Notre Dame Journal of Formal Logic
Volume 27, Number 3, July 1986

Information and Association

J. A. FODOR

1 Here is what happened. I started out to write something about the use of associative networks as models of mental processes, but it kept turning into a paper about what notion of information is the right one to use in cognitive psychology. So I thought: Very well then, I shall write something about what notion of information is the right one to use in cognitive psychology. When I set to work on that, however, it kept turning into a paper about the use of associative networks as models of mental processes.

It began to dawn upon me that perhaps there is some connection between questions about the use of associative networks as models of mental processes and questions about what notion of information is the right one to use in cognitive psychology. This idea rather surprised me since I do not remember ever having heard these two sorts of questions discussed together. (Indeed, the people who have recently had interesting things to say about the first are mostly psychologists and the people who have recently had interesting things to say about the second are mostly philosophers-cum-semanticists. For all I know, cognitive science being what it is, these two groups of researchers have never heard of one another.) So it occurred to me that perhaps I should write a paper about information and association and how views about the one are related to views about the other.

What follows is thus cartography. I want to see how some ideas that are current in cognitive science fit together to comprise a landscape. The main thesis is that two markedly distinguishable prototheories at present occupy the field, and that these express very different—perhaps irreconcilable—construals of the doctrine that minds are information processors. I have, as will become abundantly evident, my preference as between these views, but my present purposes are only partly polemical. I am also interested in getting clear what the options are and how they are assembled from their parts.

By way of prospectus, then, the views I have in mind are typified by the following galaxies of claims:

Received April 16, 1984; revised March 12, 1985

Type 1 theories

mental processes are largely associative
mental computations are "executive free"
mental processing is massively parallel
the information transmitted is the basic notion in cognitive theory
information is 'in the world'
intentionality is a nuisance
the typical explanatory constructs of cognitive psychology are semantic

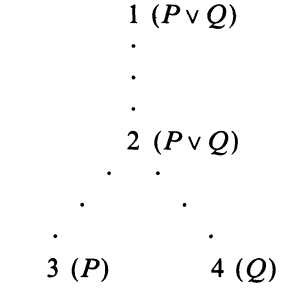
Type 2 theories

mental processes are largely computational
mental processes are executive-driven
mental processing is typically serial
the information encoded is the basic notion in cognitive theory
information is a by-product of representation
intentionality is the key to the mental
the typical explanatory constructs of cognitive psychology are proof-theoretic.

As must be apparent, none of these slogans attains the highest degree of perspicuousness. Nor, I take it, is it transparently obvious why I have them grouped the way they are. Getting all that sorted out will be the burden of the following.

2 Let's begin with the simplest sort of case. Suppose we have a 'Boolean Network' set up as follows. (Information flows 'up'—from higher to lower numbered nodes—unless otherwise specified. It may be assumed, though I shall not generally stress this, that the routes through this network are probabilistic.)

N1:



I have, arbitrarily, labeled the nodes of this network with propositional constants. Propositions can be either true or false, so the network is interpreted as exhibiting relations among the truth values of a certain set of abstract objects. But, in fact, the network might be interpreted in any way at all so long as the kinds of entities (objects, events, whatever) that are assigned to the nodes are (a) capable of assuming (at least) two 'distinguished' states (on and off; T and F; 0 and 1; instantiated and uninstantiated; etc.), and (b) the distinguished states of each of the entities depend, in the ways that the diagram specifies, on the distinguished states of the others.

So, for example, you might interpret N1 as specifying a 'neural network': To effect this interpretation, (a) a collection of neurons is assigned to each node; (b) we distinguish between two states of each of the collections (excitation and quiescence, as it might be); and (c) paths along the network are identified with routes of causation, so that the state of excitation of the objects at node 2 depends, in the ways specified, upon the states of excitation of the objects at nodes 3 and 4.

But, equally, we might imagine that the nodes of N1 represent four *people*, each of whom can be in either of two states (hands up or hands down, for example). There might be a convention that these people follow: person 2 puts his hands up just in case either person 3 does or person 4 does. My point in stressing the plurality of possible interpretations of N1 is to make clear that there is nothing peculiarly *mental* (or even biological, or even physical) about the pattern of dependencies among distinguished states that N1 specifies. More generally, if a notion of the 'information' in a system can be defined in terms of such relations, then that notion will exhibit a desirable sort of ontological neutrality: since it depends only upon *patterns of relations* of distinguished states it is *ipso facto in*dependent of the ontological status of the entites that are in those states. This is important for psychologists because notions like information seem to be crucial in the explication of the mental and we would like our psychology not to assume that the properties of minds are *sui generis*. One way to achieve the *naturalization* of the mental would be to show that, in the sense of information in which minds are information processors, so too are a lot of other things.

