11
The Possibility of a Cognitive Architecture

Andy Brook

> What representation is cannot . . . be explained. It is one of the simple concepts that we necessarily must have.
> —Kant, *Lectures on Logic* (trans. Michael Young, Cambridge University Press, 440)

Zenon Pylyshyn has made important contributions to a great many areas in cognitive science. One of them is cognitive architecture (hereafter CA). In fact, few people have paid more attention to CA than he has. Many researchers refer in passing to classical architectures, connectionist architectures, and so on, but Pylyshyn is one of few people who has come to grips with what CA is, what properties such a beast would have.

A number of arresting claims have been made about CA. Pylyshyn (1984, 1999) himself has claimed that it is that part of cognition which cannot be explained in terms of cognitive processes, and that CA is cognitively impenetrable, that is, that it cannot be influenced by beliefs or desires or other cognitive states and cannot be changed as a result of learning—it is the invariant framework of cognition, what persists through changes in cognitive contents. Fodor (1983, 2000) has claimed that we will never find a cognitive architecture for our central cognitive system. Against this claim, Anderson (2005) says that he has an existence-proof that Fodor is wrong, namely, the computational problem-solving system ACT-R. ACT-R does many things that central cognition does and clearly has a CA, so a CA for central cognition is not only possible, one actually exists. What are we to make of these claims?

A preliminary point. The term, "cognitive architecture," is used for two very different kinds of phenomena. (Multiple uses for a single term and other forms of terminological chaos abound in cognitive science.) By "cognitive architecture" some people mean the subsystems that make up a cognitive system: perception, reasoning, memory, and so on. Other people use the term to refer to the basic components out of which cognitive

capacities and/or cognitive contents are built. Candidates include the compositional components of representations, "physical symbols," a system of weighted nodes, and so on. Yet others use the term to mean both things. A question that exemplifies the first use: "Must cognition have both procedural and declarative memory?" Some questions that exemplify the second: "Are representations the building blocks of cognition?" "Do representations all encode information in the same way or are there different encoding formats?"[1] As to the third option, people who talk about physical symbol systems and connectionist systems as two architectures have both notions in mind (perhaps only implicitly), since the two systems have both overall structure and basic building blocks.

In this chapter, I will focus on the building blocks notion of CA. The system structure issue has its interest, but the shape of the issue is fairly clear there, at least if described at the level of generality adopted above. Put another way, the conceptual situation with system structure is less vexed than is the building block issue.

What We'd Like a Story about a CA to Do

"Vexed? What's the problem? Obviously, representations are the building blocks of cognition. What more needs to be said?" Alas, a lot more. To begin with, there are some major problems about what representations are. For example, when I am conscious of the world and my own states in a single act of consciousness, how many representations do I have? I am representing many things, but that does not automatically determine how many representations I have. Indeed, there is a case for saying "just one," as James (1890) pointed out. If so, what about representations that are not part of current unified consciousness, something true of most memories, for example? Here "many" may be the better answer. And so on (Brook and Raymont, forthcoming, chapter 8). But we can set these problems aside. Even if we could develop a clear account of what representations are—how to individuate and count them and so on—representations still could not be the ultimate building blocks of cognition.

First, as Fodor put it in a pithy comment about propositions, "If they're real, they must really something else" (1990 13). A single

representation, no matter how "single" is cashed, is just too big a unit to be the basic unit of a science. In addition, the question, "What are representations made up of?," is just too interesting and too relevant to the question, "What is cognition like?," for representations to be treated as a rock-bottom primitive in cognitive science.

Second, we already know that radically different accounts can be given of the "architecture" of representations, for example, the account in terms of integrated bundle of information contained by a vehicle of some kind of classical cognitive science (and virtually all philosophy of mind until recently) and the account that holds that representations are distributed throughout the hidden nodes of a connectionist system. Since these differences of architecture are cognitively relevant—that is to say, would affect how the respective systems would perform cognitive tasks—they have to be viewed as part of cognition, and therefore on a plausible notion of CA (yet to come) as reflecting differences of CA.

Third and most important, representations have a number of kinds of component, each of which could have its own architecture. In particular, we can distinguish vehicles of representation—acts of seeing, acts of imagining, and so on—from the object or contents of representations—what is seen or imagined. We can also distinguish a third element, the apparatus for processing representations, the overall cognitive system. A fourth might be whatever it is in such a system that interprets representational content.

So if we could find the CA of the various elements of representation and representing just identified, we'd be done? Who knows?—we need to step back a step or two. What do we want a CA to do for us? And how could we recognize one if we saw it? Compare chemistry. In chemical reactions, the building blocks are atoms. Atoms are not the smallest units there are, not by orders of magnitude, but they are the smallest units that enter into chemical reactions. (Suppose that that is true, anyway.) Moreover, and more to the point, when they combine, chemical reactions ensue, in such a way that when we understand how atoms combine into molecules, interact with other atoms and molecules, and so forth, we see that the interactions in question must result in—or be—some larger-scale chemical phenomenon of interest.[2]

Atoms have other properties that we expect of building blocks. Their relationships to other atoms can be affected by chemical reactions, but they themselves are not thus affected. Moreover, atoms are universal across the domain of chemical reactions, so that all chemical reactions can be understood as combinations of and transformations in the relationships among atoms. In addition, the properties of atoms vis-à-vis other atoms are systematic and orderly (electron structure, atomic number, and so on). Finally, atoms have properties that allow us to systematically connect the properties that make them building blocks of chemistry to a science of a fine-grained structure the bits of which by themselves do not enter into chemical reactions, namely, particle physics.

