Connectionism, Generalization and Propositional Attitudes:

A Catalogue of Challenging Issues[*]

John A. Barnden

Computing Research Laboratory & Computer Science Dept
New Mexico State University
Box 30001/3CRL
Las Cruces, NM 88003-0001
(505) 646-6235 jbarnden@nmsu.edu

Running Head: Connectionism, Generalization and Propositional Attitudes

[*] This work was supported in part by grant AFOSR-88-0215 from the Air Force Office of Scientific Research and grant NAGW-1592 under the Innovative Research Program of the NASA Office of Space Science and Applications.

1. INTRODUCTION

1.1: Refining the Debate

My goal is to bring up some neglected issues involved in the business of applying connectionism (mainly of the distributed variety) to the high-level cognitive tasks that symbolic artificial intelligence (AI) has been applied to, such as reasoning and natural language processing. Most of the points I raise present something of a problem to (certain styles of) connectionism, although one or two are about opportunities for connectionism to contribute something distinctive to the achievement of those tasks.

Therefore, I present neither a condemnation of connectionism nor a championing of it. Connectionists may well be able at some stage to produce systems that overcome the problems I present. My claim is merely that the problems are important, that their solution may require mechanisms beyond those that are currently being entertained, and that there is more to certain central notions such as *systematicity* and *structure sensitivity* (explained below) than is currently being given credit for in the connectionist literature. On the other hand, connectionists may well encounter difficulties in trying to exploit the opportunities I mention. Thus, both the problems and opportunities are potentially rich issues for further research. It would be a mistake to think that the lack of a championing of either symbolicism or connectionism in this chapter means that the discussion does not advance the symbolicism/connectionism debate. On the contrary: in seeking to achieve genuine advances in the debate it is valuable to look carefully at the true variety, subtlety and complexity of the issues involved, rather than to assume that the primary goal is to prove that connectionism is better than symbolicism or the other way round.

As I aim only at cataloguing and clarifying some issues, I do not present much in the way of detailed suggestions about how to overcome/exploit the problems/opportunities, let alone present a specific system. This is despite my own work on developing a structured connectionist system aimed precisely at "closing the gap" [Barnden 1988, 1989c; Barnden 1991; Barnden & Srinivas, in press]. Although I will make occasional reference to this system, the article is not aimed at establishing that Conposit has special leverage on the problems and opportunities (which is not to say that I think it does not!).

1.2: Generality, Systematicity and Structure-Sensitivity

The issues I will be discussing are to do in one way or another with *generalizations.* One of the central advantages claimed for connectionism is that it supports automatic generalization, by virtue of activation pattern similarity effects, away from the specific situations it has seen in the past. I do not impugn this claim, but present some types of generalization that are not catered for in the current connectionist literature and that present interesting challenges. These types are summarized in section 1.5.

The generalization issues are bound up with the notions of *systematicity* and *structure-sensitivity*, which are of great topical importance, notably in regard to the gap between symbolicism and connectionism. Fodor & Pylyshyn (1988: p.46ff) cast systematicity of *inference* as a matter of a system being able to do all inferences of a particular logical type, not just some of them (unless the deficiencies are just a matter of resource bounds being exceeded). A system would fail to be systematic if, for instance, it could infer from `P ∧ Q ∧ R` to `P` but could not infer from `P ∧ Q` to `P`, or vice versa. In fact, it is reasonable and useful to extend the notion to apply to inferences that are not just a matter of content-free logical deduction. For instance, the inference that someone can hear from the premise that he or she can speak is systematic if it applies to anyone under any description. Systematicity of inference is actually just a special case of the more general notion of systematicity of *processing.* This more general notion is implicit in Fodor & Pylyshyn (1988: especially pp.37ff).

Systematicity of *representation* [Fodor & Pylyshyn 1988: pp.39–41] is, broadly speaking, the ability of a system just to be able to *represent* proposition P given (a) that it can represent proposition Q and (b) that P and Q are sufficiently similar in some defined sense. For example, under systematicity of representation we would require that if the system can represent *John loves Mary* then it should also be able to represent *Mary loves John.* A more complex example is that if the system can represent *John loves Mary* and can also represent definite descriptions like *the man with the red hat and a long nose* then it should automatically be equipped to represent *The man with the red hat and long nose loves Mary.*

Both systematicity of inference (or processing in general) and systematicity of representation are qualities of *generality* in a system. In learning systems they can also be qualities of *generalization.*

Structure-sensitivity is the ability of a system to process representations in accordance with the structure of the items they represent. For instance, the conversion of an internal representation of an active sentence into the corresponding passive sentence is an operation that must be sensitive to the structure of the active sentence. The types of systematicity currently being focused on in the high-level connectionist literature generally rest on structure-sensitivity, as they are to do with the representation and inference of items with significant internal structure.

1.3: Propositional Attitudes

A special purpose of the chapter is to point out how various issues that are more or less strongly bound up with *propositional attitudes* — beliefs, intentions, hopes, desires and the like — pose problems for connectionism. I must declare a special interest here. In previous research I have been much concerned with propositional attitudes [see, e.g., Barnden 1983, 1986, 1989a,b; Ballim, Wilks & Barnden, 1990, 1991]. However, until now I have not brought this work into any significant contact with my concern with high-level connectionism. The interaction has led to the issues discussed in this chapter. They are general issues extending beyond the particular concerns of propositional attitudes, but they take on important special forms in the propositional attitude arena, and some of my arguments appeal to special points about propositional attitudes.

The main focus in propositional attitude research is on the semantics, and use in inference, of natural-language *attitude reports.* A typical example of such a report is the *belief report*

John believes that all of Susan's brothers are stupid.

The portions of this article that address propositional attitude issues are concerned primarily with how a cognitive agent X is to interpret incoming attitude reports and to represent and reason about other agents' mental states. The article does not address the formal semantics of attitude reports or the philosophical question of what attitudes are.

The propositional attitude area is one of the most central and troublesome *within* the traditional symbolic paradigm. Propositional attitudes are central in the semantic/pragmatic theory of natural language — see Barwise & Perry (1983), Cresswell (1985), Fauconnier (1985) and Mates (1950) for just a few pieces of evidence for this. As for their importance for AI and connectionism (localist or distributed), it is obvious that propositional attitudes are important in *mundane* multi-agent interaction scenarios, and that they are often explicitly reported in *mundane* natural language discourse. They are also held to play a central implicit role in the communicative acts effected by natural language discourse: see, e.g., Allen (1983), Cohen & Levesque (1985), Grosz & Sidner (1986), Sperber & Wilson (1986), and Wilks & Bien (1983).

Given the central importance of propositional attitudes and natural language reports of them, it should be an important objective for high-level connectionism to represent them and reason about them. This is especially so because of the high degree of subtlety and intricateness that has characterized technical attempts, within the traditional symbolic sectors of philosophy, linguistics and AI, to address propositional attitude representation and reasoning. See Barwise & Perry (1983), Creary (1979), Cresswell (1985), Hobbs (1985), Maida (1988), Perlis (1985), and Zalta (1988) for a small sample of complex representational/notational proposals.

It is surprising, therefore, that so far the symbolicism/connectionism debate has given very little detailed, technical consideration to the question of how to represent or reason with propositional attitudes. See, however, Gasser (1989) for a connectionist system that deals to some extent with the goals, intentions and purposes of various classes of agent; and see Rey (1988) for some claims that propositional attitudes present problems for connectionism. Connectionism and similar techniques have been used to handle belief maintenance (truth maintenance, belief revision) and degrees of belief [see, e.g., Craddock & Browse 1986, Pearl 1986]. However, such work has been concerned with maintenance/degrees of belief within only the artificial cognitive agent X that one is designing, rather than X's interpretation of belief reports about *other* agents.

1.4: Beliefs Eliminated?

Connectionism has also been used by eliminative materialists in philosophy as a basis for claims that the notions of belief and so on, which are often dubbed as folk-psychological, should not be part of a mature philosophy and scientific psychology, and may even disappear from our language and thought ultimately [see, e.g., Churchland 1986]. However, this point, even if justified, has very little impact on the concerns in the present article. Consider a cognitive agent X (person, artificial system, or whatever). Let us grant for the sake of argument that it is not philosophically/scientifically accurate or useful to characterize X as operating on the basis of beliefs and so on. However, X is still faced with interpreting sentences uttered by other agents, and these utterances are replete with explicit mention of agents' attitudes. Furthermore, a human speaker S of such an utterance has folk-psychological views of the agents Y she or he is talking about; and it is important for X to be able to think in S's terms about those Ys in order to achieve a coherent understanding of what S is saying. This is so no matter how philosophically/scientifically ill-conceived S's folk-psychological views are.[1] Therefore, the scientifically-respectable characterization of an agent's mental states, be that agent X, S or a Y, has little if anything to do with how X itself should represent *its* interpretations of the utterances of S about the mental states of Ys. Those utterances do not, even implicitly, have anything to with scientifically-respectable characterizations of mental states, except *perhaps* if S happens to be a thorough-going eliminative materialist. And, as emphasized above, it is a cognitive agent X's interpretations of an S's utterances about mental states that are the main target of the parts of this article that are addressed specifically at propositional attitudes.