Suppose, for example, that there is some sense of information in which information is transmitted from node to node in realizations of network N1. In some (*NB: but not in all*) such realizations, the mechanism for the transmission of that information may itself include the occurrence of mentation. Thus, in the case where the nodes are people who raise and lower their hands, part of the story about why the guy at node 2 raises his hands may involve reference to thoughts he has about what the guys at nodes 3 and 4 are up to. It might be that 2 says to himself: '3 has raised his hands, and the convention is that I raise my hands if 3 or 4 does, so I must raise my hands.' But, though this gives a characteristically mental tone to the account of how information is transmitted through *that* instantiation of N1, the conditions for *being* an instantiation of N1 (hence, *ex hypothesi*, sufficient conditions for being an information processing system) do not depend upon the occurrence of any mental processes. These conditions require (to repeat) only that certain dependencies obtain among the distinguished states of the objects at the nodes *however those dependencies are achieved.*

Well, as everybody knows, there *is* a notion of information—which I will call the *Standard* notion of information—that does satisfy the condition that the transmission of information through a network depends solely on the existence of patterns of relations among distinguished states of the objects that constitute the network. Since everybody knows about the Standard notion of information, I will not bother even to sketch it here. (People who want to see a version of the Standard notion that is worked out with an eye specifically to its application to the mental should read [3]). Suffice it that, according to the Standard

notion of information, x transmits information about y just in case (and, just to the extent that) the distinguished states of x are correlated with the distinguished states of y. Correlation makes information, according to the Standard view.

Corollaries of the Standard view:

1. Information is cheap (because lots of things are correlated).
2. Information is 'in the world' (because lots of things in the world are correlated).
3. If information about the world is 'in the head' (if, for example, the belief that it is raining contains information about whether it is raining) that must be because states of the head are correlated with states of the world.

Well, then, just what notion of an *information processor* does one arrive at if one starts from the Standard notion of information? And is it the notion of information processor that we want for our cognitive psychology? I propose to sneak up on this by slow stages. First, I want to build some intuitions, to suggest that there are aspects of what one might pretheoretically take the informational situation in N1 to be that are *not* reconstructed by the Standard notion of 'information transmitted.' I will then propose some arguments to show that these neglected informational notions are ones we actually need for our account of minds.

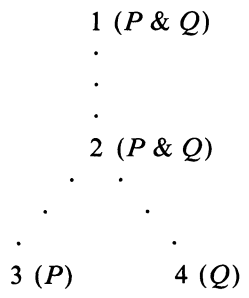
Here is an easy question to begin with: According to the Standard notion, what information is transmitted from 2 to 1 in N1? Standard answer: the information that $P \vee Q$. (If N1 is a neural net, then if 2 is excited, that transmits the information that either 3 is excited or 4 is excited; and if 2 is quiescent, that transmits the information that both 3 and 4 are quiescent.) Why, according to the Standard view, is that the right answer? Because, by assumption, the true description of the state of affairs in N1 is that the distinguished states of 2 are correlated in a certain way with the distinguished states of 3 and 4 (e.g., 2 fires iff 3 fires or 4 fires) and, we are assuming, correlation makes information. Notice that, while this much is uncontentious, that is only because it is stipulative; it follows from the definition of the Standard notion of information together with the specification of N1.

Well, by stipulation then, the activation of 2 transmits information about the activity of 3 and 4; and it does so because the state of activity of 2 depends on the state of activity of 3 and 4. But, of course, the activation of 2 does not transmit the information *that* the state of activity of 2 depends on the states of activity of 3 and 4. That information is not represented *anywhere* in the network (except in the labels; a point that I will return to). A graphic way of seeing this is to imagine yourself 'in' the network at node 1. This is what the rest of the network looks like from that perspective:

2

Node 1 cannot, as it were, 'see' the network beyond node 2. So that, for example, from the point of view of node 1, network N2 looks just like network N1.

N2:



We – standing outside the networks, so to speak – can see what node 1 cannot: that whereas the situation in N1 is that the activation of 2 is effected by the *dis*junction of 3 and 4, the situation in N2 is that the activation of 2 is effected by the *con*junction of 3 and 4. Hence, standing outside the networks, we can see that the information transmitted by the activation of 2 in N1 is different from the information transmitted by the activation of 2 in N2. But node 1 cannot see this; all that node 1 can see, in either network, is node 2 going on and off.

One might put it (as friends of the Standard notion often do) that this conception makes information an "objective" quantity, that it makes information *perspective free* or *observer neutral.* The basic idea is that what information is transmitted through the network to 1 is independent of what 1 can 'see' (a fortiori, it is independent of what, if anything, 1 knows or believes). For, absolutely *all* that matters in determining what information is transmitted through the network is what situation objectively obtains in the network. Similarly, the only thing about 1 that matters in determining whether 1 *receives* the information that 2 transmits is that the designated states of 1 really do correlate with the designated states of 2. The information that 1 receives is not, one might say, in 1. If it's anywhere, it is in the correlation between the designated states of 1 and 2; i.e., it's *in the network.* And if, as might be the case, the network is in the world, then the information is in the world too.

Of course one buys this objectivity at a price. A way to see this is to distinguish between the information *transmitted* by the activation of 2 and the information *displayed* by the activation of 2. (Unlike 'the information transmitted', 'the information displayed' does not, alas, readily admit of quantification. It is not even well defined. I am not, however, building a theory. I am just building intuitions in aid of making clear why the account of information that the Standard notion offers is not the one that our cognitive science needs.) In the present case, then, we can identify the information that is *displayed* at 2 with the information that is visible from 1: i.e., information about the designated states of 2 (information that 2 is active or quiescent if we stick to the neural network interpretation). Notice that access to the information displayed at 2 does not, in and of itself, allow you to distinguish N1 from N2. To tell which network you are in, you need to know not just what information is displayed at 2 but also how the information displayed at 2 covaries with the information displayed at 3 and 4.

