasoning and Discourse Processes*, pp. 13–51. London: Academic Press.

Kuhn, T. (1970). *The Structure of Scientific Revolutions*, 2nd edition, Chicago, Illinois: The University of Chicago Press.

Lakatos, I. (1970). Falsification and the methodology of research programmes. In I. Lakatos and A. Musgrave (eds) *Criticism and the Growth of Knowledge*, pp. 91–196. Cambridge: Cambridge University Press.

Lindsay, P. H. and Norman, D. A. (1977). *Human Information Processing*, 2nd Edition. New York: Academic Press.

Lorayne, H. and Lucas, J. (1974). *The Memory Book*. New York: Stein and Day.

Luria, A. R. (1976). *Cognitive Development*. Cambridge, Mass.: Harvard University Press.

Masterman, M. (1970). The nature of a paradigm. In Lakatos, I. and Musgrave, A. (eds) *Criticism and the Growth of Knowledge*. Cambridge: Cambridge University Press.

McCarthy, J. (1980). Circumscription: a form of non-monotonic reasoning. *Artificial Intelligence*, 13, 27–39.

McClelland, J. L., Rumelhart, D. E. and Hinton, G. E. (1986). The appeal of parallel distributed processing. In *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, Volume 1: *Foundations*. Cambridge, Mass.: MIT Press.

McDermott, D. (1986). A Critique of Pure Reason. Technical Report, Department of Computer Science, Yale University, June, 1986.

Miller, G. A. and Selfridge, J. A. (1950). Verbal context and the recall of meaningful material. *American Journal of Psychology*, 63, 176–185.

Miller, G. A. (1956). The magical number seven plus or minus two, or, some limits on our capacity for processing information. *Psychological Review*, 63, 81–96.

Minsky, M. (1975). Frame-system theory. In R. Schank and B. L. Nash-Webber (eds) *Theoretical Issues in Natural Language Processing*. Cambridge, Mass, June 10–13, 1975.

Murphy, G. L. and Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review*, 92, 289–316.

Norman, D. A. and Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology*, 7, 44–64.

Norman, D. A. and Bobrow, D. G. (1976). On the role of active memory processes in perception and cognition. In C. N. Cofer (ed.) *The Structure of Human Memory*. San Francisco: Freeman.

Norman, D. A. and Bobrow, D. G. (1979). Descriptions: an intermediate stage in memory retrieval. *Cognitive Psychology*, 11, 107–123.

Oaksford, M. R. (1988). Cognition and Inquiry: The Pragmatics of Conditional Reasoning. PhD Thesis, Centre for Cognitive Science, University of Edinburgh.

Oaksford, M. R. and Stenning, K. (1988). Process and Pragmatics in Reasoning with Conditionals containing Negated Constituents. Presented to the International Conference on Thinking, University of Aberdeen, August, 1988.

Peterson, L. R. and Peterson, M. (1959). Short-term retention of individual items. *Journal of Experimental Psychology*, 58, 193–198.

Pollard, P. and Evans, J. S. B. T. (1981). The effects of prior belief in reasoning: an associational interpretation. *British Journal of Psychology*, 72, 73–82.

Putnam, H. (1974). The 'corroboration' of theories. In P. A. Schilpp (ed.) *The Philosophy of Karl Popper*, Volume I, pp. 221–240. La Salle, Illinois: The Open Court Publishing Company.

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: a critique of mental imagery. *Psychological Bulletin*, 80, 1–24.

Pylyshyn, Z. W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science*. Montgomery, Vermont: Bradford.

Reiter, R. (1985). On reasoning by default. In R. Brachman and H. Levesque (eds) *Readings in Knowledge Representation*. Los Altos, California: Morgan Kaufman.

Rumelhart, D. E., Smolensky, P., McClelland, J. L. and Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. Chapter 14 in J. L. McClelland, and D. E. Rumelhart (eds) *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, Volume 2: *Psychological and Biological Processes*, pp. 7–57. Cambridge, Mass.: MIT Press.

Shepard, R. N. (1967). Recognition memory for word, sentences and pictures. *Journal of Verbal Learning and Verbal Behavior*, 6, 156–163.

Stenning, K., Shepherd, M. and Levy, J. (1988). On the construction of representations for individuals from descriptions in text. *Language and Cognitive Processes*, 2, 129–164.

Stenning, K. and Levy, J. (1988) Knowledge-rich solutions to the binding problem: a simulation of some human computational mechanisms. *Knowledge Based Systems*, 1, 143–152.

Wason, P. C. (1966) Reasoning. In Foss, B. (ed.) *New Horizons in Psychology*. Harmondsworth, Middlesex: Penguin.

Wason, P. C. and Johnson-Laird, P. N. (1972). *Psychology of Reasoning: Structure and Content*. Cambridge, Mass.: Harvard University Press.

Wason, P. C. and Shapiro, D. (1971). Natural and contrived experience in a reasoning problem. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 23, 63–71.

Young, M. N. and Gibson, W. B. (1962). *How to Develop an Exceptional Memory*. Chilton: Radnor, Pa.

Correspondence and offprint requests to: Mike Oaksford, University College of North Wales, Department of Psychology, Bangor, Gwynedd LL57 2DG, Wales, UK