Meaning and Content in Cognitive Science

In Prospects for Meaning, Richard Schantz, ed. de Gruyter, Berlin & New York

(2012)

Robert Cummins
Philosophy and Beckman Institute
University of Illinois at Urbana-Champaign
rcummins@uiuc.edu

Martin Roth
Department of Philosophy
Knox College
maroth@knox.edu

1. Introduction

What are the prospects for a cognitive science of meaning? As stated, we think this question is ill posed, for it invites the conflation of several importantly different semantic concepts. In this paper, we want to distinguish the sort of meaning that is an explanandum for cognitive science—something we are going to *call* meaning—from the sort of meaning that is an explanans in cognitive science—something we are not going to call meaning at all, but rather *content.* What we are going to *call* meaning is paradigmatically a property of linguistic expressions or acts: what one's utterance or sentence means, and what one means by it. What we are going to call *content* is a property of, among other things, mental representations and indicator signals. We will argue that it is a mistake to identify meaning with content, and that, once this is appreciated, some serious problems emerge for grounding meaning in the sorts of content that cognitive science is likely to provide.

2. Representation and Indication.

Cognitive science appeals to two main sorts of things that have contents: representations and indicator signals. We are going to spend some space describing these and then turn to why it is dangerous to think of their contents as meanings.

2.1. Indication. In the theory of content, ‘indication’ is used to talk about detection. Familiar examples include thermostats, which typically contain a bimetallic element whose shape detects the ambient temperature, and edge detector cells.[1] Other examples include the lights in your car’s dashboard that come on when the fuel or oil level is low, and magnetosomes, which are chained magnetic ferrite crystals that indicate the direction of the local magnetic field in certain anaerobic bacteria.

When thinking about detection, it is important to distinguish the mechanism that does the detection from the state or process that signals that the target has been detected. The cells studied by Hubel and Weisel—the so-called edge detectors—are indicators, but the pattern of electrical spikes they emit when they fire are indicator signals. Similarly, thermostats are indicators, while the signals are currents that result when the bimetallic element bends so as to close an open circuit.

2.2 Representation. Familiar examples include maps of all kinds, scale models, graphs, diagrams, pictures, holograms, and partitioned activation spaces. Cognitive maps (see below for discussion) are paradigm examples of what we mean by representations in the mind/brain. They are structured, and their content is grounded in that structure rather than in correlations with other events or states.

2.3. Indication vs. Representation. Though causal and informational theories of

[1] David Hubel and Torsten Wiesel (1962), who discovered such cells, write: “The most effective stimulus configurations, dictated by the spatial arrangements of excitatory and inhibitory regions, were long narrow rectangles of light (slits), straight-line borders between areas different brightness (edges), and dark rectangular bars against a light background.”

representational content generally assert that representational content is, or is inherited from, indicator content, indication and representation should be kept distinct. For starters, indication is transitive, whereas representation is not. If S3 indicates S2, and S2 indicates S1, then S3 indicates S1. If a photo-sensitive cell pointed at a light in your car's dashboard is attached to an audio device that plays a recording "the water temperature is high" whenever the light goes on, then, if the light indicates low oil pressure, so does the recording. Notice that what it indicates has nothing to do with what it means. Representation, on the other hand, is not transitive. A representation of the pixel structure of a digitized picture of the Sears Tower is not a representation of the building's visual appearance, though the latter may be recovered from the former because a representation of the pixel structure encodes the building's visual appearance.

The transitivity of indication implies that indicator signals are arbitrary: given transitivity, in principle anything can be made to indicate anything else. Because indicator signals are arbitrary, systematic transformations of whatever structure the signals may have cannot systematically alter their contents. But structural transformations can systematically alter the contents of representations, and such transformations are what make representations useful. Consider, for example, software that "ages" a digitized image of a face, i.e., returns an image of that face as it is likely to look after some specified interval of time. Nothing like this could possibly work on an input that was required only to indicate a certain face—a color, say, or a name—because there is no correlation between the physical characteristics something must have to be a signal that indicates the appearance of a face at age 18 and the physical characteristics of that face at age 18. It follows from the nature of indication that the structural properties of an indicator signal have no significance. Indicator signals demonstrate that their targets are there, but are silent about what they are like. Representations, on the other hand, mirror the structure of their targets (when they are accurate), and thus their consumers can cognitively process the structure of the target by modifying the structure of its representation. But unlike indicator signals, representations are typically silent about whether their targets are "present." Only incidentally and coincidentally do they detect anything.

Because edge detector cells all generate the same signal when they detect a target, you cannot tell, by looking at the signal itself (e.g., the spike train), what has been detected. Rather, you have to know which cells generated the signal. This follows from the arbitrariness of indicator signals, and is therefore a general feature of indication: indicators are source dependent in a way that representations are not. In sum, then, because indication is transitive, arbitrary, and source dependent while representation is intransitive, non-arbitrary and not source dependent, indication and representation are different species of content.