The distinction between the information transmitted and the information displayed is easily obscured when you label the diagrams of a Boolean network. For, each label – given its intended interpretation – actually *displays* the information that is transmitted by the activation of the corresponding node. The label on a node says how the activation of that node depends upon the activation of the rest of the network. It is thus of primary importance to understand that, in Boolean networks (and other executive-free information processing systems) *the labels are for us, not for the machine.* This is not a philosophical gloss; it is *strictly* true. There is nothing in the operation of a Boolean network, qua Boolean network, that is sensitive to the character of the labels. Qua Boolean network, all that matters is the connectivity of the nodes and their instantaneous states of excitation. To put it slightly differently, each node can 'see' the states of excitation of its neighbors; but it cannot see their labels.

One reason this is so hard to keep in one's head is that one tends to be misled by features that some *but not all* instantiations of Boolean networks have in common (hence, of course, features that the notion of a Boolean network *per se fails* to reconstruct). So, for instance, it is possible to think of instantiations of a Boolean network where the information at a node *is* displayed as well as transmitted; cases where the labels *do* matter for the operation of the network.

Consider an instantiation of N1 where the nodes are people and they exchange information by displaying flashcards on which formulas ('P', 'Q', and '$P \vee Q$', as it might be) are inscribed. So, when 2 is activated (when guy 2 holds up the flashcard that reads '$P \vee Q$'), he not only transmits, but also displays, the information that $P \vee Q$. When (and only when) an event displays the information that it transmits, we can say that the event *encodes* that information. Intuitively speaking, the fact that the information is encoded, and not just transmitted, has striking consequences. For, though the guy at 1 still cannot see the network beyond 2, *he nevertheless can tell what the network is like beyond 2 assuming that he can 'read' the display.* In short: once we introduce a notion of information encoded (to contrast with the Standard notion of information transmitted) we see that there is room for a corresponding notion of a display being read (to contrast with the Standard notion of information being received). And, just as information can be transmitted even when no information is encoded, so information can be received without any display being read (think of N1 as instantiated in a neural net). All that receiving information requires is correlation between designated states of the receiver and designated states of the source. God knows what reading a display requires; the least you need is access to a code.

I have been nagging about the distinction between the information transmitted and the information displayed, and about the heuristic status of the labels in diagrams of Boolean networks, because it is only when one keeps these points in mind that one sees the profound inappropriateness of the Standard notion of information for reconstructing the cognitive scientist's notion of an information processor. I suppose that the fundamental intuition about information processors is this: they are systems whose behavior in a given situation is determined *by the character of the information that is available to them* in that situation. We want a notion of information that will let us hang on to that intuition, and I am claiming that we cannot get one by identifying the pretheo-

retical notion of the information available with the Standard notion of the information received.

The basic reason is what we have already seen: while the (distinguished) state of a node in a network depends only on the states of the nodes it can 'see' (i.e., the local nodes to which it is connected), the information received by a node depends upon the state of the entire network, 'visible' or otherwise. It is thus perfectly possible to have two networks which display the same information to a given node while transmitting different information to that node; as, indeed, we saw that networks N1 and N2 display the same information to node 1 (viz. information about the state of activation of node 2) although the information that gets transmitted through the networks is that $P \vee Q$ in one case and that $P \& Q$ in the other. Well, to put it in a nutshell, what determines what information is (intuitively) *available* at a position in a network depends on what information is *displayed* at that position, not on what information is *transmitted* at that position.[1]

The best way to see this—this still being all just intuitive and pretheoretical—is by thinking about examples. So, here is Johnny walking down Elm Street past the open windows of his neighbors, through which the morning news is audible. As he passes the window of number 7, he hears the announcer say 'It's raining here'; as he passes the window of number 9 he hears the announcer say 'It's raining here'. Let us suppose, however, that the radio in number 7 is tuned to a station in Chicago and the radio in number 9 is tuned to a station in Tulsa. Then: (a) the information *displayed* is the same at number 7 and at number 9; but (b) the information *transmitted* is different since, presumably, the signal at number 7 is correlated with the weather in Chicago (but not with the weather in Tulsa) and the signal at number 9 is correlated with the weather in Tulsa (but not with the weather in Chicago). Question: what is the information available to Johnny-qua-information-processor, such that his behavior in the current situation is determined by the availability of that information? Answer: surely it is the information displayed, not the information transmitted. Ceteris paribus, it would be a *miracle* if the information *transmitted* determined Johnny's behavior because, to put the point crudely, there is nothing in the situation to tell Johnny what information *is* being transmitted. (Remember: the signal at number 7 transmits the information that it is raining in Chicago. But it does not transmit the information that it transmits the information that it is raining in Chicago. Compare the case where what the announcer says is: 'it's raining here in Chicago'. Here the information that it is raining in Chicago is *encoded* as well as transmitted. Since it is encoded it follows that it is displayed. And since it is displayed, it is available to modulate Johnny's behavior.)

I do not want to say that the information available *is* the information displayed. But I do want to say something like this: the information available is the information displayed plus whatever the receiver can figure out from the display. The information that is in the Standard sense *transmitted* becomes in the pretheoretical sense *available* only when an information processor can infer the information transmitted from the information displayed.[2]

Roughly, there will be two kinds of cases where the information that is transmitted becomes available to a receiver: when the transmitted information

is encoded (hence displayed) and the receiver knows the code; and when the receiver knows how the character of the display depends upon the state of the rest of the network and is thus able to infer from the display what information it transmits. It is terribly important to understand that both these conditions constrain the receiver in ways that merely being a recipient of information transmitted does not; that is, they require more of the receiver than covariance of its designated states with designated states of the source.