At risk of belaboring the point, we should note that the argument in the previous paragraph has a more mundane analogy that may make matters clearer. If X had to interpret an S's utterances about the tourist attractions of New Mexico, it would be entirely inappropriate for X to interpret a statement S makes about desert thunderstorms in terms of some scientifically accurate theory of thunderstorms. Such a theory would be totally irrelevant to X's understanding of what S is saying, except in the special case of X taking S to be knowledgeable about the theory.

1.5: Plan of the Chapter

Section 2 argues that connectionist systems must be able to construct *explicit generalizations*, not just to generalize their responses, which is what current connectionist learning theories are aimed at. The argument rests in part on propositional attitude considerations.

Sections 3, 4 and 5 highlight three forms of generalization that have been given inadequate attention in connectionism. The first, discussed in Section 3, is rapid generalization from recent examples, by analogical matching. This rests on an ability to find matches between two or more

[1] This point is at the heart of the metaphor-based treatment of propositional attitudes in Barnden (1989b, 1990), since most folk-psychological ways of looking at mental states are infused with metaphor.

complex *short-term* representations.[2] This contrasts with the implicit matching to *long-term* information structures in most localist or distributed connectionist studies.

Section 4 discusses the second type of troublesome generalization. This is generalization of a connectionist system's representation and reasoning to cope with *anomalous combinations* of concepts, departing markedly from combinations it has been trained on. Propositional attitude contexts are among the more important types of context in which anomalous combinations arise.

Section 5 points out that reasoning must often be *embedded* within one or another type of context. A primary special case is that of reasoning within the context of a given agent's beliefs. Embedded structure-sensitive processing presents problems for the so-called *non-concatenative* representations used in some connectionist systems. The embedding of a particular sort of reasoning can be regarded as a result of *generalizing* the processing involved.

Section 6 concludes by summarizing the various issues raised.

2. EXPLICIT GENERALIZATION

The issue I address in this section is that current work on generalization and learning within connectionist systems is concerned almost exclusively with *implicit* generalizations, in a sense to be made clear. In contrast, I will argue for the need for connectionist systems to construct *explicit* generalizations. The argument depends partly on an appeal to propositional attitude representation.

2.1: Implicit and Explicit Generalization

Imagine a typical sort of generalizing connectionist system, such as a feed-forward back-propagation network, trained to output the conclusion "atheist" on examples of specific communists living in the town. Let us suppose that it succeeds in generalizing to all or most communists living in the town. Then it can make the right inference about each individual member of that class: when given a description of some individual member, we can view the system as coming out with an activity pattern representing an assertion that that person is an atheist. This assertion can then be passed to other processes; in particular, it might cause the system to output the sentence *That person is an atheist.* However, what the system has *not* done is to construct (or develop the means of constructing) an activity pattern representing the *generalization* (that all or most communists living in the town are atheists) as such.

[2] A *short-term representation* is a representation sitting in the short-term memory of some cognitive agent (whether a person or an AI system). Typically the representation will only be *temporary* in that it will only recently have been created and will be lost or destroyed after a short time. A short-term representation might, for instance, encode the agent's interpretation of some sentence that it has just heard, or a conclusion recently inferred from other short-term or long-term knowledge.

In short, a typical generalizing connectionist system is able to cope with specific new cases individually, but is not able to come up with an explicit generalization.

Why should it be desirable for systems to construct explicit generalizations? One reason is that surely one can imagine wanting to ask the trained system what it knows about communists, and expecting it to answer that all [or most] communists living in the town are atheists if in fact this is a generalization that has been trained into it. But also, and perhaps more importantly, the explicit generalization might be needed in order to feed into other inference processes. For instance, suppose the system is in some sense able to apply, or behave as if it is applying, the rule

> *If all [or most] communists living in a town are atheists then the town is eligible for a grant from the Fundie Fund.*

The crucial point about this rule is that the system must recognize that there is a generalization over a class; mere conclusions that particular members of the class are atheists are simply not enough for the rule to be applied. If, on the other hand, as a result of its training the system were able to produce an activity pattern representing the explicit generalization in question, then some subnetwork that embodies the rule could take that generalization as input.

However, we must be careful to avoid a hasty conclusion that the explicit generalizations are really needed even in the presence of the displayed rule. Suppose, as a result of training, some module G of the system maps the communist/living-in-the-town combination to an atheism assertion. More precisely, suppose that there is a *communist* feature unit, a *living in the town* feature unit, and an *atheist* feature unit, among many others. (This localist encoding assumption is made simply for ease of illustration.) G includes those three specific feature units, and is such that *yes* activation on the first two (with *don't-care* activation on all others) causes *yes* activation to appear on the *atheist* unit. The grant eligibility rule could then be respected in roughly the following way. The system would try to satisfy the condition of the rule by deliberately putting the *yes* value at the communist and living-in-the-town units — even though the system is not currently perceiving or otherwise considering any specific communist, so in effect the system is *deliberately imagining* an indefinite communist-living-in-the-town. Then, if the atheism conclusion pops out, the system causes the eligible-for-grant conclusion to appear. Thus, G has been used in place of an explicit generalization.[3]

Lest we appear to have saved the day for the typical connectionist system, note carefully that we have assumed several abilities that go beyond what is typically entertained for connectionist systems:

— The ability to deliberately imagine an indefinite individual (as opposed to merely processing an individual handed to it as input).

[3] Note however that the representation is less powerful than a typical symbolic representation would be: without further elaboration the connectionist system is not able to infer that someone is not a communist living in the town if he or she is not an atheist. However, this lack will not be exploited in our argument.

— A procedural control capability for manipulating G in the appropriate way as a a sort subroutine in the course of doing the grant-eligibility inference.

In sum, getting a connectionist system to use a rule like the one discussed, whose condition part appeals to a generalization, requires either (i) the production of an activity pattern A explicitly representing the generalization in question or (ii) the use of control facilities that go beyond those put forward in most connectionist proposals. If (i) is the case, and the system has learned the generalization from training, then that learning must have the effect not only of enabling the system to conclude the atheism of any individual communist living in the own that is handed to it, but also of constructing activity pattern A (or of developing the means of generating A on demand).

2.2 General Sentences and Beliefs

In fact, there is an obvious consideration that advances the case for option (i). The consideration is that the system needs also to be able to understand general natural language sentences such as *All the communists in the town are atheists,* and to feed the fruits of such understanding through, say, the grant-eligibility rule (even though the system itself has not been trained to believe that all communists in the town are atheists). It is natural to suggest that the system constructs an explicit generalization from the sentence. It is much more difficult now to suggest that this construction could be avoided by the G-based method. Even assuming that would be simple and quick to construct or develop G as a direct result of processing the sentence, we have the point that G is not something that the system necessarily wants to maintain outside of the context of the sentence. It may well be that the system should merely take the sentence to convey something the speaker believes, rather than to convey a truth. G must not be allowed to infect the system's own reasoning about the world, even momentarily; that is, we have to devise machinery for stopping the system using G to conclude for itself that the town is eligible for the grant. Also, it is very possible that there is no need for the system to remember later on that the speaker believes that all the communists in the town are atheists. So, G would have to dismantled in some way (unless we were prepared to tolerate the system developing an indefinite number of useless modules like G).

In sum, a generalization communicated in natural language to the system may well need to be recorded as merely being a speaker belief, and therefore barred for affecting the system's reasoning from its own beliefs, and, moreover, may well not need to be given a long-term representation by the system. Both of these points militate against reliance on a G module and in favor of production of an activity pattern representing the generalization. If this is accepted, then for the sake of uniformity in the system, if the system *learns* the generalization from individual instances it should also develop (a means of generating on demand) an activity pattern that represents the generalization. That is, option (i) at the end of section 3.1 is to be preferred over option (ii). This requires a major extension of the capabilities of current connectionist learning systems.