3. Meaning in Cognitive Science

We will discuss a number of reasons why it is dangerous to think of the sorts of contents had by representations and indicator signals as meanings. The first is that meaning is linked to understanding. Meaning is idle unless someone understands what is meant. But the representations and indicator signals of cognitive theory are not supposed to be *understood*; they are supposed to be computationally *processed.* Equating content with meaning engenders a regrettable tendency to think that the processing of a representation or indicator signal is supposed to amount to *understanding* it.[2] Critics are quick to point out that it doesn't, and they are right. Cognitive Science hopes to explain what it is to understand an utterance by appeal to the processing of representations and indicator signals, so the notion of content applied to representations and indicator signals had better not presuppose understanding anything. We trivialize the problem of understanding understanding, and hence of meaning, if we suppose that utterances are simply translated into sentences of the language of thought which are already understood simply because it is the language of thought.

Another reason why it is dangerous to think of contents as meanings is that it suggests that a theory of content is, or is something that grounds, a *semantics* for content. This would be harmless were it not for the fact that semantics now means, for all intents and purposes, specifying references and truth conditions of the sort famously recommended by Davidson in "Meaning and Truth" (1967). With the publication of that seminal article, meanings came to be references and truth conditions, and semantics came to be the now familiar combinatorial truth-conditional semantics pioneered by Tarski (1956). As a consequence, the idea that mental representations or indicator signals have meanings became the idea that they have references and truth-conditions—what else is there, after all?—and the theory of content was seen as the attempt to say what fixes the references and truth-conditions of the things cognitive processes process (Fodor, 1990).

If you want to have truth-conditional semantics, however, you need your bearers of meaning to have logical forms, so you need them to be language-like. The idea that mental representations and indicator signals have meanings thus leads, through the

[2] Notice, too, that this makes any attempt to equate understanding a sentence with information flow a non-starter. If you tell me something I already know, I can understand what you say although the information associated with your utterance is zero.

Davidsonian Revolution, to the Language of Thought (LOT). This is a Bad Thing. It is a Bad Thing because, so far as we know, the representations and indicator signals required by Cognitive Science don't have logical forms, and are not candidates for truth-conditional semantics. They are, in this respect, in good and plentiful company. Pictures, scale models, maps, graphs, diagrams, partitioned activation spaces, magnetosomes, tree rings, fish scale ridges, sun burns, idiot lights and light meters all have *contents*, and none of them are candidates for truth-conditional semantics.[3]

There is another route to the idea that content is meaning. You start with propositional attitude psychology, aka folk psychology, aka BDI psychology (for **b**elief, **d**esire, **i**ntention psychology). Propositional attitudes require propositional contents, and propositions are the sorts of things expressed by sentences and only sentences. So, if you think the mind is a BDI machine, you are going to think there must be a language of thought (LOT), as Fodor famously and correctly argued (1975). This language of thought actually has to be a *language*, in that it has to be in the proposition expressing business. Moreover, as Fodor has also famously argued, if we are going to accommodate the productivity of the attitudes, that language is going to have to have a combinatorial semantics of just the sort Davidson argued we should find for natural language. So: BDI, LOT and truth-conditional semantics are quite literally made for each other. BDI, to repeat, requires propositional contents, and LOT is the only way to have them. The only remotely plausible or worked out semantics for language (hence LOT) is the now standard truth-conditional semantics. BDI, LOT and truth conditional semantics (TCS) are thus tightly knit together. It is a neat package, and working it out has been a major

[3] LOT sentences might represent propositions (or the structure of a natural language sentence), but they do not, in the sense intended above, represent anything that isn't structured like a sentence. They might, of course, encode such a structure. See Cummins, et al. (2001) for further discussion.

industry in the philosophy of mind, an industry that, following Fodor and Pylyshyn (1988) we might call the classical program. Unfortunately, however, there are compelling reasons for thinking that *the brain doesn't work that way.*

4. Against the Classical Program

There are many compelling reasons for rejecting the Classical Program. Here are some of the most prominent concerns.

There are no portable symbols in the brain. LOT proposes to get complex representations from semantic primitives via the usual combinatorics. The semantic primitives in question are typically indicator signals. But indicator signals are source dependent. As we pointed out above, an edge detector with a vertical line as target emits the same signal as one with a horizontal line as target. Thus, LOT theories must assume that indicator signals are somehow typed, and typed by something other than their source, since this will be absent in non-detection uses in complex expressions. The occurrence of a |cat| in a detection case will be a different brain event than the occurrence of a |cat| in a |There is a cat on the mat|. What makes both occurrences of |cat| tokens of the same type? It cannot be "shape" (e.g., of the spike train), since that does not correlate with content. It is not clear what this *could* be. There appears to be no evidence that, e.g., the occurrence of a given pattern of activation in one area of the brain is informationally related to the occurrence of the same pattern in another area. However, representations, because their content is grounded in their structure, do allow for portability of representation, since the same structures in different neural circuits will share content.