THE MORAL: The notion of available information, unlike the Standard notions of information transmitted and information received, is intentional, perspectival and receiver relative. What information is available depends upon one 'objective' factor (viz. what information is on display) and one 'subjective' factor (viz. what the receiver is able to infer.) This is too sad for words, but it is nevertheless true.

Moreover, deep down, everybody knows that it is true. Suppose you have a correlation between *A* and *B* and a correlation between *B* and *C*. Then, to a first approximation, you have a sufficient condition for the transmission of information from *A* to *C* given the Standard notion of information transmission. ('Transmits information to' is approximately transitive since it is just a way of spelling 'is correlated with'. That 'makes information available to' is *not* transitive is a way of putting the point that I have been struggling to make.) If you then ask a friend of the Standard notion how the *mere* existence of two such correlations could, in and of itself, be sufficient for the *availability* of information about *A* at *C*, what he is likely to say is this: Well, what information is *available* at *C* is a matter not just of what information is objectively there, but also of what information *C* is "attuned" to.

For example: "There is a lawlike relation between smoke and fire. Situations where there is smoke are, by and large, close to situations where there is fire. And it is attunement to this relation that enables us to learn about [particular occasions of] fires from [particular occasions of] smoke" ([2], p. 12). It is not, however, clear what this 'attunement' comes to; not even as a metaphor. It is one thing for a device to be 'attuned' to a class of particulars (as, indeed, smoke detectors might be said to be attuned to smoke); presumably, to be attuned to a class of particulars is to be disposed to respond selectively to things that are in that class. But it is quite another matter to make sense of a device to being attuned to a *generalization* (e.g., to the generalization that if there is smoke there is fire). Barwise and Perry do not give us much help in understanding what tuning to a generalization might amount to. Here is what they say: "Being attuned to... [the] relation... [between smoke and fire presupposes] only the ability to detect smoke, to respond in some way appropriate to fire, and to do the latter on the occasion of the former" (p. 12). This, however, is an old and unconvincing story; one which age has not improved (see, for example, [5] and [6]).

To begin with, as people have endlessly pointed out, you cannot rely on formulas like 'a response in some way appropriate to fire' to pick out a class of behaviors since whether behavior *is* appropriate depends not just on the fire, but also on the utilities of the behavior (see, for example, "Norma", act 2, part 2: "Vanne al rogo ed il tuo scempio/Purghi l'ara e lavi il tempio," and so forth).

There is thus more subjectivity—not to say intentionality—built into talk about attunement than the unwary may at first suppose. But pass that; there is worse to come. To respond to *smoke* in a way that is actually appropriate to *fire* would not give evidence of attunement to a relation between them; at best it suggests that you have mistaken the one for the other. And at worst it suggests a sort of craziness since, quite generally, forms of behavior that are 'appropriate to' smoke are ipso facto not appropriate to fire and vice versa. Thus, one says: 'Oh dear, I fear that there may be a fire' in the presence of clear cases of smoke, but *not* in the presence of clear cases of fire; one throws a bucket of water on clear cases of fire but *not* on clear cases of smoke; show me a man who tries to roast his hotdogs in smoke and I will show you a man who ends up with bloodshot eyes and a raw weiner. Etc. You cannot, in short, take this story about 'attunement to relations' literally; and that you cannot is *very* old news.[3]

But probably you are not supposed to. Probably it is a euphemism. What the Real Story is, is something like this: to be attuned to the relation between smoke and fire is to *know that* "situations where there is smoke are, by and large, close to situations where there is fire"; and to use what you know for spotting fires is to come to *expect* fire when you detect smoke. It is behavior appropriate to the *expectation* of fire that you produce if you have detected smoke and are 'attuned to' the fact that smoke means fire. It is, however, not allowed for friends of the Standard notion to tell the Real Story since the Real Story makes it painfully clear that you need intentional apparatus (believing; expecting; Lord knows what all else) to bridge the gap between the information that is in the world and the 'available' information, the information out of which an organism acts.

That philosophers of Barwise and Perry's sophistication should be caught flirting with behaviorism suggests that something has gone very badly wrong indeed.[4] It is thus worth emphasizing that the passage just quoted is not merely a slip of the pen. Here's another one: "The school bell rings and the students learn that it is time for class to end. A certain type of sound, one they hear on different days, is systematically related to a certain type of situation, the end of class. It is this relation between different types of situations that the students become attuned to, and thereby learn that the sound of the bell means that it is time for class to end. Thus the sound of the bell, on any particular occasion, conveys the information about the end of class" ([2], p. 13).