We can now draw some morals about how we'd like a story about CA to go. We'd like it to be a story that:

1. identifies the units which when combined become but are not themselves cognitive processes;
2. in such a way that when we understand how these units interact, we see that the interactions must result in or be cognitive processes, and we understand what kind of cognitive processes they are;
3. where the units are universal across cognition or at least across one or more domains of cognition, so that all cognitive processes or all within a domain can be understand as combinations and transformations of them (plus, perhaps, other things);
4. if there are a number of kinds of such units, the various kinds systematically relate one to another; and,
5. all this helps us to explain how the physical system in which the units are located can "do" cognition—how these building blocks or more molar aspects of cognitive function are implemented in the brain. They should be points at which we can begin to bridge the gap between cognitive and noncognitive accounts.

A tall order!

Note that the CA that we seek is a functional architecture. We are interested in the functions that combine with one another and work together with other things to yield—or be—cognitive functions such as

perceiving and thinking and imagining. I wish that I could now offer a nice, clean set of suggestions for what kind of functional unit might meet the five requirements just identified. Alas, I doubt that there is anything that will fill this five-part bill. But I may be able to clarify some of the context surrounding the issue. I will offer:

- a suggestion for a criterion by which we might recognize such functional units if we encountered them;
- some reasons why a couple of other suggestions along these lines won't work;
- some arguments claiming to show that Fodor's reasons for rejecting the possibility of a CA for central cognitive systems do not work; and,
- some reasons, I hope new reasons, for fearing that Fodor's conclusion is nonetheless right, that cognitive systems do not have a CA, not cognitive systems remotely like us at any rate.

Criteria for Being a CA

At least three suggestions, all due, more or less explicitly, to Pylyshyn (1984), have been made about what constitutes an item of CA, about the criteria for being a CA. They are related. Says Pylyshyn, we have got "down" to CA when:

1. The items in question remain unchanged, on the one hand through cognitive change (they are rearranged but are not changed), and on the other across implementations of cognition on different kinds of system.
2. The items in question are cognitively impenetrable. To be cognitively impenetrable is to be closed to influence by learning or by the semantic contents of beliefs or desires or other representations.
3. A cognitive account cannot be given of the items in question.

Start with (1). The first clause, unchanged through cognitive change, captures the idea that cognitive architecture is the framework or format of cognition, that within which cognitive change takes place. The second clause, across implementations, captures the idea of multiple realizability. On the standard functionalist picture of cognition, cognitive functioning can be "realized" in (can be done by) systems that otherwise have very different structures. Clocks can be realized in structures as different as sundials, hourglasses, weight-and-pulley systems, spring-and-escapement systems, the vibrations of a quartz crystal, and the wavelengths of certain elements. Like telling time, arithmetic, that is, encoding and manipulating numbers, can be done with pen-and-paper, a four-function calculator, an abacus, and so on. Similarly, thinking, remembering, perceiving, and so on could in principle be done by mammalian brains, silicon chips, quantum computers, and who knows what else. Or so many cognitive scientists claim.

There are problems with the first clause of (1). A great many things remain unchanged through cognitive change, the functioning of the glucose-distribution mechanism in the brain, for example. So the things that are unchanged must be elements of cognition. Alas, the latter notion is nothing more than the notion "cognitive architecture" in different words, that is, the target of our search in the first place. So modified, is the remain-unchanged requirement right? Not that I can see. So long as cognitive activity is built out of them, why should it matter whether the building blocks of cognition are (sometimes, under certain conditions) changed as a result of cognitive activity? (The cognitively impenetrable requirement, that is to say (2), would require—and explain—the remain-unchanged requirement; that may, indeed, be what gives (1) whatever attractiveness it has. See below.)

The second clause, about multiple realizability, does not work, either. Consider representations. For a host of reasons, it makes sense to say that serial computers, neural networks, and brains all have representations and can represent some of the same things using roughly the same forms of energy as inputs (i.e., in roughly the same sensible modalities). (I leave aside Searlean qualms about whether computers can really represent—it is enough for my purposes that all three function as

though they represent in ways relevant to cognition.) Yet the three have very different architectures—binary encoding, encoding distributed across multiple nodes, and . . . who knows what, but neither of the above. Since how the representations are encoded is relevant to how they function cognitively in their respective systems, then even if kinds of representation of kinds of things can be constant across cognitive systems, their architectures need not be.

The attraction of (2), the idea of cognitive impenetrability, again lies in the distinction between the framework or structure of cognition, which it seems should remain unchanged as cognitive functioning develops and changes, and what learning and changes of belief and desire can change. (2) would be an explanation of why cognitive activity does not change CA. However, the explanation would be only as good as (2) itself. How is (2) supposed to work? Is cognitive impenetrability meant to be necessary for being a CA, sufficient for being a CA, or merely distinctive of CAs?