2.3 Vague Quantification

A benefit from connectionism is that it emphasizes the importance of *incomplete* generalizations — using *most* rather than *all* — and the practicality of reasoning under their influence. This is because of the inherent fuzziness of many types of connectionist systems. A typical generalizing connectionist system need never get to the stage where literally all communists living in the town are inferred to be atheists. Rather, exceptional features, or lack of enough input features, could lead to the inference not being made. The ability to handle exceptions to otherwise emergent rules gracefully is one claimed strength of connectionism [cf. Rumelhart & McClelland 1986]. Of course, much work has gone into default rules and exceptions to them in the plausible-reasoning area of traditional AI, and most AI people probably subscribe to the view that quasi-universals are more important for most of AI than are true universals. But at least connectionism reinforces this view and gives it some alternative flesh.

Nevertheless, the current ability of connectionist systems to embody incomplete generalizations does not go very far towards coping with the general issue of incomplete generalization and other forms of vague quantification (manifested by words such as *several* and *few* in English). Consider the following sentence, spoken by John to the system:

Most communists living in the town are atheists.

Let us presume that the system decides to record this as a belief of John's, not as a new fact in the system's own long-term knowledge. Even if the G-module method had been suitable for complete generalizations, it would in any case be inappropriate here. Consider what a G-module would have to be like. It would have to produce less certain atheism conclusions than was the case in section 3.2. This would presumably mean either that atheism conclusions would only be produced stochastically, with high probability less than 1, on the setting of the communist and living-in-the-town feature units to *yes,* or that an atheism conclusion is always produced but the *atheist* unit has a degraded activation value. But in either case a *particular* number (a probability or degraded activation value) would have to be adopted. Such a number might be forthcoming in the case of a *system* belief that most communists living in the town are atheists, particularly if that belief arose from training on individual communists living in the town. The number could reflect the proportion of the training set that are atheists. However, it is inappropriate in the case of the system representing *John's* belief in that generalization. This is for two reasons.

First, we cannot assume (without considerable extra argument and evidence) that people's incomplete generalizations are encoded with the aid of any sort of numerical measure of majority. After all, it seems reasonable to entertain the possibility that a person's belief in *most X are Y* is encoded in that person's mind by means of a representation that uses a vague-quantification device analogous to the English word *most.*

Secondly, even if the connectionist system could assume that John is indeed using a numerical measure of majority in his belief, it is illegitimate (without extra argument and evidence) for us to allow the system to assume a particular value for that measure. It appears very likely that different people have different numerical criteria for using the word *most.*

The sentences we have looked at in this subsection have not been attitude reports, although the implicit layer of speaker belief has been important to our argument. Of course, the considerations arise even more centrally in the case of attitude reports, such as *Mike believes that most communists living in the town are atheists.* As an aside, it is worth noting that the interpretation of vague quantification in belief reports has been largely pushed aside in propositional attitude research. This is especially surprising in view of the fact that the strict quantifiers have been very extensively studied in that research, and yet are far less important for practical, and even philosophical, purposes. When one studies belief reports one should be concerned, *even as a philosopher*, with the sorts of belief people really have, not with over-idealized sorts of belief.[4]

3. SHORT-TERM STRUCTURE MATCHING

This section deals with the first of three types of generalization that a cognitive system needs to be able to perform but that have not been shown to be well catered for by current connectionist systems, whether localist or distributed. The type considered in this section is to do with *analogies* between two or more novel propositions conveyed by discourse. This brings in the problem of *short-term structure matching*, which has not been studied enough within either localist or distributed connectionism.

The issues also argue for an extension of the usual notion of *systematicity of inference* [Fodor & Pylyshyn 1988: p.46ff]. As treated in the connectionist literature, systematicity of inference and the closely related issue of structure-sensitivity have exclusively been a *long-term* matter in a certain sense. For instance, the inference from speaking to hearing is naturally taken to be a matter of long-term knowledge. Hence, it is natural in connectionism to assume that the systematicity is immanent in some weight settings that implicitly capture the system's long-term knowledge, whether this results from training or from the hand-coding of rules. In the latter case, it is assumed that systematicity is to do with similarity of current inputs to large numbers of past training instances.

However, there is a type of generalization involving systematicity and structure-sensitivity that is to do with *short-term similarity* — similarity between two individual inputs arriving close together in time — and that involves *on-the-fly associations* as opposed to learned associations. Suppose the system inputs

Micky hates his sister for being taller. He's always being mean to her.

Then suppose that soon afterwards the system inputs

Billy hates his brother for being taller.

Surely it is reasonable to expect the system to infer at least tentatively that Billy is always being mean to his brother.[5] And, this should happen *even if* the system has never before encountered

[4] Of course, this concern takes us into the field of psychology, and a consequent psychologicalization of propositional attitude research. See Barnden (1989a,b) for further discussion of the need for psychologicalization.

[5] This can all be wrapped within a speaker belief context, but we omit this for simplicity.

the idea of someone being mean to someone as a result of hating the latter, and even if the system is too ignorant to come up with an explanation of the hater's behavior.

Part of issue being addressed is the question of how to establish an approximate structural match between *two short-term* representations, namely the representation of Micky hating his sister and the representation of Billy hating his brother. This question has largely been ignored in connectionism because of the concentration on using a short-term structure to retrieve approximately-matching structures from *long*-term memory.[6] This focus exists even in work on applying connectionism to high-level cognitive processing. Indeed, virtually the whole field of analogical processing, connectionist or not, seems to be focused on matching target structures to source structures retrieved from *long-term* memory.

Nevertheless, besides the argument from sibling hatred, there are strong reasons for wanting short-term matching (short-term similarity). We may take three well-known types of example from the pragmatics of natural language understanding.

Consider the reasoning involved in tying together what is being said by two natural language sentences that are saying the same thing at different levels of detail, as in:

Go along this street for a while. Go for three blocks and turn right.

This example, adapted from one in Hobbs (1985), requires an understanding that just one going-along is being advocated — the speaker is not telling the listener to go along the street for a while and *then* go for three blocks. The listener must, among other things, detect the approximate match between going along the street for a while and going along the street for three blocks.

For the second example, consider the task of understanding the sentence:

Just as John hates his brother for being taller, Sally hates her sister for being thinner.

The system should presumably represent the fact that a correspondence is being set up between John and Sally, between brother-of and sister-of, and between taller and thinner. Thirdly, syntactic and semantic parallelism between two nearby parts of a discourse has also been exploited in the resolution of pronoun references [Carbonell & Brown 1988].

Of course, what might in principle be immanent in a connectionist system's weights is a *mechanism* for assessing the similarity of two working-memory structures, and such a mechanism may conceivably be learnable. However, the difficulty lies not just in the learning, but also in how to do the structure comparison at all. I have been attacking the short-term-structure-comparison problem in extending the above-mentioned Conposit system to a case-based reasoning or analogy-based reasoning version [Barnden & Srinivas, in press].

[6] The studies of Bienenstock & von der Malsburg (1987) and Gasser & Smith (1991) are exceptions.

In summary, the claim is that the notion of systematicity should be broadened to include reasoning by analogy, even when the analogy holds between two short-term structures rather than involving analogical sources in long-term memory.

4. ANOMALOUS COMBINATIONS

This section addresses another type of potentially troublesome generalization. The issue concerns *anomalous combinations of ideas* in sentences. The issues raised connect with the notions in Fodor & Pylyshyn (1988) of *systematicity of representation* [pp.39–41] and *systematicity of inference* [pp.46ff]. The discussion also briefly touches on metaphor.

The reader should bear in mind that I am not seeking to show that the mentioned type of connectionist system will definitely not be able to handle the type of generalization in question. Rather, I try only to show that the requirements of systematicity are more onerous than is made apparent by the connectionist literature. I will be touching upon one or two suggestions as to how the connectionist systems in question *might* in the *future* discharge the onus.

4.1 Speaking Bananas: An Anomalous Combination

The fundamental problem to be addressed here is the sheer arbitrariness and novelty of the way concepts can be combined in natural language sentences. Consider for instance the clause

the banana can speak.

Ridiculous at first sight, it could well appear as a sentence in a joke or a children's story (or some other type of fantasy), and be a literal truth within the framework of the joke or story. Variants of it could also be intended as metaphorical statements, a matter I go into later on in this section. The anomalous clause *The banana asked for a glass of water* could be taken metonymically in, say, a restaurant context, as conveying that *the person who ordered* the banana asked for a glass of water. A counterfactual can contain anomalous combinations — consider for instance the sentence, *If bananas could speak, peaches probably could too.* Negation provides another sort of context for anomalous combinations: indeed, the sentence *Bananas cannot speak* or *It is not the case that bananas can speak* is weird but literally true.

For my own purposes the most important way the clause *the banana can speak* could appear is as a description of someone's belief. This could arise because of an explicit attitude report, such as

Micky believes that the banana can speak.