It pays to attend closely to what 'conveys' conveys when it is used in the way that Barwise and Perry use it here. To begin with, on the reading of 'conveys the information' that the authors are most clearly entitled to, the remark that the 'sound of the bell, on any particular occasion, conveys the information about the end of class', though it comes at at the end of the passage, is not the *conclusion* of the argument; it is one of the premises. For, given the 'objective' notion of information that Barwise and Perry adhere to, that the sound of the bell conveys the information about the end of class—i.e., that it Standardly *transmits* that information—is *equivalent to* the assumption that "a certain type of sound... is systematically related to a certain type of situation...." What *does not* follow from this assumed correlation, however, is that the ringing of the bell 'conveys the information' in the pretheoretical sense of 'making the information available' to the students. It is, of course, this second, stronger sense

of 'conveying the information' that we need to explain why the students start to gather their papers together when they hear the bell. To get it, we need to assume the attunement of the students to the information that the ringing of the bell transmits. But what on earth could this attunement come to except that the students have (somehow) learned that when the bell rings the class is over and that, on each occasion when the bell rings they (somehow) use what they have learned to infer that they have come to the end of the class. Notice, once again, that (a) you must have the information available—not just the information transmitted—if you are to predict the behavior that ensues; (b) 'the information transmitted' does not determine the information available; (c) what information is available, unlike what information is transmitted, is receiver relative.

Here, then, is the bad news in brief: you can have an objective notion of information—one that puts the information 'in the world'—but it will not do the work of the pretheoretic notion of 'the information available' to an information processor. Alternatively, you can have a notion of 'the information available,' but it will not be receiver neutral *and it will not be naturalistic either* because it will depend on what the receiver knows and is able to infer. What we do not have—what, for all we now know, we *cannot* have—is a notion of information that is both 'objective' and appropriate to behavioral explanation. Would that this were other, but it is not and loose talk about attunement will not make it go away: the notion of attunement is either blatantly behavioristic (and therefore hopeless) or implicitly intentionalistic (and therefore useless). In short, as things now stand, the notion of information that is required for cognitive science is *nonnaturalistic, unreduced, and intentional through and through.* A fortiori, the 'objective' notion of information does not reconstruct the intentional one. For all we now know, nothing like it ever will.

On the other hand: It is one thing to argue, as I have just finished doing, that the Standard notion of information is not the one you need for cognitive science when the information processor under analysis is a *whole organism* chock full (as whole organisms are wont to be) with beliefs, desires and other such prima facie irreducibly intentional states. It is quite another to argue, as I now propose to do, that the *"subpersonal"* information processors—whose operations, according to psychologists, underlie and account for the cognitive capacities of whole organisms—are not plausibly construed according to the Standard notion either. That the project of reconstructing the intentional apparatus of belief/desire psychology in terms of an 'objective' concept of information should fail is not, perhaps, surprising. Beliefs and desires, after all, *do* seem to be in the head; in the head is *exactly* the right place to cause behavior from. But it might nevertheless be possible to reconcile the objectivity of Standard information with the subjectivity of the propositional attitudes by taking a rather different tack. It might be possible to construe beliefs and desires as 'emergents' out of the activity of subpersonal (presumably neural) information processors, and perhaps the Standard notion of information, with its correlative apparatus of associative networks, will find a home in the analysis of these subpersonal mechanisms. Certainly nothing that has been said so far would deny that this is so.

In fact, the view that mental processes are performed by networks of associated elements is currently quite fashionable in cognitive science. The connection between this idea and the correlational account of information should be clear from the previous discussion. Since the nodes in a network interconnect, activating some of them is causally sufficient for activating others. The pattern of causal determination implicit in the structure of a network in turn implies patterns of correlation among the designated states of its nodes. Correlation makes Standard information, so the flow of activation from node to node can be interpreted as the flow of Standard information through the network. Endless variations of this basic proposal are possible; thank goodness the details do not matter much for what follows.

There are a number of arguments that are supposed to show that associative networks are prima facie plausible candidates for modeling subpersonal cognitive processes. For one thing, because they are massively parallel such networks can be very fast: in principle, the only limiting conditions are the size of the network and the speed with which excitation can be transmitted from node to node. Then again, stress is sometimes placed on the *physiological* plausibility of the view that subpersonal information processes are associative. Neurons, to a first approximation, go on and off; and they do so in consequence of excitation that flows through the brain from one cell to another. So perhaps the brain is a congeries of associative networks. (Much of the discipline known as 'cognitive neuroscience' consists of speculations at about this level of sophistication.)

Slightly more convincing is the a posteriori demonstration that some rather detailed properties of mental processes involving list search can be modeled by associative nets. Lexical access—specifically the perceptual identification of printed letters and words—has provided most of the parade cases. A number of the phenomena of lexical access suggest connections among items in the internal lexicon. The 'priming' of words by their synonyms, the 'word superiority' effect in letter recognition, and the fact that certian sorts of input distortion (such as the misspelling of 'certain' earlier in this sentence) are usually tolerated effortlessly, all suggest that the lexicon is some sort of a network with the nodes connected by both 'vertical' relations (like constituency) and 'horizontal' relations (like synonymy). The basic idea is that if the lexicon is indeed a network, then these lexical access phenomena can perhaps be explained by appeal to the spread of activation horizontally and vertically from node to node. A good deal of detailed attention has been paid this possibility, and it turns out that, with sufficient parametric tinkering, many of the more robust chronometric properties of lexical access can be accommodated (see, for example, [1]). Whether this is because the lexicon really is a network or only because the available models provide enough degrees of freedom to accommodate damn near anything, I leave it to the reader to decide. In what follows I will have nothing at all to say about the use of networks as models of subpersonal processes of list search. I do, however, have a few unsympathetic remarks to make about the use of networks as models of subpersonal *inference*; and these points are quite closely related to ones that turned up earlier in the discussion.