On a standard account of visual illusions, impenetrability could not be sufficient. The usual example illusion is the Müller-Lyer arrowhead illusion, in which two lines of the same length look to be of different lengths when one line has arrowheads pointing out, the other arrowheads pointing in. Such illusions are either completely impenetrable cognitively or can be penetrated only partially and incompletely. Even when subjects know that the lines are the same length, indeed even if they have measured them and have demonstrated to themselves that they are the same length, the lines will still appear to be of different lengths—and this will continue either forever or at least for many, many trials. Yet visual illusions are not part of CA on anybody's notion of CA. They may be a near result of something in visual CA but they are not themselves a part even of visual CA. (For an interesting exchange on visual illusions, see Churchland 1989, 255–279, and Fodor 1990, 253–263.)

If so, it would follow immediately that impenetrability is not distinctive of CA, either. A property P cannot be distinctive of a class of objects O unless the presence of P is sufficient for the presence of an O. The reverse is not true: P being sufficient for O is not sufficient for P being distinctive to Os. P may be sufficient for other things sometimes, too. But P must be sufficient for Os if P is distinctive to Os.

That leaves necessity. Is impenetrability at least necessary for something to be part of CA? The answer is less clear but it is far from clearly "yes." Consider an analogy. It is plausible to suggest that within the narrow domain of logic, the CA is symbols and rules, and within the narrow domain of arithmetic, the CA is numbers and rules. (We leave open the question of whether some more universal CA underlies these specific CAs.) Yet how rules in the two domains are applied can clearly be influenced cognitively. How they are applied can be influenced by false beliefs about how the rule applies and training in an incorrect version of the rules, that is, learning. Other influences such as tiredness, cognitive overload, distraction, and so on, can affect the application of such rules but they are not likely to count as cognitive. Clearly the two just mentioned do. The same is true of how symbols and numbers are perceived, and perception is cognitive if anything is. Why should the same not be true of other CAs?

One can think of at least two ways in which the impenetrability test might be patched up. (a) Rather than being utterly impenetrable, perhaps the test should merely be that a CA is something that is not usually penetrated. Or (b), rather than being utterly impenetrable, perhaps a CA is not cognitively penetrated when the system is functioning properly—as designed, in such a way as to be truth-preserving, or whatever. ((b) is just (a) with a normative spin.)

There is something to these revisions to the impenetrability criterion. Part of the CA of a computer (CA of the building-block kind that we are examining) is how inputs and instructions are formatted. Such a format would not usually be penetrated by the operation of the computer. It could be but it would not be. And any such penetration would probably not be compatible with the computer operating as designed, not for very long anyway.

Impenetrability faces another problem, however—it does not give us much guidance with respect to finding CA. What feature of a CA would impenetrability be reflecting? Why should mere penetrability rule out having such a feature? If something is cognitively impenetrable, perhaps that would be enough to show that it is noncognitive. But being

noncognitive is not enough by itself to ensure that something is an element of CA.

So what about (3), Pylyshyn's suggestion that we have got "down" to CA when we cannot give a cognitive account of the units in question? What do we, or should we, mean by "a cognitive account" here? Pylyshyn seems to have in mind an account in terms of beliefs and desires and similar things—in terms of a person's reasons for what they think and feel and do. I think that this proposal is too narrow; it would leave us with an implausibly large list of items for CA. Why? Because there are many influences plausibly thought to be cognitive that involve little or no influence of beliefs, desires, and the like. Our tendency to complete word-stems in line with a stimulus that we have just encountered even if we have no awareness of having encountered it is an example. This is a semantic, therefore on a plausible notion of "cognitive," a cognitive influence, but no belief or desire is involved. The process at work is semantic implication or semantic association or something along those lines.

In response, let us broaden Pylyshyn's suggestion to include accounts in terms of semantic relationships of all kinds. In addition to providing reasons for belief or action, Pylyshyn's preferred kind, semantic relationships include semantic implication and semantic association, and perhaps others. What ties the group together is that they are all either true–false or truth-preserving, or evaluable using these concepts.[3] His thought would then become something like this. Even though a cognitive system is what brings semantic relationships into being, we will not be able to give a semantic account of the behavior of the elements out of which that system is built. On this test, CA would be elements out of which semantic systems are built that are not themselves semantically evaluable (nor, therefore, cognitive, by the concept of the cognitive we have just developed). This strikes me as fine as far as it goes. Finally!

Central Cognitive Processes and CA

If cognitive accounts are accounts in terms of semantic (not just syntactic, synaptic, or whatever) relations, then CA would be the architecture of the

semantic. Interestingly, this is exactly what Pylyshyn says (1999, 4). The content of representations is semantic content: It can be true or false, accurate or inaccurate, its lexical and propositional meaning has implications for what the content of other representations should be, and so on. One way to pin down what is meant by "semantic content" here is to say that the semantic content of a mental representation (MR) or other representation consists in the information in the MR (or . . .) about some object, event, or state of affairs that the MR would cart around with it even in the absence of the object, event, or state of affairs. If this content–vehicle distinction is sound, an account of the semantic must contain two parts: one about semantic content, what is seen, heard or imagined, and one about semantic vehicles. And our question now becomes, Is there an architecture for the semantic?

Let us start with a challenge to the very idea that cognition, or at least the brain, could deal with semantic relationships. Both Fodor (1994) and Dennett (1987) have urged that the brain, and therefore presumably the cognitive system, cannot process semantic content, only "syntactic" structures of various kinds. (By "syntactic," they don't mean syntax. They mean "physically salient," i.e., detectable by something that can detect only properties such as shape, spacing, order, and so forth.) As Dennett has put it (1987, 61), a semantic engine is impossible, so the most that a brain can do is to mimic this impossible thing. (Mimic? What is it to mimic something that is impossible? But we will let that pass.)

