Précis of "The World in the Head"

Robert Cummins

Themes and Constraints

The World in the Head (Cummins, 2010) is a collection of sixteen papers spanning over two decades. Six of them are co-authored with various colleagues.[1] Three have not been previously published. All of them are papers in the philosophy of mind. More particularly, they are about representation and psychological explanation. All of these papers are colored by two convictions that, for me, amount to methodological constraints. The first is that the philosophy of mind should be a branch of the philosophy of science, where the sciences in question are psychology, neuroscience and biology, especially evolutionary biology. More particularly, the philosophy of mental representation should, first and last, account for the explanatory role representation plays in the sciences of the mind. The second constraint is really a consequence of the first. It is that neither the semantics of propositional attitude sentences, nor our "intuitions" concerning the application of mental terminology to various real or hypothetical situations, should constrain the science or our attempts to make sense of it.

"Philosophers of mind," I wrote in the Preface to *The World in the Head,* "come in many flavors, but a useful broad sort when thinking about the philosophy of mental representation is the distinction between philosophers of science, whose science was one of the mind sciences, and philosophers of language, whose interest in the verbs of propositional attitude led to an interest in the contents of beliefs" (p. v). I am deeply suspicious of the latter approach because it inevitably leads to what I call *semantic poaching.* There is a tempting line of argument running from the semantic analysis of sentences involving "mental" terminology to conclusions about the functional structure of the

[1] Jim Blackmon, David Byrd, Pierre Poirier, Martin Roth and Georg Schwarz contributed to "Systematicity and the Cognition of Structured Domains." Pierre Poirier contributed to "Representation and Indication." Jim Blackmon, David Byrd, Alexa Lee and Martin Roth contributed to "Representation and Unexploited Content." Martin Roth contributed to "Meaning and Content in Cognitive Science." Denise Dellarosa Cummins contributed to "Biological Preparedness and Evolutionary Explanation," and to "Cognitive Evolutionary Psychology Without Representational Nativism." Names on the papers appeared in alphabetical order. The papers were a group effort in each case.

mind. One starts, for example, with a truth-conditional analysis of belief sentences, and argues, let's say, that 'believes' is a three place relation between a believer, a proposition, and a sentence in LOT (the Language of Thought) that expresses it (e.g., Fodor, 1981). One then notes that some belief attributions are true. Since the truth condition for belief attributions requires a belief relation between a believer, a proposition, and an expression in LOT, it seems that anything to which one can truly attribute beliefs must harbor psychological states with precisely that structure. And, on the assumption that the brain realizes psychological states, we get a conclusion about the functional organization of the brain *without having to do a single experiment!*

This is too much for way to little. No sane epistemology of science will grant a neuroscience license, or a cognitive psychology license, on the basis of truth-conditional semantics and the "truism" that people have beliefs. This issue is discussed explicitly in chapter 5 ("Methodological Reflections on Belief") and underlies the discussion in chapters 3, 4, 6, 7 and 10 (see below).[2]

An early attempt to exhibit the wide gap between belief attribution and what is actually represented by a cognitive system is "Inexplicit Information" (Chapter 6), a paper that elaborates on one of Dennett's examples of unrepresented belief content:

> In a recent conversation with the designer of a chess playing program I heard the following criticism of a rival program: 'It thinks it should get its queen out early.' This ascribes a propositional attitude to the program in a very useful and predictive way, for as the designer went on to say, one can usually count on chasing that queen around the board. But for all the many levels of explicit representation to be found in that program, nowhere is anything roughly synonymous with 'I should get my queen out early' explicitly tokened. The level of analysis to which the designer's remark belongs describes features of the program that are, in an entirely innocent way, emergent properties of the computational processes that have 'engineering reality.' (Dennett, 1978*a*).[3]

The thought here, as it often is with Dennett, is that belief attribution is far more Rylean than the now orthodox truth-conditional treatment of belief can accommodate. My paper identifies and describes the following ways in which we can get information affecting performance without explicit representation:

[2] A more recent discussion of this issue can be found in Cummins and Roth (2011).

[3] The passage needs a small correction: we are not interested in what is represented in the program, but with what representations are created when it is executed. The point is that no representation with the content *get the queen out early* or anything comparable is generated when the program runs.