Even quite simple networks, like N1 and N2, can be viewed as devices for making inferences. So, when activation spreads from node 3 or node 4 to node 2 in N1, the network can be thought of as computing the inference from *P* or

Q to $P \vee Q$; when activation spreads from nodes 3 and 4 to node 2 in N2, the network can be thought of as computing the inference from P and Q to $P \& Q$; and when activation in N2 spreads 'downwards' from node 2 to node 3 or 4, the network can be thought of computing the inference from $P \& Q$ to P or Q. Given some ingenuity and enough nodes, *any* inference that is valid in propositional logic can be computed by a network essentially similar to these. So, perhaps one's ability to perform such inferences reduces to the activation of associative networks available at the neural level. This proposal is not irrational, but I believe it to be profoundly misguided.

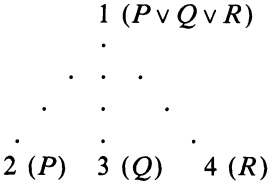
To begin with, though it is true that for any propositional inference there exists a network which computes it, it is also true that there is considerable indeterminacy—there is, if you like, no matter of fact—about which inference a given network is computing. Thus we said that N1 computes the inference from P or Q to $(P \vee Q)$. But it would have been equally right to say that it computes the inference from, say, $(P \vee (R \& -R))$ or Q to $((P \vee (R \& -R)) \vee Q)$; or, indeed, that it computes any inference generable from 'P or $Q \rightarrow P \vee Q$' by substituting logical equivalents in the premise or the conclusion. This is just to say that the only constraint the structure of the network places upon its interpretation is that if activation flows from node $|A|$ to node $|B|$, then the inference from the proposition assigned to A to the proposition assigned to B must be valid. There are, in general, lots of ways (indeed, an infinity of ways) of assigning propositions to nodes consonant with this condition. Of course, the *labels* assigned to the nodes in a network *do* tell us which inferences the network computes. But (have I mentioned this before?) *the labels are for us, not for the machine*; nothing in the behavior of the network qua network depends on how we label it. Or, to put it the other way around, if there is an indeterminacy about what inference a network is computing, then there is, of course, the same indeterminacy about what labels are the right ones for its nodes.

Why does this matter? Well, the proposal under consideration was that our capacity to make propositional inferences reduces to (or is modeled by, or is explained by—talk any way you want to here) the postulated associative networks. But this suggests that, from the psychological point of view, it is all one whether one is inferring from P or Q to $(P \vee Q)$ or from, as it might be $(P \vee (R \& -R))$ or Q to $(P \vee Q)$. Whereas, of course, it is precisely from the psychological point of view that it is *not* all one which inference one is drawing. It is entirely conceivable, for example, that someone for whom the former inference is fast and obvious might find the latter inference obscure and slow. This is, as must be evident, a paradigmatic intentionality problem. Networks do not slice mental states and processes 'thin enough': Whereas networks distinguish inferences up to logical equivalence, it appears that what is in one's head sorts them out with finer grain. (Not surprisingly, precisely the same point holds for 'the information transmitted'. It does not distinguish between logical equivalents either, so that if '. . . transmits the information that P' is true, so too is whatever you get by substituting for P a logically equivalent formula. As Dretske once put it (personal communication), information in the Standard sense is "propositional" rather than "sentential". This is exactly right assuming that logically equivalent sentences express the same proposition. The present point is that mental processes—like drawing inferences—appear to be sentential rather than

propositional. That is bad news for networks and bad news for the Standard notion of information.)

The argument just walked through is precisely the sort that leaves psychologists dry eyed. So here is another, less philosophical sounding but deriving, as we will see, from much the same considerations. Pick an inference that N1 can compute; say P or Q to $(P \vee Q)$. Imagine that Baby has learned to draw that inference, and that his learning to do so reduces to his having, as it were, 'grown' a neural instantiation of N1. Question: what does he have to grow to learn the inference from P or Q or R to $(P \vee Q \vee R)$. Answer: *he has to grow a whole new network.* E.g., N3. And, indeed, this will be true whenever we want to add

N3



a new class of inferences to Baby's repertoire; it will be true even when the new inferences are—intuitively speaking—of the same 'form' as ones that Baby has already mastered. Why is this so? Because the notion of the form of an inference can get no grip in an associative net. But why is *that* so? Because the nodes in a network *have* no form (if you prefer, the network treats all logically equivalent formulas as having the same form). But do not the *labels* have form? Yes, indeed; but (I am *sure* I have mentioned this before) the labels are for us, not for the machine.

The point, then, is that we know about some (indeed, we know about infinitely many) arguments that they are valid in virtue of their form. But networks do not know this. So there is something about our inferences that networks do not reconstruct. Notice that this is *not* just a 'performance/competence' argument. My point is not that, since you need to grow a new network for each new form of valid argument, and since there are infinitely many valid forms of argument, it follows that networks cannot reconstruct our logical competence. That is, I think, quite a good argument, but it too is of a kind that leaves psychologists unmoved (excepting, perhaps, very sophisticated psychologists). 'For,' they say, 'after all, there is nobody who can actually recognize the validity of more than a finite number of arguments. Especially *babies*! Babies *never* recognize the validity of more than a finite number of arguments, so it does not *matter* whether our theories about Baby represent his logical competence as finite... and so on, and so forth, blah, blah, blah.' I think that I am growing old. I no longer wish to discuss performance/competence arguments.