A good deal of perception and cognition is not propositional. There are, basically, two routes to this idea. The first is that, given what we know about brains, they don't seem to be much like symbol systems (Rumelhart, 1989; Sejnowski, Koch, and Churchland, 1988). The second is that, given what we know about BDI-LOT machines (e.g. PLANNERS, of the sort studied extensively in what Haugeland (1985) calls GOFAI), the computational demands involved in anything but the most simple tasks don't seem compatible with what we know about brains (e.g. Feldman and Ballard's 100-step-program (1982)). Moreover, the currently most prominent argument for LOT, namely the systematicity of thought, either begs the essential question by presupposing BDI, or is an equally good argument for non-propositional representation of domains that do not themselves have a propositional structure, namely everything other than language (Cummins, 1996a).

The Classical Program is not easily reconciled with non-human cognition. Human brains evolved, and they are, at bottom, not that different from other mammalian brains. This is not to deny that there are critically important differences between humans and non-human primates and other mammals. Nor it is to deny that the evolution of mind has been shaped by the particular adaptive problems facing each species. It is rather to emphasize how unlikely it is that human brains are fundamentally BDI-LOT machines, while non-human brains are not. The alternative is to suppose either that non-humans have no minds, or that all mammalian brains (perhaps all brains) are BDI-LOT machines. Neither option is attractive.

"Naturalizing" truth-conditional semantics for the mind hasn't fared well. It is safe to say that there is no consensus concerning the "right" way to ground a truth-

conditional semantics for LOT. It is time to consider the possibility that this is not our fault. A lot of smart people have worked hard on this for quite a while. Maybe they just haven’t got it yet. But maybe it is just a bad problem.

Truth-conditional semantics, and the increasingly arcane metaphysics that is required to sustain it, has become so complex that it is no longer remotely plausible to think that knowing what a sentence (or utterance or whatever) means is knowing its truth-condition. Consider the disputes between continuant theorists and those who support temporal-part (four-dimensionalist) theories about objects and their persistence. According to four-dimensionalism, ordinary physical objects do not exist fully present from moment to moment; rather, ordinary physical objects are space-time “worms” with various spatiotemporal parts that are related in various ways, e.g., causally (Heller, 1990). On this account, the truth conditions of such claims as “The chair I am sitting on now is the same chair I sat on yesterday” rest on whether the spatiotemporal part(s) of an object I am sitting on now and the spatiotemporal part(s) of an object I was sitting on yesterday belong to the same “worm.” Suppose this is all correct; it follows that for a child to understand the claim that the chair she is sitting now is the same chair she was sitting on yesterday, she must know a truth condition involving four-dimensional worms, temporal parts, and various relations between those parts. Or consider claims about modality. If David Lewis (1986) is right, then the claim that I could have had Cheerios for breakfast this morning is true just in case my counterpart in some possible world does have Cheerios for breakfast. Thus, a child’s understanding of that claim requires knowing, among other things, about counterparts and possible worlds. Defenders of TCS should find these consequences embarrassing, or at the very least provide some plausible motivation for thinking that children have implicit knowledge of the metaphysics of four-dimensional worms and counterparts in possible worlds (other than to salvage TCS). Note, by the way, that this is not a knock on four-dimensionalism or counterpart theory; rather, it is meant to illustrate the ways in which wedding the metaphysics to the semantics can make implausible demands on what the folk know.[4]

5. How does content relate to meaning?

We are not going to defend any of the above claims in detail. But if you think any or all of them have a reasonable chance of being true, then it behooves you to consider alternatives to BDI and LOT and truth-conditional semantics for the mind. You should, we think, start thinking seriously about the following questions:

1. How should we understand mental content if we abandon the “classical” picture?
2. What implications does a non-classical conception of mental content have for our understanding of *meaning*?

We have already said something about the first question, and that is largely motivates the second question. If the mind is not, at bottom, a propositional engine, then how is propositional thought possible? Or, to put the problem somewhat differently, how can we understand language if truth-conditional semantics correctly describes linguistic meaning, but does not correctly describe mental content? After all, it would seem to be a truism that to understand a sentence expressing the proposition that the Eiffel Tower is in Paris, you have to be able to have the thought that the Eiffel Tower is in Paris. But surely, to have the thought that the Eiffel Tower is in Paris, you have to get into a mental state

[4] But see Jeff King (1995) for a possible way to get around these worries.

whose content is the proposition that the Eiffel Tower is in Paris. Language users, it seems, *must* harbor propositional attitudes, and hence must have a LOT whose semantics mirrors that of language.

But is this apparent truism actually true? How can this picture possibly be accurate if, as the cognitive sciences seem to be telling us, mental contents are *not* propositions? If language expresses propositions—if meanings are truth-conditions—then there has to be a mismatch between what goes on in your head and what you say, and between what you say and what goes on in *my* head. Imagine, for a moment, that the mind is a picture processor. Given the rather obvious fact that a picture is not worth any number of words, this seems to be a case of massive communication failure, what Cummins called forced error (1996b). We could, it seems, give a kind of reverse Fodorian argument: Cognitive Science says our mental states do not have propositional contents. But we do understand language. Hence the standard semantics for language must be wrong. This is temptingly radical, but not to be seriously recommended by anyone who is not prepared to abandon the standard semantics for language.