Even though some good philosophers have advanced this claim—Dretske also seems to accept or at least to presuppose it—not much by way of an argument is ever given for it. This is strange. On the face of it, the claim seems to be flatly false: we seem to process semantic information directly and effortlessly, indeed to do so every waking minute. If all that computational theories can account for is the processing of physically salient properties, so much the worse for such theories—but maybe what we really need is a less myopic picture of computation in the brain.

Even the philosophers who urge that the brain cannot *process* semantic content allow that the brain is *aware* of semantic content. But that requires that it be able to detect it. If so, skipping the details and cutting to the chase, it seems to me that what Fodor et al. should have said is that we

can process semantic content only via the processing of "syntactic" information, that semantic content must be built out of the "syntactic" (i.e., the nonsemantic) in some way—in which case, syntactic structures would not mimic semantic structure, syntactic structures would make up semantic structure. And the job of finding a CA would be precisely the job of finding these semantic-composing but themselves nonsemantic elements. Put another way, it is silly to deny that we are aware of semantic content—but it would be far from silly to say that the content that we thus know is built out of the nonsemantic.

There are some well-known arguments, due to Fodor, that whatever brain processes can do, the processing activity involved could not have a CA.

As we saw earlier, more than one kind of function is involved in representing. The same is true of the semantic. There is semantic content—what sentences say, the content of representations, and so on. There are vehicles for semantic content—representations, sentences. There are mechanisms for processing semantic content—roughly, perceptual, memory, and reasoning abilities. As we said earlier in connection with representation in general, there may also be whatever it is in a cognitive system that interprets semantic content. In the terms of this scheme, the arguments of Fodor's that we will now consider focus on the mechanisms for processing semantic content.

He mounts two arguments that there could be no general, perhaps simply no, CA of processing semantic content. One is found in (Fodor 1983), the other in (Fodor 2000). (Whether Fodor mounted the arguments in the service of this implication or even meant them to have this implication does not matter.) The argument in (Fodor 1983) starts from the idea that our central cognitive system is what he calls isotropic and Quinean. "Isotropic" means that information housed anywhere in the system could in principle increase or decrease the degree of confirmation of any belief in the system. "Quinean" means that if a belief comes into conflict with something that would tend, ceteris paribus, to disconfirm the belief, then ceteris need not remain paribus. To the contrary, there will always, or if not always certainly generally, be adjustments to be made

elsewhere in the system, indeed all over the system, that would reduce or remove the tension.

Why do these observations entail the impossibility of a CA? The first move is to urge that we cannot even imagine a computational system able to handle the computational load that would be entailed by these two features of the central system. Even a heuristic account could not do the job; it could not explain how we know which heuristic to apply. (This latter point is perhaps made more clearly in Fodor 2000, 42, than in 1983.) If so, the central system could not be computational, not by any notion of the computational that we now have or can imagine. The second move is to urge that computational cognitive science is not only all the cognitive science we have; it is all the cognitive science that we have any notion of how to formulate. From this it would follow that we have no notion of how the central system is built or what a CA of such a cognitive system could be like.

But is the case so clear? The mathematics for determining probabilities is, or certainly seems to be, isotropic and Quinean. Indeed, probabilistic reasoning is a large part of central processing in general. Yet a strong case can be made for saying that, here as elsewhere, the mathematics has an architecture: numbers and variables, formation rules, transformation rules, and so on. The fact, if it is a fact, that central cognition is holistic in the two ways that Fodor identifies does not seem to be enough by itself to rule out the possibility of a CA.

So what about the argument in Fodor 2000? It is considerably more subtle (also hard to figure out). The first move is to argue that the syntax of an MR is among its essential properties (all claims by Fodor discussed in this paragraph are taken from chapter 2). If only syntactic properties of an MR can enter into causal cognitive processes, it would follow that only essential properties of an MR can enter into causal cognitive processes. And from this, says Fodor, it would follow that cognitive processes are insensitive to "context-dependent" properties of MRs.

The last statement, however, is not true, says Fodor. As conservatism about belief-change and appeals to simplicity show, we often use widely holistic properties of cognition in the process of determining which cognitive move to make. Which theory counts as

simplest, which belief change counts as most conservative, will hinge on what other theories we accept. If so, the role of an MR can change from context to context. But the role of the syntax of an MR cannot change from context to context. If so, the cognitive role of an MR does not supervene on the syntax of that MR. Since the computational theory of the mind (CTM) models cognitive processes as causal relationships among syntactic elements, CTM cannot account for context-sensitive properties of syntax. So CTM is in trouble. Furthermore, since CTM is the only halfway-worked-out model of cognition that we have, we have no idea how to model context-dependent properties of cognition. If so, once again we have no notion of how the central system is built or what a CA of such a cognitive system could be like. QED.