This sentence could be true because of Micky being, say, a tiny tot or a grown-up lunatic. But the utterance of merely *The banana can speak* by a speaker S is *implicitly* embedded in an attitude context. The simplest possibility is for the hearer, H, to take the utterance to be a sincere act of assertion and infer that *S believes that* the banana can speak. For illustrative clarity, however, I will use belief reports like the one just displayed and ignore the speaker.

We have thus mentioned seven types of context in which an anomalous combination of ideas such as *banana* and *speaking* could plausibly appear — jokes, children's stories, metaphor, metonymy, counterfactuals, negation, and attitude contexts. We can sum up by means of the aphorism that *it is not anomalous for mundane discourse to contain anomalous combinations of ideas.*

Before going on I should stress that anomaly is a matter of degree, and that I am concerned not just with extremely anomalous combinations, as in the banana example, but also with less extreme examples. An extreme example is used for the sake of illustrative vividness.

4.2 Systematicity of Processing

The problem I wish to present is that anomalous idea-combinations may cause a connectionist system to fail to perform systematic structure-sensitive processing that we would surely want it to perform. This problem may arise in a class of connectionist systems that has been recently advanced as a way of performing structure-sensitive inferencing. The systems supposedly learn to perform structure-sensitive processing of statements[7] by virtue of being trained on input/output pairs that exemplify the required processing. The class includes the systems of Blank, Meeden & Marshall (this volume) and Chalmers (1990). To fix ideas, let us consider the system presented in Chalmers (1990). It consists of two three-layer backpropagation subnetworks:

(1) a Recursive Auto-Associative Memory (RAAM) network [Pollack 1988, 1990];

(2) a transformation network.

The RAAM takes what I will call an "external" encoding of three-word active-voice sentences, such as *John loves Michael*, on the bottom layer, produces an "internal" distributed encoding on the middle (hidden) layer, and can furthermore decode the latter encoding into (a close approximation to) the original external encoding on the top layer. The representation vectors on the bottom and top layers are split up into three segments, one for each word; but the representation vector on the hidden layer is non-concatenative (monolithic or holistic) in that it cannot be broken down into parts corresponding to the parts of the sentence. Also, by an iterative use of the same network, passive-voice sentences like *Michael is loved by John* can be given an internal encoding on the middle layer. Notice carefully that the RAAM is a backpropagation network, and is *trained* to be able to do the external-to-internal encoding and internal-to-external decoding on a specific corpus of sentences.

The transformation network is *trained* to take *internal* encodings of active sentences and produce *internal* encodings of the corresponding passive sentences. These internal encodings are taken from the hidden layer of the RAAM. The interesting observation Chalmers makes is that,

[7] For simplicity, we use the term "statement" to cover both propositions and natural-language sentences.

even when the transformation network is trained on only half of the sentence corpus used to train the RAAM network, the transformation network generalizes correctly to the remaining half.

The claim is that that the transformation network learns to be sensitive to the structure of sentences, without needing to decode the internal representations into representations of the parts of the sentences. Thus, the processing in the transformation network is "holistic" or "direct" — it does not need to proceed via the *external* representations of the sentences.

My worry is that networks of this sort may not generalize to sentences whose concept-combinations are sufficiently different from the combinations appearing in the training corpus. Notice that in the Chalmers experiment mentioned above the training corpus was a randomly selected half of all the sentences in the RAAM's training set. What is of more interest, and of direct relevance to the speaking-banana problem, is the transformation net's performance on sentences *different* from those in the RAAM's training set. In fact, one of Chalmers's experiments involved training the RAAM encoder/decoder only on a random 40 of the 125 possible active sentences and the corresponding 40 of the possible passive sentences. The 40 active/passive pairs in the RAAM training set were used to train the transformation net as well. What happened was that the RAAM correctly generalized to *only* about 84% of a test set consisting of a further 40 active and 40 passive sentences [Chalmers 1990, p.58]. Chalmers claims that this generalization rate is "remarkably good," but it is not clear what this value judgment is based on, given that the system is only a toy one, no evidence is given that the generalization rate will not drop lower in a bigger system, and no evidence is given that an 84% ability to encode novel propositions is adequate for cognition.[8] Moreover, and not surprisingly, the imperfections in the RAAM encoding led to errors in the transformation net, which only generalized to 65% of the test set. (The figure went up to about 70% in a different training regime for the transformation net.) In the experiment discussed before, the RAAM was trained on *all* the possible sentences and the transformation net did generalize perfectly. But the significance of this is unclear, in that in general we surely want cognitive systems to be able to perform structure sensitive inference on structures other than those the system has specifically been trained to represent.

In sum, Chalmers' results, though certainly interesting and giving us cause for *hope*, are hardly sufficient to justify a firm conclusion at the present time that his style of network is powerful enough to cope with the full problem of systematic structure sensitivity. In particular, his results do not provide any particular reason to think that his sort of system will cope well with the speaking banana issue.

We should also note that his results do not cover the case where the agent or patient phrase of a sentence is a complex description such as *the man with the red trousers and blue hat.* With regard to this sort of example, the structure sensitivity and systematicity we want is coupled with an ability to *ignore* structure — that is, to be able to handle the agent and patient phrases without any regard whatsoever to *their* internal structure.

[8] I am presuming this familiarity from the fact that the selection of 40 random sentences out of the 125 active ones is very likely to involve the five possible agents, five possible actions and five possible patients several times each.

Blank, Meeden & Marshall (this volume) present a very similar approach for applying simple transformations to statements. In one of their systems, the transformation network was trained to directly convert internal representations of statements of the form *X chase Y* to the internal representations of corresponding statements of the form *Y flee X.* Training on 16 chase statements allowed it to generalize moderately well to 8 chase statements not in that training corpus. (The transformation is reported as being successful 87.5% of the time on the latter 8 statements. Presumably this means it performed correctly on 7 of the statements.) Another of the Blank *and al.* systems is meant to detect "reflexive" statements like *junglebeast smell junglebeast.* However, the finding was that the net failed to see that statements of the form *tarzan V tarzan* and *jane V jane* were reflexive, even when V was a verb that appeared in reflexive statements in the training corpus. The authors suggest that the reason for the failure is that neither *tarzan* nor *jane* appeared reflexively in the statements in the training corpus. Indeed, the authors say: "... it seems that the network wasn't easily finding a strong relationship between words that it hadn't seen together before..." In short, a double appearance of *tarzan* in a statement was an *anomalous combination* as far as the net was concerned, and caused it to fail.

To return to the example of speaking bananas, suppose a connectionist system of the sort in question has been trained to make the plausible inference that if someone can speak then they can also understand language. That is, the proposition *X can speak* for many different people X has been paired with *X can understand language* in the training set. The danger is that the internal representation of *The banana can speak* may be sufficiently different from that of *X can speak* for any person X that the desired output, *The banana can understand language,* may simply not be generated (or may only be generated with very low confidence, or with a greatly corrupted activity pattern, etc.). After all, the representations of all the Xs may involve high activity of a *person* feature or pattern, which would be lacking in the representation of *the banana.* Now, in some systems the representation of the individual people, actions, etc. themselves would be learned during the course of processing statements [Lee, Flowers & Dyer 1989; Miikkulainen & Dyer 1989]. Then, entities get similar representations to the extent that they appear in similar contexts in statements. With the the types of training ordinarily used, banana representations and people representations are not likely to end up being very similar. Again, in systems where the entity representations are fixed in advance by the system designer, the representations are likely to be (micro)feature based: then, if bananas and persons do not share many features, their representations may be insufficiently similar for the desired generalization to bananas to occur. Of course, bananas and people *are* similar to the extent of being organic physical objects of a medium sort of size. Hence, we should remember that instead of bananas in our examples we could use types of entity that were even more distant from people — molecules, clouds, the wind, for instance.

4.3 Systematicity of Representation

Quite apart from systematicity of inference, an even more radical difficulty raised by speaking bananas is systematicity of *representation* [Fodor & Pylyshyn 1988: pp.39–41]. It needs to be shown that connectionist systems that learn a mapping from an external representation to an internal representation exhibit sufficient systematicity of *representation*, let alone systematicity of inference. (The systems at issue now include simple recurrent sentence-processing networks such as those of Elman 1988, 1989 and Harris & Elman 1989 as well as the RAAM-based systems mentioned above.) The crucial point, already implied above, is that the *mapping from external to internal representations rests on similarity to training instances.* Thus, we are led to wonder whether the statement *the banana can speak* can even be *represented* internally in an adequate manner (given training only on people as speakers), let alone subjected to inferencing.