We can begin to buzz ourselves out of this bottle by noting that communicative signals do not *have* to share a semantics with the messages they communicate. A simple and familiar example of this is the transmission of pictures by pixilation. To send a grey scale picture, you need a signal system that is capable of specifying position-intensity value pairs. The picture sent, however, has a content completely disjoint from the contents of the signals.

This example demonstrates that successful communication does not require that the message communicated have the same content, or even the same *kind* of content, as the signals that communicate it. Communicative systems can be, as it were, recipes for assembling representations whose contents are utterly disjoint from the contents of the recipes themselves. So, accepting truth-conditional semantics for language doesn't *force* you to accept it for the mind. You cannot simply read off properties of mental content from properties of linguistic content—meaning—given only the fact that we understand language. In principle, linguistic signals *could be* recipes for assembling pictures (or maps or graphs or all of these or something else entirely) in your profoundly non-propositional head. This would allow us to have our truth-conditional semantics for language and a biologically realistic cognitive science too. If understanding a sentence with the content *that the Eiffel Tower is in Paris* doesn't require having a mental state with that (propositional) content, then meaning could be just what Davidson said it was, and the mind could still be what biology says *it* is.

But could this possibly be true? Isn't it just obvious, empirically, if not logically, that when I understand a sentence expressing the proposition that the Eiffel Tower is in Paris, that I have the thought that the Eiffel Tower is in Paris? We think it is. But we don't think this seemingly obvious fact (if it is a fact) runs very deep. We think Plato was right about thoughts, construed as propositional attitudes: thinking is just talking to oneself. Thoughts, construed as propositional attitudes, only happen to creatures that speak a proposition expressing language. Human language makes it possible for humans to, as it were, emulate a BDI machine. But underneath the emulator is something very different, something we share with every living creature with a brain. We certainly seem to be BDI machines to ourselves and to each other. But, of course, we cannot tell, just by ordinary interaction, or by reflecting in our philosophical armchairs, whether the mind is a BDI machine at bottom.

Somewhere in our evolutionary past, we became capable of emulating BDI machines and of using the kind of language we do. These are, we suspect, two sides of the same coin. Speaking a human natural language *is* being a virtual BDI machine. Signing apes are virtual BDI machines as well, though, it seems, much simpler ones than we are. Dennett (1991) was perhaps right to suggest that cognitive consciousness, as opposed to consciousness of the sort involved in perception, bodily sensation and emotion, is fundamentally tied to language. It is not just to each other that we appear as BDI machines; we appear that way to ourselves, for our cognitive introspective awareness is linguistically mediated.

By thinking about thought in this way, we still give BDI a central role in explaining meaning, just as the stored chess program plays a role in explaining your computer's chess moves. However, this perspective on BDI removes its centrality in cognitive science, allowing us to focus more clearly on the primacy of content. Additionally, by construing BDI as a virtual machine, our commitment to truth-conditional semantics for propositional attitudes does not require us to think of the contents of the underlying architecture as having a truth-conditional semantics; rather, we can think of meaning and content as being semantically disjoint.

It has, we think, been a huge mistake to mix up the foundations of Cognitive Science with the problem of intentionality and meaning. Again, propositional attitude psychology is to blame. One thinks that the only things that have intrinsic meaning or aboutness are the propositional attitudes. Everything else is derived, meaningful only by convention. But why should one think that the propositional attitudes are intrinsically meaningful? Well, if you think we are, at bottom, BDI systems, then they have to be, because it cannot be a matter of convention what our mental states are. But if we are not BDI systems at bottom, then it *might* be a matter of convention what the contents of our propositional attitudes are: they inherit their contents from the public language sentences used to express them, and the meanings of those *are* a matter of convention.

Cognitive Science needs to explain all of this. In particular, it needs to explain how a biological brain can emulate a BDI machine. But Cognitive Science does not need anything like the notion of meaning as a primitive. What it needs is an understanding of how information can be acquired, stored and manipulated in a way that gives rise to intelligent and adaptive behavior, including, in the case of humans, and, perhaps some other creatures, the ability to emulate BDI machines and use a propositional language.

6. Non-semantic composition: The assembly of complex visual representations.

As we have seen, there are strong reasons for doubting that complex representations in the brain could be *semantically* composed of source-free primitive constituents as LOT requires. In particular, it appears highly unlikely that indicator signals somehow become uniquely typed and source-independent, allowing them to float free from their roles in indication and play their required roles in complex LOT expressions as primitive terms referring to what they indicate under some set of special content-fixing conditions. How does a radically non LOTish human brain emulate a BDI machine? It seems pretty clear that something like this begins to develop in humans during the first eighteen months of life. But it seems equally clear that the brain is not fundamentally a BDI machine, and, consequently, that most of its adaptive business, including, in our case, emulating a BDI machine, gets done by processing non-propositional representations and indicator signals.

Because the combinatorics of TCS has such a strong hold on our thinking, it is worth pausing to emphasize that TCS is not the only game in town. A different picture emerges of the relation between indication and representation and of the composition of complex representations generally, if we examine the role the sort of indicators discovered by Hubel and Weisel play in the construction of visual images. One account of this is to be found in recent research by David Field and Bruno Olshausen (Field, 1987, 1994; Olshausen and Field, 1996, 1997, 2000).