Life is not so gloomy, thank goodness. Fodor twists his way through a couple of moves to try to save CTM, for example, building the whole of the surrounding theory into each MR (so that the essential syntax of the MR is the syntax of the whole theory) and concludes that such moves are hopeless. Who could disagree? What he didn't notice, however, is that there is a problem in his very first move. As we have seen, the term "syntax" as used in recent analytic philosophy is distressingly ambiguous. Sometimes it means, well, *syntax*—the kind of thing that Chomsky talks about. But sometimes it means all the physically salient properties of MRs and other representations. Unfortunately for Fodor, it is at most *syntax*, the real thing, that could be essential to MRs. There are lots of other physically salient properties—order, shape, and spacing, for example—that are not essential to MRs, that are context sensitive, and that computational processes could detect and make use of, including all the relationships of each MR to other MRs.

To get his gloomy result, Fodor would have to show that cognitive, context-sensitive properties could not supervene on any physically salient properties, not just syntactic ones, properly so-called. It is not clear how this could be done. Certainly he has not done it.

So where are we? At a bit of an impasse. On the one hand, Fodor's arguments, which are the leading arguments against the possibility of a CA, do not work. On the other hand, we do not have the foggiest notion how to think about CA.

Cognitive Architecture of What?

Let us return to the idea that a CA tells us what semantic content and processing is made of. If we want to know what a CA of the semantic might be like, the first thing we would need to know on this view is: What is the semantic like? Even Fodor would allow that our MRs have semantic properties, and that we often know them, even if CTM cannot account for them, indeed, even if, as Dennett (1987) maintains, the brain cannot process them, only syntactic analogues of them. So what has to be added to physically salient structures for them to take on semantic properties?

Fodor's own answer is: causal or nomological links to things in the world. Even without going into the details, it is clear that this approach faces major obstacles. It offers no account of semantic entailment. It has a huge problem accounting for nonreferring terms (e.g., "Santa Claus") (Scott 2002). It has next door to nothing to say about terms whose semantic role is other than to introduce things, events, or states of affair into representations (articles, etc.). And it has even less to say about semantic vehicles.

Semantic vehicles are the vehicles that provide our means of access to semantic content. Thus, an act of seeing is our means of access to the bluebird seen, an act of hearing is our means of access to the melody heard, an act of imagination is our means of access to a warm, sunny beach imagined.

Indeed, if our earlier suggestion about the structure of the semantic is right, a complete account would have to contain two more parts. The third would concern semantic processing—the cognitive equipment that processes the content in a given vehicle and relates it to other contents and vehicles, in memory, beliefs, and motives in particular.

A fourth would be this. The semantic content of an MR—what it represents with respect to what it is about—is not intrinsic to the representational vehicle in question. MRs not only represent something, they represent it to someone. And what they represent to the person who has them is a function, in part, of what else is going on in that cognitive system and in particular to how the cognitive system interprets the "raw

materials" that have made their way into the vehicle in question. Perhaps the biggest failing in Fodor's account is that he has nothing to say about this aspect of the semantic.

There is a lot of hand-waving in the remarks I just made. Let me be a bit more specific. On the issue of what this interpretive activity might be like, Wittgenstein and Davidson are interesting guides. (My saying this may surprise some because Wittgenstein is often thought to be unremittingly hostile to cognitive science, and Davidson only slightly less so.) The kind of representing that most interested Wittgenstein (1953) is the kind that requires the mastery of a "rule"—a capacity to judge that something is something or to attribute properties to something. Wittgenstein argues that such mastery cannot consist, first of all, of behavior or being disposed to behave in any way, or of a mental mechanism of any sort. The reasons are the same: Mastering a "rule" gives us the capacity to judge an infinite number of present, future, and merely possible cases, and a rule can be used correctly or incorrectly. No behavior, disposition, or mechanism could provide either by itself. However, mastery of a "rule" cannot consist in learning and then interpreting in accord with a proposition either. First, to apply the proposition, one would first have to interpret it, that is, figure out what it implies for the case in hand, which puts one on the edge of an infinite regress. Second, it will always be possible to think of yet-unencountered cases about which what to say is "intuitively" clear to us but about which the rule is silent.[4] So what does mastery of a rule, a judgmental capacity, consist in?

Just what was Wittgenstein's answer is, to say the least, controversial, but here is one reading that fits the texts fairly well. We have mastered a "rule," can apply it correctly, when the way we judge cases using it accords with how others do so. It is others' agreeing with us that makes our usage correct. This holds for all of present, future, and merely possible cases. (Once we have a base of agreement, we can then use the capacities thereby attained to generate new judgments. About these there need not be agreement.)

Davidson (2001) fills out this idea of intersubjective agreement in an interesting way. He introduces the idea of triangulation and urges that,

not just to have grasped a rule but to have judgmental capacities at all, there must be triangulation. What he has in mind is that to have judgmental capacities, two or more organisms must be able find in others' and their own behavior a pattern of making roughly the same judgments about objects, events, and states of affairs in shared perceptual environments.

If there is anything to Wittgenstein's and Davidson's story, there is not going to be any mapping between semantic content and physically salient structures of any kind in the brain—vehicles, circuits, or anything else, of any kind short of insanely complicated, anyway. If so, the prospects of a CA of semantic content are bad. Worse is ahead.

Externalism

Many philosophers now believe that some essential element of representational content consists of something outside the representation, some relationship between the representational vehicle and something. This is *externalism* about representational content, the view that at least part of what is required for something to represent, to be about something, is external to the act of representing. The content of representations consists in part of some relationship between the representational vehicle and something else. (Fodor's view was already an example of this.) Hilary Putnam (1975, 227) famously said, "meanings just ain't in the head." Note that there are two issues here, whether meaning extends beyond the head (an issue to which we will return) and whether it extends beyond individual representations (but not necessarily the head). All forms of externalism maintain at least the latter.