The potential connectionist difficulty with systematicity of representation is an important point, because it holds good even if one were to deny the desirability of the above type of systematicity of *inference* (X-can-speak ⟶ X-can-understand language). The system should at least be able to respond to the sentence *Bananas can speak* with something like *But that's stupid — bananas can't speak.* The point about this is that the ability to form the response on the basis of the input sentence itself requires significant systematicity of representation (and processing). The sentence has to be noticed as being stupid and the negation of it has to be formed; moreover, the fact that the proposition that bananas can speak is (let us assume) a belief of the speaker's has to be represented. We should demand that the system be able to make responses like *But that's stupid — bananas can't speak even if* we allowed the system to get away with not being able to perform deeper types of processing, such as inferring that Micky probably also believes that bananas can understand language.

A similar question of systematicity of representation is also brought up by the sentence *It is not the case that bananas can speak.* It seems reasonable to require the system to be able to respond to this sentence appropriately, even when there has so far been no mention of bananas or speaking in the current discourse. Even if the reasonable response is *I knew that!* the system must have some way of representing the content of the sentence, and part of this task is to deal with the combination of the banana idea with the speaking idea.

To sum up, a system should be able to represent, notice and analyze weirdness, not pretend it isn't there.

A symbolic system would at least not have any serious problem on sheer representation, since it would have the ability to form essentially arbitrary combinations of symbols. Of course, it might be that the system, when it encounters *bananas can speak* for the first time, embodies a selectional restriction to the effect that only people are allowed to be put in the agent position of a can-speak predication inside the system, whether that predication is couched as a logic expression, semantic network fragment, frame instance, or whatever. However, this prohibition would reside in some particular rule or data structure, which could now be explicitly noticed, and then thrown away or weakened.

We see here an advantage of the symbolic system's *separation of representation of structure from representation of components*: arbitrary combinations can be formed because that formation process is (or can easily be made to be) independent of the nature of the items combined. Thus, for example, the conventional data structuring techniques of pointers, sequential allocation and associative addressing can be used to link together any sort of information one likes. This is in contrast to the relevant connectionist systems, where the encoding of associations among parts of a statement is inextricably linked with the encoding of what is associated.

An advantage of connectionist systems that "merely" implement standard symbolic processing is of course that they maintain the separation of structure from component. A pure example of such a system is provided by the Conposit system [Barnden 1988, 1989c; Barnden 1991]. Here there is a working memory consisting of a two-dimensional array of (active) registers into which any symbols (from the set of symbols embodied in the system) can be placed. This arbitrariness of symbols sitting in registers at any given moment is parallel to the arbitrariness of bit-strings in computer memory cells, and confers upon Conposit an immunity to the speaking-banana problem.

4.4 Some Possible Retorts

Here I look at some suggestions for how the speaking-banana challenge might be met by the type of connectionist system under discussion. The suggestions are all of some interest and promise, but would require a substantial amount of work to be turned into convincing demonstrations that the problem can be avoided.

(1) One might try to deal with the speaking-banana problem by suggesting a dual system in which there was a similarity-based connectionist subsystem of the sort portrayed above, for coping with non-anomalous cases, and another system that coped with anomalous cases and worked on different principles. One might claim that the latter worked in a more cautious mode, and perhaps would be more likely to involve conscious deliberation and puzzlement. However, I maintain that this suggestion should be put aside until other alternatives have been investigated and rejected, so as to avoid getting into the position of having to duplicate the large parts of the inference-making apparatus as well.

(2) It is conceivable that if the connectionist system were trained with examples of the form *Xs can speak* where the Xs included a wide variety of things (people, cars, walls, clouds, ...), but *excluded* the particular case of bananas, the system might learn to ignore totally the specific nature of the X in the way it forms its internal representations and processes them. It would then be able to deal with the X = banana case even though bananas were very dissimilar to *all* the training Xs. Similarly, one might wish to claim, if one is interested in connectionist models of mind, that children do get enough of the right sort of practice with strange combinations of concepts. However, such a claim would need to be backed up by firm empirical evidence.

(3) It has been suggested that even if the author of the connectionist system were to grant that it should be able to deal with speaking bananas, there is no reason for that author to agree that it should be able to handle much more abstract agents of speaking, as in *Monday speaks.* (Assume for the sake of argument that it is known that *Monday* is the day name, not a person's name.) Thus, there are reasonable limits to the semantic anomaly of combination that we should expect a system to handle. However, notice that a joke or riddle, say, might start as follows:

Monday said to Tuesday, 'I'm more important than you because I'm earlier in the week.'

Notice that this immediately leads to presumptions in the hearer that Tuesday can hear, can understand language, and can speak. We would not be *at all* surprised, for instance, to hear now that Tuesday replied, *No, it is I who is more important because ...* The point is that despite the (putatively) much greater anomaly of *Monday said* as compared to *The banana said* the possible and desirable inferences are very much the same in both cases — and, incidentally, very much the same as when a *person* is reported as saying something. That is, there is *a priori* evidence to suppose that the degree of plausibility/desirability of inferences made by humans is not well correlated with the degree of anomaly in the statements on which the inferences are based.

(4) The Monday example also shows that the speaking-banana problem cannot be dismissed by claiming that bananas and people *are* sufficiently similar, despite my comments above, for the connectionist systems to be able to generalize sufficiently well to speaking bananas. For instance, Jay McClelland [personal communication] has made a claim on the following lines. Bananas and people, both being organic physical objects of a medium sort of size, enter into enough similar interactions (e.g., both can be kicked) for a connectionist system of *realistic* size to have similar enough representations of bananas and people to be able to cope with speaking bananas. I do not deny that this *may* be so — but it is something that needs to be established, not merely wished. And, even if it turns out to be true, it does not readily extend to the Monday example.

(5) McClelland [personal communication] has also suggested that a connectionist system with a very rapid learning algorithm could train its RAAM or whatever *on the spot* to handle new combinations of concepts. This is an interesting possibility, but it is not clear that the system would not *also* have to apply more training to its other subnetworks (such as the transformation net in the Chalmers case). It is then not clear how the system is to know which particular subnetworks to do this to.

(6) One might claim that no learning system, connectionist or otherwise, human or otherwise, could be expected to make the desired inferences in the banana case, for example the inference that the banana can think and hear, if it had *only* ever seen the person cases. Certainly, the problem I am pointing out does generalize to some extent beyond connectionism. However, it is rather easier to see how a non-connectionist system using standard symbolic representations could make the generalization to bananas even though only trained on human speakers. Suppose a symbolic system has constructed the following explicit rules as a result of its training:

(A) `if X is a person and X can speak then X can understand language`

(B) `if X is not a person then X cannot speak`

(C) `if X is not a person then X cannot understand language`

Let us suppose that the system is able to tolerate a merely-partial match of a rule condition part. Then, because of rule (A), it is plausible to suggest the system could come out with *the banana can understand language* with some moderate degree of confidence, when presented with *the banana can speak.* Rule (B) would apply but could be discounted in the present reasoning context as a result of producing a conclusion that contradicts the premise that the banana can speak. Now, it is reasonable to suppose that a rule with the same condition part as an already-discounted rule itself becomes discounted or at least suspect. This is on the heuristic that if one predication about a certain type of situation turns out to be faulty, other predictions about that type of situation also become faulty. Under this assumption, rule (C) becomes suspect, and is therefore not able to defeat the tentative conclusion from rule (A).

The central point is that the system can easily be made to tolerate a merely-partial match of a compound condition part, such as that of rule (A). On the other hand, connectionist systems of the sort in question have internal representations that do not manifest the syntax of the represented statements. That is, those activity patterns do not have a structure of subpatterns corresponding to the structure of the statement. (See also van Gelder 1989, 1990 on this *non-concatenative* type of system.) The representing of structure is only implicit in some way in an internal representation, so that the representation of the structure is *intimately mixed in with* the representation of the individual parts. Therefore, it is not easy to see how the connectionist system could emulate the symbolic system's partial-match reasoning without also generating nonsensical inferences. Certainly, it might be that the internal representation of *the banana can speak* is sufficiently similar to internal representations of statements like *X is a person and X can speak* for the system to be able to produce the desired conclusion, *the banana can understand language.* However, if the notion of "sufficiently similar" is loose enough for *that* to be the case, might it not be that the internal representation of some totally irrelevant statement like *Blurg is a horse and Blurg can neigh* also just happened to be similar enough to *X is a person and X can speak* for the conclusion *Blurg can understand language* to pop out? That is: a representation scheme in which *structural* aspects of a statement are intimately mixed in with the representation of the *components* of a statement makes for difficulty in providing a notion of *sufficiently similar* between two statements S and T that simultaneously satisfies the following requirements:

(i) it is allowable for S to match only a part of T (or vice versa)

(ii) but the parts of S and T that do match must match pretty exactly.

By contrast, a standard symbolic representation separates structure from components and therefore has no problem in simultaneously satisfying the requirements.