Natural images contain much statistical structure as well as redundancies (Field, 1987), but early visual processing effectively retains the information present in the visual signal while reducing the redundancies. In the 1950s, Stephen Kuffler (1952) discovered the center-surround structure of retinal ganglion cells' response, and Joseph Atick (1992) showed that this arrangement serves to decorrelate these cells' responses. As Horace Barlow (1961) had suspected, sensory neurons are assembled to maximize the statistical independence of their response. Olshausen and Field recently showed that the same is true of neurons in the primary visual cortex. While Hubel and Wiesel discovered that neurons in the primary visual cortex are sensitive to edges—thus their functional description as edge detectors—they did not know what the functional relevance of this structure was (1988). According to Olshausen and Field, edge detection allows neurons in the primary visual cortex to respond in a maximally independent way to visual signals, thus producing sparsely coded representations of the visual field. They demonstrated this by constructing an algorithm that could identify the minimal set of maximally independent basis functions capable of describing natural images in a way that preserves all the information present in the visual signal. Because natural images typically contain edges, and because there are reliable higher-order correlations (three-point and higher) between pixels along an edge, it turns out that natural images can be fully described as composites of about a hundred such basis functions (see figure 1). Given the statistical structure of natural images in the environment, there was sure to be such a set of functions, but the striking thing is that these basis functions are similar to those Hubel and Wiesel found 40 years earlier: spatially localized and oriented edges. Recently, O'Reilly and Munakata (2000) showed how to train a neural network using conditional principle components analysis to generate a similar set of basis functions.

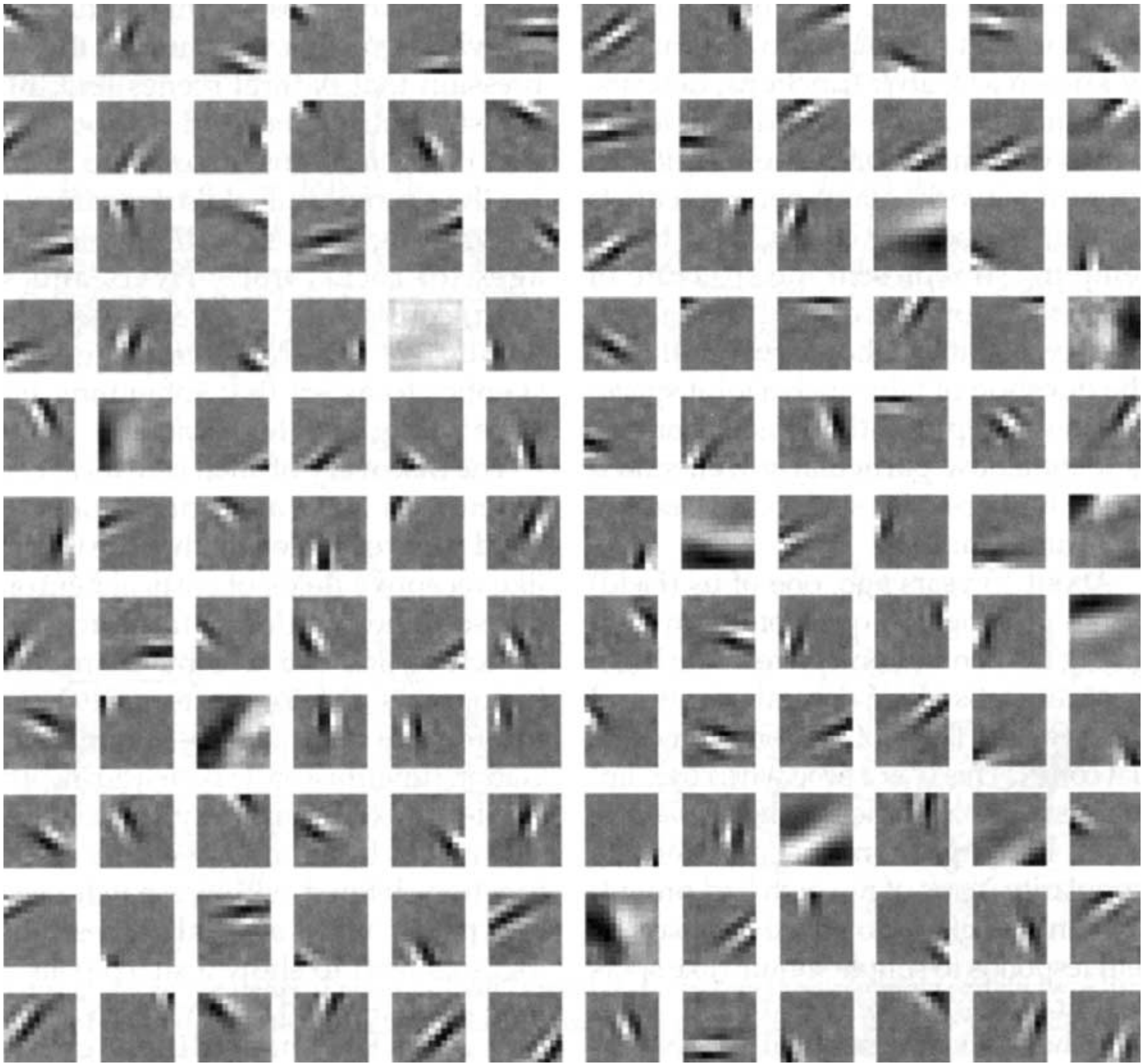

Figure 1: Optimal basis function set to represent natural images (from Olshausen & Field 2000).