There have been quite a few candidates for what the external element is. For Dretske (1995), it is the *function* of a symbol, symbol-stream, instrument, or the like, that is external to the representational vehicle in question. When I imagine something, while all the information about what I imagine is or can be carried by the representational vehicle (Dretske 1995, 103–104, 114), what determines what the representation has the function of representing is not carried by the representation. For Burge, the content of a concept and therefore of any representation using

the concept is in part what is known or believed about the concept in one's society, not just by oneself (Burge 1979). For Putnam, the content of a concept and therefore of any representation using it is in part determined by the real nature of the kind of thing to which the concept refers, not just by what a given user of the concept happens to know (Putnam 1975). For Fodor, as we saw, the representational vehicle has to be in a causal or nomological (he goes both ways) relationship to something. And so on. Whatever the story, meanings extend beyond representational vehicles and many of them extend beyond the head. (We introduced this distinction just above.)

To be sure, not everything about representational content could be external to the representation. Here is one of the ways in which Dretske, for example, tries to strike the right balance. On the one hand, clearly there is something about representation that is internal to the cognitive system that has the representation. Dretske captures that this way:

> When I close my eyes, I cease to see [the world around me]. The world does not vanish but something ceases to exist when I close my eyes. And this something has to be in me. (Dretske 1995, 36)

Furthermore, it is plausible to think, the contents of beliefs and desires have to be internal to me at least in part to explain the effects that they have on my behavior. I go to the fridge because I want something to drink and remember there being juice there. If the contents of the desire and memory were not internal to me, how could they—and why would they—affect my behavior as they do? Other phenomena are relevant, too. When I move to a new location, most of my external links change. Yet my mental contents move with me and most of them do not change: My thoughts, imaginings, emotions, memories, and so on continue to have the same contents. Even in a sensory deprivation chamber, I would still have thoughts and feelings, meanings and beliefs, even though I am cut off from all current causal contact with everything outside the tank. And so on. There is something important about representation and consciousness that is entirely "in the head."[5]

On the other hand, paying attention to how a representation is implemented in the brain is not likely to tell you either what you are representing or how you are representing it. Dretske uses analogies to argue for this. Staring at the face of a gauge is not the way to discover what information it provides or how it provides it (1995, 109). Peering at a meaningful symbol is not the way to discover its meaning or how it provides this meaning. Looking at neurons will not tell us what they do. (Neurons are all pretty much alike, or at any rate fall into a small handful of different kinds.) If so, why should we expect to learn what a psychological state represents by "peering" at it—even from the inside?[6]

> What something represents, what it has the function of indicating, are not facts one can discover by looking in the cortex, at the representation. One could as well hope to figure out what, if anything, a voltage difference between two electrical contacts in a computer means (represents, signifies) by taking ever more exact measurements of the voltage. (Dretske 1995, 37)

Dennett (1978, 47) mounted a similar argument many years ago. Suppose that we devised a brain scanner with sufficient "resolution" to find the sentence, "America is the world's greatest country!" written in Hans's brain. We couldn't tell, merely from discovering the sentence, whether Hans is jingoistic, ridiculing the idea, using the sentence as an example of a proposition in a philosophy class, just liked the sounds of the words, or what.

To illustrate the point, consider an intention to refer to something. We are able to pick out and refer to particular examples of a kind of thing and distinguish them from even qualitatively identical other examples of this kind. No representation by itself contains anything that we could use to achieve such reference, or so it is plausible to claim. Rather, what accomplishes singular reference, as Perner (1991, 20) calls it, is that one intends to refer to that object and no other. We could not pick out referential intentions by gazing at the contents of representations.

Here is how, for Dretske, representing something requires something outside the representational vehicle. A thermometer has the

function of representing, and so does represent, temperature, even though its indicating activities might correlate with other factors, level of charge in a battery or behavior of electrons in silicon, for example. Function is assigned by something outside a representation. The same is true for becoming a representation. Whether the indicator being at a certain place means so many degrees Fahrenheit or centigrade or . . . or . . . is determined by factors external to the thermometer, the factors that allow us to interpret marks and squiggles on it as indicators of particular temperatures.

On the externalist picture, then, representational content is fixed not just by the structure of the representational vehicle and the information entering it but also by complex relationships between that vehicle and the rest of the cognitive system. Beliefs about functional assignments, beliefs about causal and/or nomological connections, perhaps the functions and connections themselves, and who knows what else will enter into fixing representational content (not to mention good old-fashioned interpretation and intersubjective agreement, as we saw in the previous section).

If externalism of any form is true, the prospects for a CA of semantic content dim still further. Various instances of representational content consist of so many kinds of relationship between representational vehicle and what is around it, some vehicle–brain, some vehicle–world, that it is unlikely that the various contents will have many common constituents, will be built out of any single CA. Indeed, some of the relationships claimed to be relevant don't have any obvious simpler components at all—function assignment, for example. In addition, as Dretske puts it, even if representational content is globally supervenient on, for example, neural activity, so that for every change in representational content there will be some neuronal change, representational content will not be locally supervenient. That is to say, the contents of individual representings cannot be mapped onto specific, localized circuitry—which dims the prospects of them consisting of a single set of components even further.