However, the present argument is not one of them. Rather, it is that there is something I know about inferences, something that I use routinely in validity checking, that network models cannot know and cannot use: viz. that arguments of the form A or $B \ldots \rightarrow A \vee B \ldots$ are valid in virtue of their form, and that the argument P or $Q \rightarrow P \vee Q$, and the argument P or Q or $R \rightarrow P \vee Q \vee R$

both reduce to the form A or $B \ldots \rightarrow A \vee B$. Networks cannot know this and cannot use it because there is no sense to the question 'what is the form of the argument that this network computes?' There is, of course, sense to the question 'What is the form of this *node label*?'. But—to put the point minutely differently from the last time—the labels are not part of the network.[5]

And while we are blocking misunderstandings, here is another one that ought to be avoided. The present argument is *not* that networks suffer from some computational incapacity as compared to machines with other kinds of architectures. Specifically, it is not that there are arguments of the form P to $P \vee Q$ that you cannot compute with a network. On the contrary, since there is a network that arbitrarily approximates any given deterministic Turing machine, you can use a network to compute *any* computible function. The interesting psychological question is not, therefore, about the generative capacity of network models; in fact, mere generative capacity is rarely what chooses among computational models in psychology. What is crucial is usually the ability of the model to capture important generalizations about how we think. In the present case, the issue is whether networks have access to information about arguments that apparently *is* available to us; information about the *form* of the arguments. And the answer appears to be that they do not, for, on the one hand, in a network the form of the argument being computed is represented only in the node labels; and, on the other hand, networks have no access to the labels on their nodes.

This paper is beginning to turn back upon itself, which suggests that it is getting to be time to stop. I hope the general pattern is now clear: The idea that information processors process information-as-Standardly-construed comports with the idea that information processors are typically associative networks. These ideas exhibit interlocked inadequacies: What is wrong with networks is that they transmit information without encoding it; what is wrong with the Standard notion of information is that while what we need is 'the information encoded' what it gives us is only 'the information transmitted'.

To allow ourselves a notion of 'information encoded' is to admit a new degree of freedom into our theory since we can now distinguish between various ways of encoding what is, from the Standard point of view, the same information transmitted. There is every reason to think we need this extra degree of freedom because there is every reason to think that psychological processes are sensitive to the character of the encoding of the information that organisms receive. Indeed (as I have argued elsewhere) there is every reason to think that psychological processes are sensitive *only* to the character of the encoding of the information that organisms receive; this is a way of putting the claim that psychological processes are sensitive to *syntactic* variable (like form) but not to *semantic* variables (like the information-objectively-transmitted). The appeal to 'attunement' is best viewed as an attempt to explain how mental processes *could* be sensitive to the information transmitted without regard to how—indeed, whether—the information is encoded. But as we saw, the appeal to attunement, closely scrutinized, comes to rather less than nothing very much.

It is sometimes said in praise of networks that they finally do get the ghost out of the machine. Unlike other computational models that cognitive scientists have been attracted by, networks require no executive; no little man in the head

whose job it is to assess the stimulus and plan the response. And, though Turing taught us that there is no *principled* objection to executive driven computers, the practical fact is that the hardest problems of cognitive theory tend to be problems of executive control. Maybe the reason that the homunculus has seemed so intractable is that, in point of fact, he is not there.

This may, for all I know, be right. One of the functions executives perform in machines that have them is solving coordination problems; determining the order in which the computational capacities of the machine will be exploited. Nothing I have said here shows that good simulations of people have to have executives in *that* sense. For all that I have argued, the mechanisms of coordination may be 'distributed'; for all that I have argued, problems of coordination may all be solved 'architecturally', i.e., at the level of the fixed structure of the machine.

But there is another thing that executives do; roughly and metaphorically, but close enough for our purposes they are there to read the labels on the nodes; to ensure that the computational consequences of exciting a node are specific to the information that the node displays. It has been the burden of my plaint that you need the labels because information has its behavioral effects only qua encoded; transmission is not good enough. But if you need the labels, then you also need a guy to read them.

Is there, then, *no* use for the Standard notion of information transmission in cognitive science? Yes, do not despair. There must be an internal code, for the sorts of reasons we have just reviewed. And that code must be semantically interpreted because beliefs and desires have contents and truth values. If, however, internal formulas have a semantic interpretation, there must be something that determines what their semantic interpretation is. It may be, it *just* may be, that what fixes the interpretation of formulas in the internal code is the covariance of their tokenings with tokenings of situations in the world. That is: it may be that what fixes their interpretation is the information about the world that they Standardly transmit. This is a dim hope, but as things now stand it is our best hope.[6] Is it, then, excessive optimism to suppose that progress will lead in the direction of a unified theory; a theory which, on the one hand, does justice to the richness and profundity of intentional phenomena and, on the other, unites the previously disparate—and sometimes apparently opposed—insights garnered from research in psychology, semantics and computation theory? And, if you are prepared to buy that, could I maybe show you something in a nice preowned car?

NOTES

1. The patient reader has earned a review of the terminology. Currently in play are the following notions: the information *transmitted*, the information *received*, the information *displayed*, the information *encoded* and the information *available*. For expository purposes, I assume that all except the last of these are to be applied to information processes in *networks*. Well then:

   Node i transmits information to node j iff the distinguished states of i and j are (let us say, causally) correlated (with the causation running from i to j).