A final retort is discussed in the next subsection.

4.5 Personification and Metaphor

Elman [personal communication] has suggested that the system should *personify* Monday or the banana in our examples above. But what might this involve? Confining attention for brevity to the Monday example, I mention three representative possibilities (though no doubt there are others).

(a) Monday and Tuesday are viewed *metaphorically* as people. The metaphor is cashed out in terms of a traditional analogical mapping between part of the days-of-the-week domain and part of the human-being domain. The individual item correspondences involved in this mapping are then treated as connectionistically implemented bindings.

(b) Monday and Tuesday are again viewed *metaphorically* as people. However, no bindings are involved. Rather — and this is Elman's version of the suggestion[9] — the connectionist representation of Monday is "deformed" by being made more like the representation of a person. (In the simplest case this might involve turning on the *person* feature unit or, say, *has-mind* and *has-mouth* feature units. These units would of course normally be off in the representation of a day.) The representation of speaking might also be deformed temporarily to make it more like something that could be attributed to bananas. For instance, speaking could be relieved of the normal attribute of being produced by an animal mouth.

(c) A further suggestion that has been made to me is that the system forms a representation of a person wearing a placard on which the name *Monday* is written, and similarly another person with *Tuesday.* These people are the participants in the discourse reported in the example.

Possibility (a) is circular because it assumes the ability to represent bindings between arbitrary types of things — that is, it assumes that arbitrary combinations of concepts can be formed. This is the problem we started with. Possibility (c) appears viable, though of course it implies a major amount of machinery over and above the sort typically entertained in connectionist systems currently. An specific worry about the method is that it may go *too* far in the personification, in that we may need specific properties of *days* still to be operative in the example. The carry over of day properties to the two imagined people is not a necessary consequence of (c) as stated, so that extra assumptions must be made. One would be to regard the two people metaphorically as days, but then we are back with some variant of (a) or (b).

Possibility (b) may be viable, and is certainly worth investigating further. Indeed, Elman (1988) reports a recurrent connectionist network which forms distributed representations of simple sentences on its hidden layer and which in a sense performs deformations of the sort under discussion. More precisely, the system was trained on a certain corpus of sentences involving strong constraints on verb-noun relationships. Then, the weights were frozen and the corpus was fed through a final time with the modification that the totally novel pseudo-word *zog* was substituted for every occurrence of the word *man.* Under a certain way Elman presents of estimating the prototypical representations of individual words on the hidden layer, it turned out that the representation of *zog* bore the same hierarchical clustering relationship to the other words as the word *man* did. (This does not quite amount to saying, though, that the *zog* representation was closely similar to the *man* representation.) Thus, there is a sense in which the expectations of the system for seeing *man* in certain contexts cause *zog* to be treated much as if it had been *man.* This effect is certainly promising from the point of view of this section, but one should realize that there

[9] Cf. also the cup-of-coffee example in Smolensky (1988).

is a lot left to be demonstrated. It must be shown that the internal representations of the *zog* sentences are actually similar enough to the corresponding *man* sentences to be able to support similar inferences.

Possibility (b) is interesting not least because it is an important opportunity for connectionism to contribute something distinctive to high level cognition. We need to exercise caution about it, though. I conjecture that when we interpret the sentence *Micky believes that bananas can speak* we do *not* necessarily commit ourselves to making proper sense of the complement (the clause following the *believes that.*) That is, all we do, unless we decide to think about the sentence carefully, is to assume that *there exists a way in which* Micky sees bananas as able to speak, without committing ourselves to any *particular* way, and therefore without needing to effect the distortions just discussed. It is conceivable, then, that what we need in a good connectionist model of human cognition, and also perhaps in a good connectionist artificial intelligence, is a way for the system to turn off the mutual sensitivity of concepts that are being combined. That is, the mutual sensitivity had better not be *too* automatic. If this is so, then the system should be able to form a straight combination of incompatible concepts, a trivial task for a traditional symbolic system.

Here we see as way in which hearers' processing of just *Bananas can speak* might differ, at least in degree, from their processing of that same anomalous combination when it is embedded in an attitude report, such as *Micky believes that bananas can speak* or *Susan wants Micky to believe that bananas can speak.* The more explicit layers of attitude context there are, the less the pressure to make full sense out of the idea of bananas speaking. Unless the hearer has some particular motive for thinking about Micky's belief state or Susan's desire state, the hearer can refrain from delving into the idea of bananas speaking. There could indeed be such a motive: one that is independent of the discourse, or one that is introduced by the exigencies of coherently linking the sentence to other sentences in the discourse. Notice, however, that it might be that the speaker's only intention in reporting Micky's belief was to emphasize the weirdness of Micky's beliefs, in which case no deep understanding of the belief would be necessary. Of course, the attitude-free sentence *Bananas can speak* is embedded in an implicit speaker-attitude context, but in this case it is more likely that the statement will link up in some meaningful way with other speaker statements. The speaker might go on to say, *So I'll get that banana over there to tell my wife I've gone to El Paso.* The hearer would need to reason that the speaker also believes that bananas can understand language and have memories.

A further complication is that variants of the banana example can have mundane metaphorical interpretations. For instance, the sentence *The bananas spoke to John of luxury* could mean that the presence of the bananas in a certain household indicated to John that the household was a luxurious one, in a context in which bananas were luxury items. To worse the complication, however, one might claim that metonymy is involved here rather than metaphor. The use of *spoke* could be viewed as a metonymic way of referring to the normal *result* of speaking, namely the implanting of information in the mind of the hearer. But if metaphor is indeed what is involved, it might be seen as requiring a deformation of the concept of banana and/or the concept of speaking. These deformations would be favorably looked upon in an interactionist theory of metaphor [see Waggoner 1990 for a review].

5. EMBEDDED STRUCTURE-SENSITIVE PROCESSING

The idea of non-concatenative *versus* concatenative representations was used to in the previous section. A non-concatenative representation, like that in the middle layer of a RAAM, is in a sense holistic, not having any natural structural similarity to whatever is represented. Rather, all we assume is that there are well-defined mechanisms *of some sort* for transforming a non-concatenative representation of an item X into representations of the parts of X, and for creating the representation of X out of the representations of its parts. Let us call these two types of transformation the *analysis* transformations and the *synthesis* transformations respectively. In the non-concatenative case, these transformations are not just a matter of breaking the X representation down into its parts or of bunching the representations of X's parts together, respectively.

A large part of the current interest in the internal representations in systems like those addressed in the previous section is based on their being not only (i) non-concatenative but also (ii) susceptible to being processed "directly", that is, without first being unpacked into the corresponding "external" representations by means of the analysis transformations. With regard to (ii), if the internal representations could be used in inference or other processing *only* indirectly, then they would lose much of their point, even in the eyes of their proponents. And, it is certainly of great interest and importance to see whether useful forms of processing can be done directly on the non-concatenative representations.

However, there are various modes of reasoning that are not readily catered for by direct processing of non-concatenative representations. These modes are various forms of *embedded* reasoning. A particular form of embedded reasoning is important in the propositional attitude arena. I will discuss that type first, and then briefly mention others.

In reasoning about the beliefs of another agent a system must often make inferences embedded within the context of that agent's beliefs, so to speak. Thus, if John believes that Sally is a chess master and that all chess-masters are clever, the system may need to make the (plausible) inference that John has concluded (and therefore believes) that Sally is clever. Now, such inferences about John on the system's part could be performed by means of reasoning schemata of the following form:

$$\frac{\begin{array}{l}\text{A believes that X is a P}\\ \text{A believes that all P are Q}\end{array}}{\text{A believes that X is Q}}$$

However, this involves a style of reasoning where the *A believes that* layer is explicitly carried through all steps, resulting in extra complexity in the necessary matching — for example, in the matching of the premise *John believes that Sally is a chess-champion* to the first line of the displayed schema. Therefore, propositional attitude theorists have often proposed [e.g. Creary 1979, Haas 1986] that "simulative reasoning" be used. In our example, this would involve steps such as the following. First, the system records the fact that it is reasoning within the context of John's beliefs. Next, it perceives that the statements *Sally is a chess-champion* and *All chess-champions are clever* match the premises of the schema

$$\begin{array}{l}\text{X is a P}\\ \text{All P are Q}\\ \hline \text{X is Q}\end{array}$$

The system then draws the conclusion that Sally is clever. It now backs up out of the John belief environment, and is able to report that *John believes that Sally is clever.* (The backing-up can be accompanied by a reduction of confidence in the conclusion, given that people do not always draw conclusions from their beliefs.)