To see how visual representations are assembled out of these basis functions, consider a vector of V1 cortical cells connected to the same retinal area via the same small subset of LGN cells. Each cell has a receptive field similar to one in the minimal set of basis functions in Figure 1 above. An incorrect way to understand the assembly of such visual representations is as the activation of a subset of such basis functions whose content is based solely on the information each cell receives from the relevant retinal region. If this were the case, cells in the primary visual cortex would serve to indicate activity in the LGN and, ultimately, of features in the visual field. Indeed, it would be simple to determine what is currently observed if the basis functions were completely independent and the noise factor was known, but neither is true. As a result, many distinct visual representations are compatible with the information present in the visual field. Lewicki and Olshausen (1999) have shown, however, that it is possible to infer a unique visual representation if the system has information about the probability of observed states of affairs and the probability of visual representations given observed states of affairs. Instead of assembling a visual representation from a set of indicator signals alone, the visual system may construct the representation from indicator signals and relevant probabilistic information about the visual environment.

The account that emerges here consists in the assembly of an image from a set of indicator signals that are surprisingly few in number, indicate multi-point correlations between adjacent pixels in the (whitened version of) the input, and whose success as detectors of their proprietary targets is due, to a large extent, to recurrent circuitry that effectively computes a Bayesian function in which the prior probabilities are determined by the features of neighboring areas of the developing image. Most important for our purposes, perhaps, their representational content is semantically disjoint from the image they compose in the same way that pixel values are semantically disjoint from the representational content of a computer graphic. Like maps and scale models, such representations have their content as a consequence of their geometry rather than their origins. Because such representations have the source-independence that indicator signals lack, they are candidates for the disciplined structure sensitive transformations that representations afford. Since such representations literally share the static and dynamic structure of their targets, properties of the targets can be learned by transforming the representations in ways that reflect the ways in which nature constrains the structure of, and structural changes in, the targets. As a result, faces can be aged, objects can be rotated or "zoomed" and three dimensional projections can be computed. None of this is possible in a system in which visual representations are semantically composed from constituents whose contents are determined by their indicator role.

7. Escaping the grip of propositional representation.

It is useful, in this connection, to think about the above example in relation to another relatively well-entrenched form of non-propositional representation, namely cognitive maps (Tolman, 1948). The hypothesis of cognitive maps is designed to explain the striking ability of living creatures such as rats and humans to get from point A to point B, via a novel route, without directions. The hypothesis is based on an analogy with the use of written maps by humans: if you have a map, you can get from point A to point B via a novel route without directions, and there seems to be no other way to do it. You can get specific directions, or you can get a map and get from anywhere to anywhere (in principle). If you can get from anywhere to anywhere via novel routes, and there is no external map in the offing, or, as in the case of rats, no ability to use one, then there must be an internal map. This might be something like a memorized map, but, instead of having its origins in an external map, it is built up from a combination of exploration and background knowledge about such things as which direction you need to go, and how streets or paths or whatever are laid out. Smallville USA, in which streets are laid out in square blocks with numerical names in order running north-south, and, say, presidential names in order running east-west, is a particularly transparent example. If you know this fact about Smallville, and you know your numbers and presidents, and you can see at what intersection you are located when you are at an intersection, you can get from anywhere to anywhere. When we do this sort of thing, we more or less consciously have a kind of street map schema in mind. But there needn't be any reflective consciousness involved. We may do the whole thing unreflectively, as the rats presumably do. However it is done, it seems unavoidable that, if one can get from point A to point B via a novel route without directions, a map, however schematic, and however available or unavailable to introspection, must be there, somehow, underlying the ability.

Cognitive maps are paradigmatic examples of mental representations. To do their causal work in us and in rats, and their explanatory work in Cognitive Science, they do not need any special causal connection with the environment, nor do they need to have any historical properties. These were all rung in to ground *meaning*, not to underwrite their explanatory power. What they must have instead, and all they must have instead, is a structure that is reasonably similar to the route structure of the environment the traveler happens to be in.