Control-implicit: An appliance repair person nearly always checks the power supply before doing anything else. It is thus built into the routine that, at any step beyond step one, the power supply has been checked and is ok. The current state of control—that one is beyond step one—carries the information that the power supply is ok.

Domain-implicit: I could give you a set of directions for getting to my house from yours, and another for getting from your house to Paul's, but you could not, without executing them, determine so much as whether Paul and I live in the same place. I could do this by, for example, relying exclusively on turn left, turn right, counting intersections, and counting houses. In such a case, the location of my house just isn't going to be a consequence, in any sense, of premises supplied explicitly. The only way you could use the program to figure out where my house is would be to execute it, either in real space or using a sufficiently detailed map. The information in question is as much in the map or geography as it is in the program; the program is completely domain dependent for its success. Nevertheless, given the terrain, the program does carry the information that my house is at a certain place: if you follow it and wind up at the bus depot, you have every right to complain that I gave you the wrong information.

Rules and Instructions: A chess program might have an instruction that says, in effect, "If it is possible to deploy the Queen, deploy the Queen." If it always opens Pawn to King Four, this will happen on move two of every game. But, of course, the instruction does not occur in any database accessed by the program.[4] This contrasts with the case in which we have a production system that has a production like the instruction above. In that case, we have a system that contains a database of rules it follows when an antecedent is matched.

The moral of all this, for present purposes, is that without some experimentation, or access to system design, you cannot tell whether you are looking at a case of explicit representation. We may very well be attributing beliefs to each other, and rightly so, in cases in which nothing answering to the standard truth conditions for belief sentences is on the horizon. Poachers, of course, operate without the license that experimentally supported analysis is supposed to provide.

[4] Assuming the program is not hardwired, but compiled or interpreted, the compiler or interpreter will have access to the instruction in *its* database. But the chess system will not.

Systematicity

The gap between belief talk and cognitive architecture, together with a suspicion of poaching, jointly motivate the discussion in chapters 3 and 4 ("Systematicity," and "Systematicity and the Cognition of Structured Domains"). The focus there is the argument from the alleged systematicity of thought to the Language of Thought (Fodor and Pylyshyn, 1988; Fodor and McLaughlin, 1990). The core of the argument is captured in the following pair of claims:

(i) Thought is systematic: anyone who can have the thought that *p* can have the thought that *p**, where *p** is a systematic variant of *p*, i.e., a variant of *p* gotten by permutation of constituents in a way that honors syntax.

(ii) The only, or best, explanation of (i) is that the contents of thoughts are represented in a LOT.

The argument for (ii) is straightforward: To have the thought that *p* requires having a representation of *p*. If *p* is represented in LOT, then the resources required to process *p* are sufficient to process *p**, since the syntax of both representations are the same, as are the constituents.

The thing to notice is that (i) actually assumes what the argument is supposed to prove, namely that thoughts have propositional contents with a language-like syntax. This would be relatively harmless if, as Roth and I argue in Chapter 11 ("Meaning and Content in Cognitive Science"), thinking is just talking to oneself in a natural language one understands. But, of course, this is not what defenders of the systematicity argument have in mind. The argument is supposed to be an argument for LOT, understood as, among other things, the scheme translation into which enables understanding of natural language. It is supposed to be the scheme of mental representation that underlies the mind/brain's cognitive capacities generally. So, for the purposes of this argument, it cannot be a datum that thought is systematic.

What *is* patently systematic is language.

(i') Anyone who can understand a sentence *s* can understand a systematic variant of *s*.[5]

If there is an argument here at all, it is that LOT is the only (or best) explanation of (i').

[5] One could quibble about this, and we do. But it isn't the crux of the matter.

Before looking at that argument, it is worth pausing to reflect on why (i), and the question begging it embodies, seems plausible in the first place. The answer is not far to seek: 'Jones thinks that the Eiffel Tower is in Paris' has, one would assume, the same logical form as 'Jones believes that the Eiffel Tower is in Paris', and that, as pointed out above, is bound to tempt poachers to suppose that thoughts are structured like beliefs. The temptation is to poach and read the structure of thoughts off the truth-conditional semantics of thought sentences.