The point, of course, is that this simulative reasoning procedure involves simpler reasoning schemata and simpler matching. Another important point is that the X/P/Q reasoning schema used is the very same one that the system would itself use to conclude that Sally was clever if it believed her to be a chess-champion and believed all chess-champions to be clever. There is no need for the system to have the more complex schema displayed earlier. The extra simplicity is bought at the minor expense of the system's having to keep track of the context it is reasoning within, and to strip off and restore *John believes that* layers.

Now consider a system that has a non-concatenative internal representation I1 of *John believes that Sally is a chess-champion* and a non-concatenative internal representation I2 of *John believes that all chess-champions are clever.* One could make a variety of suggestions as to how the internal representation I3 of the (plausible) conclusion that *John believes that Sally is clever* could be produced:

(a) The system produces I3 directly from I1 and I2 purely by virtue of having had the training sufficient to enable it to produce the conclusion *Sally is clever* from *Sally is a chess-champion* and *All chess-champions are clever.* That is, there is automatic generalization to the case of that inference being embedded within a belief context.

(b) The system has had training as in (a), but goes through the simulative reasoning procedure described above: it uses an analysis transformation to produce internal representations for *Sally is a chess-champion* and *All chess-champions are clever* from I1 and I2 respectively, produces *Sally is clever* from these new internal representations, by generalization from its training, and then uses a synthesis transformation to construct I3.

(c) The system has been trained on structurally similar examples of embedded inference (e.g. on the inference from *Mike believes that Susan is a musician* and *Mike believes that all musicians are refined* to *Mike believes that Susan is refined*), and this is sufficient for generalization to occur to our John/Sally example.

Now, method (c) is undesirable because it introduces a scale up problem. Not only must analogous training be done in the case of (some) other types of attitude, rather than belief, but it must also be done for non-attitude forms of embedded reasoning. Perhaps worse, separate training would have to be done for each extra layer of attitude that is added. Thus, it is reasonable to draw the plausible conclusion that *John intends George to believe that Sally is clever* from *John intends George to believe that Sally is a chess-champion* and *John intends George to believe that all chess-champions are clever.*

It would be very nice if (a) were feasible. But, until it (and the corresponding approaches to other types of embedded reasoning problem, below) are shown to be feasible, we cannot rely on it. The problem increases with the depth of embedding. Thus, a non-concatenative internal representation of, say, *John intends George to believe that Sally is a chess-champion* is only going to be weakly similar as an activation pattern to a non-concatenative internal representation of *Sally is a chess-champion.* Therefore, the chances of direct processing of the latter representation generalizing to correct direct processing of the former are questionable.

The remaining method, (b), is antithetical to the spirit of non-concatenative representations, as was implied above. One might as well use a concatenative representation in the first place (or at least a hybrid representation, whereby, in representing *A believes that Z*, the combination of A, belief and Z is done concatenatively but Z is represented non-concatenatively).

Within-attitude reasoning is an especially important form of embedded reasoning, but there are others. For instance, counterfactual reasoning raises similar issues. In making the argument *If Sally were a chess-champion she would be clever* (using a background assumption that chess-champions are clever) we need to reason within a counterfactual context in which Sally is a chess-champion. (See Fauconnier 1985 and Dinsmore 1987 for a general framework in which belief spaces, counterfactual spaces, fictional spaces and several other types of space are unified.) Another important case of embedded reasoning arises from the observation that, in ordinary propositional logic, if S is a subformula of F and S is logically equivalent to a formula T, then F is logically equivalent to the formula G obtained from it by replacing some or all occurrences of S in F by T. For example, F and G might be

$$\mathtt{p} \longrightarrow (\mathtt{q} \wedge (\mathtt{u} \longrightarrow \mathtt{v}))$$

$$\mathtt{p} \longrightarrow (\mathtt{q} \wedge (\neg\mathtt{v} \longrightarrow \neg\mathtt{u}))$$

respectively, with S being $\mathtt{u} \longrightarrow \mathtt{v}$ and T being $\neg\mathtt{v} \longrightarrow \neg\mathtt{u}$. Notice that S can be be a *small* and/or *deeply embedded* component of F, making approaches similar to (a) and (c) even more questionable. Further, the structure of non-S parts of F is completely arbitrary.

The Conposit implementational-connectionist system [Barnden 1988, 1989c, 1991] uses concatenative representations throughout, and could therefore cope with embedded reasoning in a relatively straightforward way. For instance, the John/Sally belief context example would be handled in outline as follows. Each statement of the form *A believes that Z* is broken up into two sub-representations within Conposit's working memory: one part which can be thought of as *A believes that x* for a variable x, the other part being Z together with a binding of it to x. Reasoning can proceed directly on the Zs; and this does not even require the Z parts to be broken out first, as they are already separate from the *A believes that x* parts. Thus, simulative reasoning presents no special problems. (And it is straightforward to keep track of the contexts.)

One moral of this section is that *structure-sensitivity* [Fodor & Pylyshyn 1988] of processing, which is a primary challenge to systems using non-concatenative representations, is not just a matter of the *whole* structure of a represented item. Thus, with reference to the propositional logic example just above, the structure-sensitivity involved in mapping from S ($\mathtt{u} \longrightarrow \mathtt{v}$) to T ($\neg\mathtt{v} \longrightarrow \neg\mathtt{u}$) is more than a matter of transforming S as a free-standing structure to T. Rather, the structure-sensitivity involves the ability to map any F containing S *as a part* to the corresponding G. Similar observations hold for other sorts of embedding, such as belief contexts. To go back to Chalmers' sentence passivization network (see section 4.2), what is needed is for the system to be able to do not just passivization but also embedded passivization, such as transforming *John believes that Michael loves Bill* to *John believes that Bill is loved by Michael,* and so on at deeper levels of embedding.

Hence, direct processing of non-concatenative representations of structured objects is at a severe disadvantage. The particular object part that is to be subjected to structure-sensitive processing on a given occasion is intimately mixed in with the other parts, requiring the direct processing mechanism to somehow preserve those other parts while manipulating the part in question. The corresponding preservation in the case of concatenative representation is trivial.

7. CONCLUSION

We have looked at various challenging issues to do with getting connectionism to cope with high-level cognitive activities such a reasoning and natural language understanding. The issues are to do with various facets of generalization that are not commonly noted. We have been concerned in particular with the special forms these issues take in the arena of propositional attitude processing. The main problems we have looked at are:

(1) The need to construct explicit representations of generalizations, not just generalize correctly to individual cases.

(2) The need to be able to match two or more complex short-term information structures, to enable rapid generalization from recent examples rather than from long-term memories.

(3) The need to represent and reason with anomalous combinations of concepts.

(4) The need to perform embedded reasoning. This presents special problems for systems using non-concatenative representations.

We also touched on vague quantification in attitude report complements. Neither this topic nor that of analogies between short-term structures (point 2) has been adequately addressed in the symbolic framework, let alone in connectionism.

One opportunity we saw for connectionism to contribute something distinctive to the realization of high-level cognition lies in its support for the automatic distortion of the concepts involved in anomalous combinations, including those arising in metaphor, to make them fit together appropriately (assuming the combinations can be dealt with at all).

There are some other connections between connectionism and propositional attitude research that we have not discussed. As noted in section 1, one connection already studied by others is the possibility of connectionism accounting naturally for *degrees* of belief, desire and so on. Another connection is the opportunity for connectionism to provide a way of handling *intermediate* types of attitude. The usual attitude types (belief, intention, hope, etc.) are almost always cast in AI, linguistics and philosophy as being clearly distinguishable from each other. Nevertheless, there is reason to think that the distinctions are fuzzy, and that intermediate types are possible. What I have in mind here is, say, a state intermediate between believing something to be true and wishing it to be (cf. wishful thinking). Another example is provided by the sentence *He half hoped she would come and half dreaded it.*

The opportunities and problems covered are put forward as things worth being optimistic about or pessimistic about, respectively. They are not put forward as decisive arguments for or against connectionism. The hope is that this chapter contributes to a greater understanding of the connectionist/symbolist gap by presenting some unusual issues and by throwing new light on some well known ones.

ACKNOWLEDGMENTS

I am deeply indebted to John Dinsmore for creating the opportunity for me to develop the lines of thought in this chapter. I thank Jeff Elman, Tom Eskridge, David Farwell, Mike Gasser, Margarita Gonzalez, Rocio Guillen, Cathy Harris, Steve Helmreich, Trent Lange, Jay McClelland, Tim van Gelder, Jin Wang, Fuliang Weng, and Yorick Wilks, as well as John Dinsmore and anonymous reviewers, for useful questions, criticisms and suggestions.

REFERENCES

Allen, J.F. (1983). Recognizing intentions from natural language utterances. In M. Brady & R.C. Berwick (Eds), *Computational Models of Discourse.* Cambridge, MA: MIT Press.