It has seemed to many that nothing could count as a cognitive map (or any other representation or indicator signal), unless it is "usable" by the traveler or one of its subsystems. After all, if the map were, say, etched into the inner surface of the rat's skull, it wouldn't do much good. But that is a misunderstanding of the same sort we have been warning against: representations like maps (or indicator signals, for that matter), do not need to be understood or grasped or used by the systems that harbor them to count as contentful. To see this, it suffices to consider the fact that it must be possible to learn (or develop, or evolve) the ability to exploit content one cannot *currently* exploit. Since you cannot learn (or develop, or evolve) the ability to exploit content that isn't there, there must be unexploited content. Indeed, it must be possible for an individual to harbor representations aspects of which that individual cannot even *learn* to exploit, if we are to allow, as we surely must, for the possibility that the species might evolve the ability to exploit that content in the future. For example, all neural network models of learning presuppose that the brain learns to exploit previously unexploited structure in its representations, for the process of weight adjustment made over time makes no sense unless we assume that the representational content of the input pattern remains the same throughout learning. It is precisely such unexploited content in the input patterns that the network is learning to use. But if a network can learn a task, it can evolve the same ability. Neither the learning nor the evolution makes sense if we suppose the representations don't represent unless and until they are thoroughly exploited (Cummins, et al. 2006).

All sorts of creatures harbor cognitive maps. They are not sets of beliefs. They do not have propositional contents. They are not candidates for truth-conditional semantics. They do not have to be understood, or "grasped" or even exploitable by the creatures that harbor them to have the contents they do. They need no causal connections with the environment or historical properties (which, to repeat, was all about fixing reference, anyway) to have the contents they do. They just need to share geometrical properties with the environment. They need, in short, to be reasonably good maps. And maps, while complex, are not built up out of semantic primitives whose meaning combine to yield the content of the complex.

A similar lesson emerges when we think about how content is determined in connectionist networks. Recently, Paul Churchland has characterized the representations that he thinks underwrite such cognitive abilities as recognizing faces and representing grammatical structure as points, regions, or trajectories in neuronal activation spaces, and contrasts such representations with language-like schemes that express propositions (1998).

What are the representational contents of such trajectories and partitioned activation spaces? Consider a network that learns to discriminate families in terms of facial resemblance (1998). The points in such a space are what Churchland calls prototype points. They are centers of attraction, where the clustered points correspond to related family members. These prototype points reflect the way that training the network partitions up the relevant activation space. Furthermore, the underlying geometry is remarkably constant across different networks trained to the same task, including ones with differing input encodings and with differently dimensioned hidden layers. We can think of this geometry as representing an objective structure in the families' face space, and that trained networks discover this and represent it via a structurally similar activation space. According to Churchland, we can think of the prototype points as something like individual concepts in a conceptual space. Recurrent networks pose a different case because they allow for temporally extended representations, and are best conceived in terms of trajectories in activation space rather than points. This approach works nicely for Elman's well known grammar network (1992). For our purposes, the important point is that such contents are not propositional, they are not grounded in whatever causal connections the points may enjoy with distal properties, and they represent in virtue of shared structure.

Those of us interested in what the explanatory primitives of Cognitive Science are should concede that there is, to a first approximation anyway, no meaning in the head. Not because we are externalists, but because meaning isn't what's wanted. There *is* a kind of content in the heads (and other bodily parts, down to the molecules) of intelligent creatures, but it isn't meaning as that is generally understood. Meaning proper, whatever it turns out to be, is an explanandum for Cognitive Science, not an explanans. And for those of us who are worried about the pitfalls of trying to read off the fundamentals of cognition from the human case, meaning should probably be pretty far down on the agenda, except as it arises in connection with human language processing. And there, one must beware of the kind of translation theory of language understanding that makes LOT and BDI seem inevitable.

8. Conclusion

So: what we need at the foundations of Cognitive Science is representation and indication, and we need the concept of unexploited content: content that is there to be exploited, but which the system has not yet learned or evolved to exploit. Content, in short, that is not a function of use.

We do not need psychosemantics as generally conceived. As this is usually conceived, it makes sense only in a BDI + LOT framework. But it is likely doomed even there, because the conventions governing the semantics of the sentences that express the contents of the propositional attitudes are, obviously, just the conventions governing the sentences of a natural language, and these are going to yield meanings—references and truth-conditions—that do not match the non-convention governed mental/neural states that are, or should be, the bread and butter of Cognitive Science. You are not going to find an evolved natural content for every convention governed sentence or its constituents. Our language doesn't reduce to Aristotle's, let alone to an imagined language of thought that is presumed to differ little if at all from the code running in the brains of our Pleistocene ancestors.

What we do need (though not everyone would welcome it) is a scientific explanation of meaning. This we evidently do not have at present. But we think our prospects would improve if we were to conceive the problem as a problem about language and communication, and not as a problem about the contents of psychological states. The trend has been to derive linguistic and communicative meaning from something else, the contents of propositional attitudes. Sentences express propositions because they express *thoughts.* This, by itself, isn't so bad. The trouble comes when we go on to think that *thoughts,* being the cogwheels of the mind, must have their propositional contents intrinsically, non-derivatively. Thoughts must be things having "natural" (i.e., non-conventional) propositional contents. The conventions of language simply link linguistic expressions to intrinsically meaningful thoughts. We are now deeply committed to a BDI theory of the mind, and the LOT that goes with it. The concepts of representation and indication fundamental to the science of cognition get hijacked and pressed into service as grounders for meaning, and this grounding business becomes the goal of the theory of mental content. Linguistic meaning is then conceived as just LOT meaning worn by conventional symbols, and LOT meaning is just, well, functional role, or the role of primitive LOT symbols in detection, or…. We completely lose sight, in philosophy, anyway, of the success that cognitive science has achieved in understanding such things as cognitive maps and edge detectors and visual images/maps. And we make our job seem easier than it is by failing to see the gap between meaning and content. It is a large gap, and it needs to be filled by something other than hand-waving.