When the poaching is disallowed, the systematicity argument emerges as an argument for LOT as an explanation of systematicity in language understanding. There is something to be said for the idea that cognizing systematic domains should be explained by appeal to representations that share structure with representational targets in those domains (Cummins, 1996). But humans and other animals cognize many systematic domains that are not isomorphic to language, so arguments exactly parallel to one that moves from (i') to (ii) will also give us reason to think that human cognitive architecture requires representations that share structure with music, space, color, as well as representations for special domains, e.g., distance-rate-time problems. Massive representational pluralism, and the accompanying massive cognitive modularity that would likely go with it, are not what advocates of the systematicity argument had in mind. But the alternative is to abandon the kind of argument that leads from the structure of the domain to the structure of its representations in the mind, and adopt the view that everything is structurally encoded, i.e., that the mind/brain utilizes a scheme that encodes structure rather than actually having it. This, of course, is precisely what connectionists such as Paul Smolensky began advocating (1987; 1988; 1991), and what prompted the systematicity argument as a response.

Meaning and Content

We need to distinguish the sort of meaning that is an explanandum for cognitive science—something Roth and I (Chapter 11) *call* meaning—from the sort of meaning that is an explanans in cognitive science—something we don't call meaning at all, but rather content. Cognitive science appeals to two main sorts of things that have contents: representations and indicator signals. 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 in V1. 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. Familiar examples of representations include maps of all kinds, scale models, graphs, diagrams, pictures, holograms, and partitioned activation spaces. Cognitive maps 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.[6]

Though causal and informational theories of 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. 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. 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. 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.

It is dangerous to think of contents as meanings, for this suggests that a theory of content is, or is something that grounds, an account of semantics. 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 truth-conditional combinatorial semantics pioneered by Tarski (1936/56). As a consequence, the idea that mental representations or indicator signals have meanings became

[6] Representations carry no information in the information-theoretic sense of the term, or, rather, the information they carry is irrelevant to their representational content. A structure isomorphic to a map of Chicago is a map of Chicago irrespective of its causal history.

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 Davidsonian Revolution, to the Language of Thought.

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.

If the mind is not, at bottom, a propositional engine, 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? 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 I call forced error (1996). 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 need not have the same semantic content as 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 content of the picture sent, however, is 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. A compelling alternative is to follow Plato in supposing that at least some thinking is talking to oneself. This allows us to have thoughts with the same structure as sentences—a LOT, in fact—but does not require meaning as understood in standard semantics to attach to anything other than natural language expressions. It leaves open the serious question of how a profoundly non-propositional brain can use and understand language, but this is surely a question worth asking, a question that LOT, construed as an hypothesis about the mind/brain's fundamental representational resource, simply begs.

Content and Use

LOT proposes, as we saw above in discussing systematicity, a representational scheme that is structurally arbitrary for everything except language and the structured propositions that were invented to fit it. A theory of content for LOT is, thus, inevitably, a use theory, i.e., a theory that derives the content of a term *t* from the things it is applied to under conditions that guarantee correct application. Non-propositional (non-symbolic) representations such as pictures, models and maps, have their content intrinsically in virtue of their structure. A relatively well-entrenched example of this kind of representation is cognitive maps (Tolman, 1948). 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 for expressions in LOT, not to underwrite the explanatory power of things like cognitive maps.

What they must have instead, and they all must have instead, is a structure that is reasonably similar to the topography 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 that has been warned 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. This is the thesis of chapter 8 ("Representation and Unexploited Content").

Representational Specialization

Chapter 12 ("Representational Specialization") looks specifically at the role of form in structured representations. Every representational scheme presupposes something about its targets. These presuppositions are built into its structure—what Kant called its form. From inside a particular scheme, its presuppositions are synthetic a priori in the sense that counter-examples to them cannot be represented in that scheme. Thus a system, call it Cubic, that represents objects in space by coloring cells in a three dimensional cube—really coloring, not assigning color symbols—cannot represent counter-examples to any of the following:

V1. Every object is colored.
V2. Every object has a determinate size and shape.
V3. No two objects can occupy the same place at the same time.
V4. Every object has a determinate location relative to every other object.
V5. Every object is a determinate distance from every other object.

Of course, Cubic does not represent any of these propositions. Indeed, Cubic cannot represent any propositions at all.[7] Its only representational targets are colored shapes in three-dimensional space. Nevertheless, *we* are in a position to see that these propositions are somehow inevitable for Cubic, even though they are, in fact, all empirical, and all false. We might be tempted to say that Cubic *thinks* that every object is colored, provided we are not thinking that thoughts must be propositional and involve language-like representations.