Note j receives information from node i iff node i transmits information to node j.

Node i displays information to node j iff node i is 'visible' to node j (that is, iff i and j are intransitively connected).

Node i encodes the information that it transmits to node j iff it displays the information it transmits to j. (There is the following degenerate case: a node both displays and transmits the information that it is in one or another of its distinguished states. It follows that it encodes the information that it is in that state. The interesting case, by contrast, is the one where a node transmits and displays (hence encodes) information about the states of nodes *other than itself*.)

Finally, information is available to a system just in case the system is in a position to act on (or out of) that information. In the most familiar cases, the system is intentional and the information that it acts on is the object of one or other of its propositional attitudes: Psmith brings his umbrella because he knows (thinks/expects/fears... etc.) that it will rain.

2. Notice that no information can be available from the display unless it is transmitted by the display. In this sense (as Dretske has correctly insisted) information transmission places constraints on any other notion of information exchange. Alas, these constraints are, from the point of view of defining 'information processor', not very revealing.

3. By the way, Barwise and Perry's account of attunement to a relation fails in the other direction too. For example, I am sometimes disposed to respond to my kept potato plant in a way that would be appropriate to a fire: viz. I pour water on it. This does *not* imply, or even indicate, that I am attuned to a relation between fires and potato plants, or that I am trying to put my kept potato plant out.

   I am prepared to stop beating this dead horse, but only if Barwise and Perry are prepared to stop trying to ride it.

4. I should emphasize, however, that Barwise and Perry are clearly not comfortable with this behavioristic account of attunment, nor are they consistent in their allegiance to it. Thus they say, [2], p. 268, that "An organism... is attuned to the constraint [that a certain kind of plant is edible for that kind or organism] provided it eats the plant if it sees it, under certain circumstances—say when it is hungry...". But we are also told that "... an organism is attuned to C if, under certain circumstances, its cognitive conditions 'follow' C" (ibid) and it turns out, on p. 269, that "'following' is nothing but a very general type of inference, where [organism] a infers from the presence of a situation of type S that there is a situation of type S'". This is perfectly reasonable as far as it goes, but it makes attunement an inherently mentalistic notion, thus undermining the project of constructing an account of intentionality that relies solely upon an 'objective' conception of information. 'What the percept means to the organism' is now a matter of what the organism can infer from the percept; this is a long way from 'smoke means fire', and it is about as far as you can get from the "Ecological Realism" of Gibson and Reid (cf. [2], pp. ix–x). It is, indeed, a way of putting the moral of this paper that the only consistent form of Ecological Realism is behaviorism.

5. It will have occurred to you that we do not, strictly speaking, need the labels to recover the notion of validity in virtue of form; we could use the form of the networks themselves. This, however, does not help with the main problem; for, as we mentioned above, none of the nodes in the network can 'see' the network; all a node can see is its neighbors. In short: just as you cannot use the form of the labels to

define validity unless there is somebody in the system who can read the labels, so you cannot use the overall form of the network to define validity unless there is somebody in the system who can look at the overall form of the network. Neither condition is met by 'executive free' systems; that, indeed, is part of what it is for them to be executive free.

6. By the way, this sort of correlational theory of meaning is *far* more plausible for mental representations than for expressions of a natural language since the tokenings of the latter—but not, presumably, of the former—are contingent upon the motivations, linguistic competences and communicative intentions of the speaker who utters them; in fact, on a whole grab bag of "pragmatic" variables. Thus, for example: Suppose Psmith notices that Mary's hair is on fire, and hence, perforce, thinks: *Mary's hair is on fire* (thereby tokening the Mentalese expression whose truth condition is that Mary's hair is on fire). Whether he then *says* "Mary's hair is on fire" (thereby tokening the English expression whose truth condition is that Mary's hair is on fire) depends, *inter alia* on whether he thinks that Mary (or some other suitably situated auditor) would be interested to know that Mary's hair is on fire. (See [4] for an indication of how complex these sorts of pragmatic considerations can become.)

   In short, the correlation between mental representations and semantically relevant situations in the world is typically better (more reliable) than the correlation between English sentences and semantically relevant situations in the world. This is because the causal chain that connects the tokenings of mental representations to events which satisfy their truth conditions is typically *shorter than* (indeed, is typically a proper part of) the causal chain which connects the tokenings of English sentences to events which satisfy *their* truth conditions. That is the principal reason why it is mental representations, not English sentences, that are the natural candidates for being the primitive bearers of semantic properties.

REFERENCES

[1] Anderson, J., *The Architecture of Cognition*, Cambridge, Harvard University Press, 1983.

[2] Barwise, J. and J. Perry, *Situations and Attitudes*, Cambridge, MIT Press, 1983.

[3] Dretske, F., *Knowledge and the Flow of Information*, Cambridge, MIT Press, 1981.

[4] Grice, H. P., "Logic and conversation," in *Semantics of National Language*, eds. D. Davidson and G. Harman, Dordrecht, Reidel.

[5] Mowrer, D. H., *Learning Theory and the Symbolic Processes*, New York, Wiley, 1960.

[6] Osgood, C., "On creating and understanding sentences," *American Psychologist*, vol. 18 (1963), pp. 735–751.