Objection: "Didn't you say that not just representational vehicles but also representational content could persist through the closing of eyes

and moves to new locations? If so, something crucial to representational content can be entirely inside the head—and the idea of CA for representational content would appear to get a new lease on life." Not a very big one, I am afraid. Distinguish:

1. Something being external to a representation

from

2. Something being external to the head (the person, subject, self, mind, consciousness).

The external element postulated by many externalists is something external only to representations, not to the head. Dretske is an example; function and even knowledge of function would be external to a representation, on his account, but they are not or certainly need not be external to the person who has the representation. If true, this would explain how we can cart representational content around with us even with our eyes closed.

What might be meant by saying that something external to a representation is still internal to a cognitive system? In my view, information is internal to a cognitive system when it is cognitively available to the system. Information is cognitively available to a system when the system can use it to structure cognitive activity: perceiving, thinking, remembering, emoting, and so on. (Just existing somewhere in the system in the way in which sentences exist in a book wouldn't be enough.) There are different ways in which a system can use information to structure cognitive activity, running from completely automatically without the system having any consciousness of what is going on, in the way that syntactic rules structure the parsing of sounds. to being known consciously and applied deliberately. On all of these alternatives, the information is internal to the system.

The external elements postulated by many forms of externalism are available to the system in this way. On Dretske's brand of externalism, for example, what is external to a representation is its function: what

information it has the function of representing. The cognitive system not only has this information but has to have it to know (or even to have beliefs about) what information is being represented by this, that, or the other representational vehicle. And it does have it. Just as we grasp what the indicating function of a gauge is, we grasp what the indicating function of a perception is. We may not be conscious of what this function is, we may not be able to describe it, but we do grasp what it is—what a given perception represents and when it is misrepresenting.[7] If so, then the element external to representations in Dretske's picture can still be, indeed often must be, internal to the cognitive system that has the representations.

Notice that the same is true on Fodor's account. For Fodor, the external element is a causal or nomological link between symbol and referent-type. No matter. To use the symbol, the system must grasp (in the special sense we have been using, which does not imply being able to articulate) the kind of thing to which the symbol refers. Thus, on Fodor's account, it must grasp to what the symbol is linked. The same is true of Burge's social externalism. In general, many of the elements said to be external to representations will nonetheless be internal to the cognitive system that has the representations.

What about the externalists who deny that the element external to representations is graspable by the cognitive system that has them—Putnam (1975), for example? Here is how his story goes. Suppose that Adam and his twin on twin earth, Twadam, have beliefs about a certain clear liquid in front of them and both call it "water." One liquid is H_2O and one is XYZ. They have beliefs about different things and so, Putnam concludes, have different beliefs. Here the element of content external to the beliefs is not graspable by the cognitive system and so is external to the system in every way. Is this a problem for our story about how, even in the face of externalism, representational content can be internal to the cognitive system whose representation it is?

No. It is far from clear that the external element here does affect the content of representations. They both believe "This [the substance in front of them] is water." However, if they don't also believe, "This is water-rather-than-twin-water" (in their respective idiolects), then their concept

of water in each case may well be broad enough to range over both substances. If so, their beliefs, representations, and conscious states would have the same content. (Brook and Stainton [1997] reach the same conclusion in a different way.)

Anyway, if some element of content is external to the cognitive system, that element could not be made up of anything in a system's CA, so it would be of no concern to us.

That on many accounts the element external to representational vehicles is nonetheless graspable by, and therefore in some way internal to, the cognitive system as a whole does not improve the prospects for there being a CA, unfortunately.[8] Why? For reasons we have already seen. To fix representational content, elements widely dispersed across a cognitive system would still be needed. This would make it unlikely that various instances of content have a common structure or are locally supervenient on anything that does.

If there is anything to the story that we have told in the previous section and this one, the prospects for representational content having a CA are bad. What about the prospects for a CA for representational vehicles, acts of representing? Here the prospects may be better—but the issue is also less important. How a given content is represented is much less important cognitively than what the content contains, is about. Why? Because whether something is seen or heard matters less for belief, memory, planning, action, and so on than the content of what is seen or heard. Moreover, since all acts of seeing, for example, are much alike, the nature of the representational vehicle won't usually have much influence on differences of representational or semantic content. Where representational modality does matter, as it does for example over whether something is seen or imagined, that is usually because the difference is an indicator of something else, in this case the status of the content as information about the world—and this status can be determined independently.

With these brief comments, I will leave the topic of a CA for representational vehicles. What about the fourth element, the thing that does the interpreting, enters into the intersubjective agreements, and so on—what used to be called the *subject of representation*? Given the

centrality of interpretation in most current accounts of semantic or representational content and of triangulation or a related notion in not a few, it is remarkable how little attention this topic has received. Whether an interpreter/agent/subject has a CA has received even less. Could this entity have a CA? The whole issue has been so little explored that it is hard to know. There is a major piece of work to be done here.

Concluding Remarks: Prospects for Cognitive Science

Many theorists have suggested that if Dretske and Wittgenstein and Davidson are right, if content has elements that are external to representational vehicles, is a result of interpretation, has an ineliminable intersubjective element, and so on, then not just the prospects for a CA but the prospects for any science of cognition at all are dim. I do not think that this bigger conclusion is warranted.