The thesis I urge is that every representational scheme is like Cubic in being structured in ways that enable representation of some targets and not others. If this is right, then it will follow that no single scheme is going to get everything right, and every scheme is bound to get some things wrong. This shouldn't make us anti-realists, of course. It should make us representational pluralists.

The ability to exploit multiple schemes makes it possible to avoid being trapped by the presuppositions of any particular scheme. Shadowing Kant for purposes of illustration, imagine adding a conceptual (propositional, symbolic) scheme to Cubic's representational repertoire. Call the result P-Cubic. It is a tempting Kantian idea to think of P-Cubic's *experience,* its epistemic access to the world, as consisting of its perceptual representations, and to point out that P-Cubic could never have an experience that directly contradicted any of V1-V5. This would not be exclusively a contingent fact about the external world, but also and correlatively a consequence of the form of P-Cubic's representational resources. V1-V5 would be synthetic for P-Cubic—not conceptual truths—but they would be a priori in the sense that no disconfirming experience is possible for P-Cubic. For P-Cubic, colored shapes in a Cartesian 3-D space would be the a priori form of outer intuition, i.e., of perceptual object representation. The form of P-Cubic's perceptual representations constrain not only what it can accurately perceive, but also, indirectly, what it is rational for it to think is true. P-Cubic could contemplate the negation of, say, V1,

[7] It might *encode* propositions (see below), but I am assuming here that Cubic has no means to exploit such a possibility.

but could never have an experience disconfirming V1. Or so goes one version of the Kantian story and its empiricist predecessors. It is beguiling, but it is fundamentally mistaken.

To see why, notice that P-Cubic's perceptual resources suffice for the confirmation of a simple mechanics. P-Cubic might well *discover* the existence of uncolored objects, their locations and their trajectories, without ever experiencing such objects as such, or even without experiencing them at all. Even a rudimentary naïve mechanics, gleaned from what *is* available in perception, could provide P-Cubic with persuasive evidence of what is not available to perception, e.g., an uncolored, hence imperceptible, wall on a pool table. What makes this possible is precisely the fact that P-Cubic can represent and evaluate propositions that are true of possibilities that it cannot experience as instances of those possibilities, but for which P-Cubic can have persuasive indirect evidence. Because P-Cubic's propositional representations do not have the same contents as P-Cubic's perceptual representations--because its concepts are not *copied* from its percepts, but do apply to the objects represented in those percepts--P-Cubic is in a position infer how things should look on the pool table if there is an invisible quarter circle barrier around one of the pockets. This, more or less, is how we generally find out about the unperceived.

It is tempting to suppose that whatever can be represented at all can be said. But a picture, for example, is not translatable into words. We can describe what we see, and construct images of what is described, but this is not translation however loose. The reason is that sentences and pictures are structured differently, and these differences embody and enforce constraints on what can be accurately represented. What we can do, of course, is encode pictures symbolically—e.g., as a matrix of RGB values. But the resulting encoding will require very different processing than the original. It will interact differently with learning and selection processes, for example. We could, generalizing, encode all of our science, replacing the graphs, models, maps, and so on with encodings drawn from a single scheme. But the result would be largely useless. Because encodings are structurally arbitrary—they do not share structure with the things they encode—similarity (nearness) in encoding space is not related to similarity in target space, and this will make learning an intractable problem. Representational resources are like other kinds of tools: specialization makes for precision and tractability, but only in their proprietary domains. This is why a Swiss army knife doesn't have ten blades or ten corkscrews.