Ballim, A., Wilks, Y. & Barnden, J.A. (1990). Belief ascription, metaphor, and intensional identification. In S.L. Tsohadzidis (Ed.), *Meanings and Prototypes: Studies in Linguistic Categorization.* London and New York: Routledge. pp. 91–131.

Ballim, A., Wilks, Y. & Barnden, J.A. (1991). Belief ascription, metaphor, and intensional identification. *Cognitive Science, 15* (1), pp.133–171.

Barnden, J. A. (1983). Intensions as such: an outline. *Procs. 8th Int. Joint Conf. on Artificial Intelligence,* Karlsruhe, W. Germany.

Barnden, J.A. (1986). Imputations and explications: representational problems in treatments of propositional attitudes. *Cognitive Science, 10* (3), 319–364.

Barnden, J.A. (1988). Conposit, a neural net system for high-level symbolic processing: overview of research and description of register-machine level. *Memoranda in Computer and Cognitive Science*, No. MCCS-88-145, Computing Research Laboratory, New Mexico State University,

Barnden, J.A. (1989a). A misleading problem reduction in belief representation research: some methodological considerations. *J. Experimental and Theoretical Artificial Intelligence, 1* (2), 4–30.

Barnden, J.A. (1989b). Belief, metaphorically speaking. In *Procs. 1st Intl. Conf. on Principles of Knowledge Representation and Reasoning.* San Mateo, CA: Morgan Kaufmann.

Barnden, J.A. (1989c). Neural-net implementation of complex symbol-processing in a mental model approach to syllogistic reasoning. In *Procs. 11th Int. Joint Conf. on Artificial Intelligence.* San Mateo, CA: Morgan Kaufmann.

Barnden, J.A. (1990). Naive Metaphysics: a metaphor-based approach to propositional attitude representation. *Memoranda in Computer and Cognitive Science*, No. MCCS–90–174, Computing Research Laboratory, New Mexico State University, NM 88003, USA.

Barnden, J.A. (1991). Encoding complex symbolic data structures with some unusual connectionist techniques. In J.A. Barnden & J.B. Pollack (Eds.), *Advances in Connectionist and Neural Computation Theory, Vol. 1.* Norwood, N.J.: Ablex Publishing Corp.

Barnden, J.A. & Srinivas, K. (in press). Overcoming rule-based rigidity and connectionist limitations through massively-parallel case-based reasoning. *Int. J. Man-Machine Systems.*

Barwise, J. & Perry, J. (1983). *Situations and attitudes.* Cambridge, Mass.: MIT Press.

Bienenstock, E. & von der Malsburg, C. (1987). A neural network for invariant pattern recognition. *Europhysics Letters, 4*(1), 121–126.

Blank, D.S., Meeden, L.A. & Marshall, J.B. (this volume). Symbolic manipulations via subsymbolic computations.

Carbonell, J.G. & Brown, R.D. (1988). Anaphora resolution: a multi-strategy approach. *Procs. COLING-88.*

Chalmers, D.J. (1990). Syntactic transformations on distributed representations. *Connection Science, 2* (1 & 2), pp.53–62.

Churchland, P.S. (1986). *Neurophilosophy.* Cambridge, MA: MIT Press.

Cohen, P. & Levesque, H. (1985). Speech acts and rationality. In *Procs. Meeting of the Association for Computational Linguistics.*

Craddock, A.J. & Browse, R.A. (1986). Belief maintenance with uncertainty. In *Procs. 8th Conf. of the Cognitive Science Society.* Hillsdale, N.J.: Lawrence Erlbaum.

Creary, L. G. (1979). Propositional attitudes: Fregean representation and simulative reasoning. *Procs. 6th. Int. Joint Conf. on Artificial Intelligence,* Tokyo.

Cresswell, M.J. (1985). *Structured meanings: the semantics of propositional attitudes.* Cambridge, MA: MIT Press.

Dinsmore, J. (1987). Mental spaces from a functional perspective. *Cognitive Science, 11* (1), pp.1–21.

Elman, J.L. (1988). Finding structure in time. Tech. Rep. 8801, Center for Research in Language, University of Califormia, San Diego, CA.

Elman, J.L. (1989). Structured representations and connectionist models. In *Procs. 11th Conf. of the Cognitive Science Society.* Hillsdale, N.J.: Lawrence Erlbaum.

Fauconnier, G. (1985). *Mental spaces: aspects of meaning construction in natural language.* Cambridge, Mass.: MIT Press.

Fodor, J.A. & Pylyshyn, Z.W. (1988). Connectionism and cognitive architecture: a critical analysis. In S. Pinker & J. Mehler (Eds.), *Connections and symbols*, Cambridge, Mass.: MIT Press, and Amsterdam: Elsevier. (Reprinted from *Cognition, 28,* 1988.)

Gasser, M. (1989). Robust lexical selection in parsing and generation. In *Procs. 11th Conf. of the Cognitive Science Society.* Hillsdale, N.J.: Lawrence Erlbaum.

Gasser, M. & Smith, L.B. (1991). Comparison, categorization, and perceptual dimensions: a connectionist model of the development of the notion of sameness. Manuscript, Computer Science Department, Indiana University, Bloomington, IN 47405.

Grosz, B.J. & Sidner, C.L. (1986). Attention, intentions, and the structure of discourse. *Computational Linguistics, 12*(3), 175–204.

Haas, A.R. (1986). A syntactic theory of belief and action. *Artificial Intelligence, 28,* 245–292.

Harris, C.L. & Elman, J.L. (1989). Representing variable information with simple recurrent networks. In *Procs. 11th Conf. of the Cognitive Science Society.* Hillsdale, N.J.: Lawrence Erlbaum.

Hobbs, J.R. (1985). Ontological promiscuity. *Procs. 23rd Ann. Meeting of the Association for Computational Linguistics,* Univ. of Chicago.

Lee, G., Flowers, M. & Dyer, M.G. (1989). A symbolic/connectionist script applier mechanism. In *Procs. 11th Conf. of the Cognitive Science Society.* Hillsdale, N.J.: Lawrence Erlbaum.

Maida, A. S. (1988). A syntactic approach to mental correspondence. *Procs. Conference of the Canadian Society for Computational Studies of Artificial Intelligence,* Edmonton, Alberta.

Mates, B. (1950). Synonymity. *Univ. of California Publications in Philosophy, 25* 201–226. [Reprinted in L. Linsky (ed.), *Semantics and the philosophy of language,* Urbana: U. Illinois Press, 1952.]

Miikkulainen, R. & Dyer, M.G. (1989). Encoding input/output representations in connectionist cognitive systems. In D. Touretzky, G. Hinton & T. Sejnowski (Eds), *Procs. 1988 Connectionist Models Summer School.* San Mateo, CA: Morgan Kaufmann.

Pearl, J. (1986). Fusion, propagation, and structuring in belief networks. *Artificial Intelligence, 29* (3), pp.241–288.

Perlis, D. (1985). Languages with self–reference I: foundations. *Artificial Intelligence, 25* (3), 301–322.

Pollack, J.B. (1988). Recursive auto-associative memory: devising compositional distributed representations. In *Procs. 10th Annual Conf. of the Cognitive Science Soc.* Hillsdale, N.J.: Lawrence Erlbaum.

Pollack, J.B. (1990). Recursive distributed representations. *Artificial Intelligence.*

Rey, G. (1988). Sanity surrounded by madness. *Behavioral and Brain Sciences, 11*, 48-50.

Rumelhart, D.E. & McClelland, J.L. (1986). On learning the past tenses of English verbs. In J.L. McClelland, D.E. Rumelhart and the PDP Research Group, *Parallel Distributed Processing, Vol. 2.* Cambridge, Mass.: MIT Press.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*, 1-74.

Sperber, D. & Wilson, D. (1986). Relevance: communication and cognition. Oxford: Blackwell.

van Gelder, T. (1989). Compositionality and the explanation of cognitive processes. In *Procs. 11th Conf. of the Cognitive Science Society.* Hillsdale, N.J.: Lawrence Erlbaum.

van Gelder, T. (1990). Compositionality: a connectionist variation on a classical theme. *Cognitive Science, 14* (3), pp.355–384.

Waggoner, J.E. (1990). Interaction theories of metaphor: psychological perspectives. *Metaphor and Symbolic Activity, 5* (2), pp.91–108.

Wilks, Y. & Bien, J. (1983). Beliefs, points of view and multiple environments. *Cognitive Science, 7*(2), 95–119.

Zalta, E.N. (1988). *Intensional logic and the metaphysics of intentionality.* Cambridge, Mass.: MIT Press.