There are more kinds of science than building models Lego-style. There is also correlational science. As recent brain-imaging work has shown, this kind of science can be exciting and revealing even without a story about building blocks and how they combine to yield target phenomena. A second kind of science flows from the point that Pylyshyn has repeatedly made (e.g., in Pylyshyn 1984 and 1999) that the only way to capture many of the similarities in cognitive function of interest to us across differences of implementation is to use the proprietary language of cognition, the language of representation, belief, desire, and so on. But phenomena thus described can be explored, models can be built, computational simulations can be created and tested, and so on with no story about CA. To return to where we began, even if Fodor is right that individual representations are too "big" a unit to be the ultimate constituents of a science of cognition, we can do a great deal of interesting and important work by treating them as though they are.

References

Anderson, J. (2005). The modular organization of the mind. Talk presented at Carleton University October 13, 2005.

Bechtel, Wm. (2005). Mental mechanisms: What are the operations? *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, 208–201.

Brook, A. and P. Raymont (forthcoming). *A Unified Theory of Consciousness*. Cambridge, Mass.: MIT Press/A Bradford Book.

Brook, A., and R. Stainton (1997). Fodor's new theory of content and computation. *Mind and Language* 12: 459–474.

Burge, T. (1979). Individualism and the mental. *Midwest Studies in Philosophy* 4: 73–121

Chomsky, N. (1980). *Rules and Representations*. New York: Columbia University Press.

Churchland, P. M. (1989). *A Neurocomputational Perspective*. Cambridge, Mass.: MIT Press/A Bradford Book.

Davidson, D. (2001). *Subjective, Intersubjective, Objective*. Oxford: Clarendon Press; New York: Oxford University Press.

Dennett, D. C. (1978). Brain writing and mind reading. In *Brainstorms*, 39–52. Montgomery, Vermont: Bradford Books.

Dennett, D. C. (1987). Three kinds of intentional psychology. In *The Intentional Stance.* Cambridge, Mass.: MIT Press/A Bradford Book.

Dretske, F. (1995). *Naturalizing the Mind*. Cambridge, Mass.: MIT Press/A Bradford Book.

Fodor, J. (1983). *Modularity*. Cambridge, Mass.: MIT Press/A Bradford Book.

Fodor, J. (1987). *Psychosemantics*. Cambridge, Mass.: MIT Press/A Bradford Book.

Fodor, J. (1990). *A Theory of Content and Other Essays*. Cambridge, Mass.: MIT Press/A Bradford Book.

Fodor, J. (1994). *The Elm and the Expert*. Cambridge, Mass.: MIT Press/A Bradford Book.

Fodor, J. (2000). *The Mind Doesn't Work That Way*. Cambridge, Mass.: MIT Press/A Bradford Book.

James, W. (1890). *Principles of Psychology*, vol. 1. London: Macmillan.

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly* 64: 354–361.

Macleod, M. (unpublished). Rules and norms: What can cognitive science tell us about meaning. Talk presented at Carleton University, November 24, 2005.

Perner, J. (1991). *Understanding the Representational Mind*. Cambridge, Mass.: MIT Press.

Putnam. H. (1975). The meaning of "meaning." In his *Mind, Language and Reality: Philosophical Papers*, vol. 2, 215–271. Cambridge: Cambridge University Press.

Pylyshyn, Z. W. (1984). *Computation and Cognition: Toward a Foundation for Cognitive Science.* Cambridge, Mass.: MIT Press/A Bradford Book.
Pylyshyn, Z. W. (1999). What's in your mind. In *What Is Cognitive Science?* ed. E. Lapore and Z. W. Pylyshyn. Oxford: Blackwell.
Scott, S. (2002). *Non-Referring Concepts.* PhD Dissertation, Institute of Cognitive Science, Carleton University, Ottawa, Canada
Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell.
Wittgenstein, L. (1967). *Zettel*. Oxford: Blackwell.

---

1. The latter question was one of the issues at stake in the mental imagery wars of the 1980s and 1990s in which Pylyshyn was a prominent participant.
2. I borrow this idea of mechanisms that must yield a target phenomenon from Levine's (1983) account of how we close explanatory gaps, gaps in our understanding of how and why something happens. For interesting suggestions about the kind of mechanisms that we should seek here, see (Bechtel 2005).
3. Some would argue that there is more to semantic evaluation than being true–false or truth-preserving, including more kinds of satisfaction-conditions. The issue is not important here.
4. Drawn heavily from Mark Macleod, unpublished, with thanks.
5. It is not often noticed that the situations just described would not break the link to the external element on all forms of externalism. Causal links would be broken but nomological links would not. Social links would be broken but links to social practices would not. Functional links might be broken but function-assigning histories would not.
6. Compare this remark by Wittgenstein (1967, §612): "What I called jottings would not be a *rendering* of the text, not so to speak a translation with another symbolism. The text would not be *stored up* in the jottings. And why should it be stored up in our nervous system?"
7. Chomsky's (1980) way of putting the point that I am trying to make here is to say that one *cognizes* the function.
8. This external–internal mix does help in other places. It can be used to show that externalism is no threat to the view that consciousness is a kind of representation, for example (Brook and Raymont, forthcoming, ch. 4).