Evolution and Mind

Thinking of the mind in terms of the propositional attitudes encourages what Elman et. al. call representational nativism (1996). The problem, in a nutshell, is that most learning tasks are next to impossible for a propositional engine. As a consequence, the poverty of the stimulus argument looks to apply nearly everywhere: if the goal of learning is to acquire a symbolically specified theory whose tacit application accounts for performance, learning is just not going to happen unless a good deal of the theory to be acquired is already in place. Those theories, moreover, will need to be domain specific, as Cosmides and Tooby (1994) correctly argued. Massive modularity and representational nativism are thus joined at the hip, but as Elman et. al. pointed out, the resulting combination is difficult to reconcile with both the amount of plasticity in the infant and the fundamental similarity of all mammalian brains. Chapters 13 ("Biological Preparedness and Evolutionary Explanation") and 14 ("Cognitive Evolutionary Psychology without Representational Nativism") argue that an approach that assumes powerful learning biases in a neural network architecture can accommodate neural plasticity, while providing a coherent framework for selection of developmental programs that lead to what Pinker (1997) calls a Swiss army knife model of adult cognition. This is what nativism—the kind required by developmental neuroscience and evolutionary psychology—looks like when we abandon a conception of the mind and cognition grounded in the propositional attitudes.

Psychological Explanation

Chapters 15 ("Connectionism and the Rationale Constraint on Cognitive Explanation") and 16 ("'How Does It Work?' vs. 'What Are the Laws?'") are explicitly about psychological explanation, the former about its content, and the latter about its form.

Begin with the content: Cognition is normative—it is a matter of getting things more or less right. Thus, it seems that cognitive capacities must be grounded in a kind of automated epistemology, i.e., cognitive outcomes must be explicable as the exercise of a process that embodies a rational for the task. Yet connectionist systems appear not to embody such rationales. Rationales, as we have been taught to think of them, are propositional and hence symbolically represented. Connectionism thus encourages us to investigate the possibility of non-symbolic rationales.[8]

Now for the form: The term 'effect' is ambiguous in science. Compare the following:

(i) Shrinking polar icecaps is an *effect* of global warming.

(ii) People tend to remember only the thesis of a philosophy paper. This is because philosophers are taught to state their thesis at the beginning of a paper, and at the end. This is an example of the primacy-recency e*ffect*.

Psychology is mostly in the business of discovering and explaining effects in the second sense. *Effects, as I shall call them, are law-like specifications of regularities that characterize special systems such as 16-month-old human infants and adults with frontal lobe damage. *Effects don't explain anything. You cannot explain why someone hears a consonant like the speaking mouth appears to make by appeal to the McGurk *effect. That just *is* the McGurk *effect (McDonald and McGurk, 1978). Knowing the *effect, I can predict that a subject will report the consonant the speaker's mouth appears to make, but prediction, as this example shows, isn't explanation, despite the fact that we sometimes say that a particular *effect specification explains or accounts for the data. The latter is just a potentially misleading way of saying that the data confirm the *effect specification.

*Effects are explananda, and are explained by appeal to the designs—typically functional analyses—of the systems that exhibit them. To explain the McGurk *effect, we don't need a more general or basic set of laws from which to derive the *effect specification; rather, we need to know how the speech perception system works, how it is designed. Nevertheless, *effects are not the primary explananda for psychology. The primary explananda are capacities such as the capacity to learn a natural language or to see depth. These also are to be explained by uncovering the design of the systems that have them. A model of a system's design that is put forward to explain a psychological capacity will predict some *effects and not others. For example, a popular 17th century model of depth perception assumed that two eyes and a fused image were required. But the illusion of depth produced by perspective drawing, or by looking along a railroad track, requires only one eye. Thus the perspective and texture gradient *effects showed that the binocular model couldn't be the whole story.

[8] But see Roth (2005), where it is persuasively argued that "classical" rationales may be implemented in connectionist systems, with the relevant epistemological/logical dependencies implemented as geometric relations rather than as relations of causal dependency.

Consciousness and AI: No resolution from the armchair.

The oldest paper in this collection is a paper on consciousness (Chapter 1, "What is it Like to be a Computer?"). Written in the mid 1970s as a response to John Searle's Chinese Room Argument (1980) but never previously published, the paper uses a version of the famous Hubert-Yoric set-up from Dennett's "Where Am I?" (1978*b*) to argue that one could discover introspectively, from a first person perspective, that one was, in fact, a computer. The set-up is this: Dennett's brain—Yoric—is removed from his body, with which it communicates by wi-fi. A computer—Hubert—receives the same signals from Dennett's body as Yoric does. Hubert is then programmed to be input-output to Yoric.[9] We now suppose that a switch is installed that switches the source of commands to Dennett's body between Yoric and Hubert. My version of the story then proceeds as follows: You are the subject of this experiment, not Dennett. We explain the set-up to you, and hand you the switch. A possible outcome is that flipping the switch makes no more difference to you than flipping a switch from a bin of switches in a hardware store. Should that happen, you would have to say that you know what it is like to be a computer: it is like being you. Indeed, functionally, you have been one all along.