References

Atick, J. 1992. "Could information theory provide an ecological theory of sensory processes?" *Network* 3:213-251.

Barlow, H.B. 1961. "Possible principles underlying the transformation of sensory messages." W.A. Rosenblueth (ed.) (1961). *Sensory Communication*. Cambridge, Mass.: MIT Press, pp. 217-234.

Churchland, P. 1998. "Conceptual similarity across sensory and structural diversity: the Lepore/Fodor challenge answered." *Journal of Philosophy* 95: 5-32.

Cummins, R. 2002. "Truth and meaning." In Joseph Keim-Campbell, Michael O'Rourke and David Shier (eds.), *Meaning and Truth: Investigations in Philosophical Semantics*. New York: Seven Bridges Press.

—1996a. "Systematicity." *Journal of Philosophy* 93: 591-614.

—1996b. *Representations, Targets, and Attitudes*. Cambridge, Mass.: MIT Press.

Cummins, R., J. Blackmon, D. Byrd, A. Lee, C. May, M. Roth. 2006. "Representation and unexploited content." In G. McDonald and D. Papineau (eds.), *Teleosemantics*. New York: Oxford University Press.

Cummins, R., P. Poirier. 2004. "Representation and indication." In Phillip Staines and Peter Slezak (eds.), *Representation in Mind*. Elsevier: 21-40.

Cummins, R., J. Blackmon, D. Byrd, P. Poirier, M. Roth, G. Schwarz. 2001. "Systematicity and the cognition of structured domains." *Journal of Philosophy* 98: 167-185.

Davidson, D. 1967. "Truth and meaning." *Synthese* 17: 304-323.

Dennett, D. 1991. *Consciousness Explained*. New York: Little, Brown.

Elman, J. 1992. "Grammatical structure and distributed representations." In S. Davis (ed.), *Connectionism: Theory and Practice*. Vol. 3 of *Vancouver Studies in Cognitive Science*. Oxford: Oxford University Press: 138-194.

Field, D.J. 1987. "Relations between the statistics of natural images and the response properties of cortical cells." *Journal of the Optical Society of America, A*, 4:2379-2394.

—1994. "What is the goal of sensory coding?" *Neural Computation* 6:559-601.

Feldman, J., D. Ballard. 1982. "Connectionist models and their properties." *Cognitive Science* 6: 205-254.

Fodor, J. 1975. *The Language of Thought*. New York: Crowell.

—1990. "Psychosemantics, or where do truth conditions come from?" In W. Lycan (ed.), *Mind and Cognition*. Oxford: Basil Blackwell.

Fodor, J., Z. Pylyshyn. 1988. "Connectionism and cognitive architecture." *Cognition* 28: 3-71.

Haugeland, J. 1985. *Artificial Intelligence: The Very Idea*. Cambridge, Mass.: The MIT Press.

Heller, M. 1990. *The Ontology of Physical Objects*. Cambridge: Cambridge University Press.

Hubel, D. H. and T. N. Wiesel. 1962. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex". *Journal of Physiology*. 160:106-154.

Hubel, D.H. 1988. *Eye, Brain, and Vision*. New York: Scientific American Library.

King, J. 1995. "Structured propositions and complex predicates." *Nous* 29 (4): 516-535.

Kuffler, S. 1952. "Neurons in the retina: Organization, Inhibition and excitatory problems." *Cold Spring Harbor Symposia on Quantitative Biology*. 17: 281-292.

Lewicki, M.S. and B.A. Olshausen. 1999. "A Probabilistic framework for the adaptation and comparison of images codes." *Journal of the Optical Society of America, A*, 16:1587-1601.

Lewis, D. 1986. *On the Plurality of Worlds*. Oxford: Blackwell.

Olshausen, B.A. and D.J. Field. 1996. "Emergence of simple-cell receptive field properties by learning a sparse code for natural images." *Nature* 381:607-609.

—1997. "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vision Research* 37:3311-3325.

—2000. “Vision and the coding of natural images.” *American Scientist* 88:238-245.

O’Reilly R.C. and Y. Munakata. 2000. *Computational Explorations in Cognitive Neuroscience*. Cambridge, Mass.: MIT Press.

Rumelhart, D. 1989. “The architecture of mind: a connectionist approach.” In Michael Posner (ed.), *Foundations of Cognitive Science*. Cambridge, Mass.: The MIT Press.

Sejnowski, T., C. Koch, P.S. Churchland. 1988. “Computational neuroscience.” *Science* 241.

Tarski, A. 1956. “The concept of truth in formalized languages.” In *Logic, Semantics, and Metamathematics*. Oxford: Oxford University Press.

Tolman, E. 1948. “Cognitive maps in rats and men.” *The Psychological Review,* 55(4): 189-208.