Whether or not the mind is a computational process is, therefore, an empirical question, and *a priori* arguments against computational accounts of consciousness, such as Searle's Chinese Room argument, cannot be probative. The relevant empirical science was a long way off when the paper was written, and remains a long way off. This doesn't mean no one should be studying consciousness. It just means that it should be done with a healthy respect for the fact that we don't have any very clear idea what we are talking about.

I have been told that this is an application of the very appeal to intuitions scouted above and critiqued in Cummins (1999). I don't think so. It is, of course, a thought experiment, but it isn't designed to elicit intuitions, but to demonstrate the in-principle empirical testability of a computationalist theory of mind from a first-person perspective. In this respect, I think it is on all fours with various influential thought experiments in the sciences themselves, e.g., those employed by Galileo and Einstein.

[9] Updating a bit, we can suppose Hubert is a neural network trained to respond like Yoric by using Yoric's responses as a training set.

Conclusion

Mainstream philosophy of mind has been hamstrung by a focus on the propositional attitudes and semantic poaching. An increasing proportion of the philosophy of mind is philosophy of the mind sciences: most philosophers of mind, these days, know a lot of the science, and take it to be their business to explain that science, and to contribute a bit of Ur-science at the frontiers. They are increasingly unsympathetic to the kind of semantic poaching that characterized the height of the theory of content debate. This is not always a welcome development in the rest of philosophy, but, then, neither was the advent of scientifically informed philosophy of physics or biology.[10]

References

Cosmides, L., and Tooby, J. (1994). 'Origins of Domain Specificity: The Evolution of Functional Organization.' In L.A. Hirshfeld and S.A. Gelman (eds.), *Mapping the Mind: Domain Specificity in Cognition and Culture*. Cambridge: Cambridge University Press.

Cummins, R. (1996). *Representations, Targets and Attitudes*. Cambridge, MA: MIT Press.

Cummins, R. (1999). 'Reflections on Reflective Equilibrium'. In Ramsey, W. and M. DePaul (eds.) *The Role of Intuition in Philosophy*. New York: Rowman & Littlefield.

Cummins, R. (2010). *The World in the Head*. New York: Oxford University Press.

Cummins, R., and Roth, M. (2011). 'Intellectualism as Cognitive Science'. In Eva-Maria Jung and Albert Newen (eds), *Knowledge and Representation*. Stanford, CA: CSLI.

Davidson, D. (1967). 'Truth and Meaning'. *Synthese* 17: 304-23.

Dennett, D.C. (1978*a*). 'A Cure for the Common Code'. In *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press.

Dennett, D.C. (1978*b*). 'Where Am I?'. In *Brainstorms: Philosophical Essays on Mind and Psychology*. Cambridge, MA: MIT Press.

Elman, J., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., and Plunkett, K. (1996). *Rethinking Innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Fodor, J. (1981). 'Propositional Attitudes.' In *RePresentations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge, MA: MIT Press.

Fodor, J., Pylyshyn, Z. (1988). 'Connectionism and Cognitive Architecture: A Critical Analysis'. *Cognition* 28: 3-71.

[10] I wish to acknowledge the substantial help I received from Martin Roth in preparing this précis. Some of the prose, and much of the structure, is his.

Fodor, J., McLaughlin, B. (1990). 'Connectionism and the Problem of Systematicity: Why Smolensky's Solution Does Not Work'. *Cognition* 35: 183-204.

Fodor, J. (1990). 'Psychosemantics, or Where Do Truth Conditions Come from?' In W. Lycan (ed.), *Mind and Cognition*. Oxford: Basil Blackwell.

McDonald, G., McGurk, H. (1978). 'Visual Influences on Speech Perception Processes'. *Perception and Psychophysics* 24/3: 253-7.

Pinker, S. (1997). *How The Mind Works*. New York: W.W. Norton & Company, Inc.

Roth, M. (2005). 'Program Execution in Connectionist Networks'. *Mind and Language* 20 (4): 448-467.

Searle, J. (1980). 'Minds, Brains, and Programs'. *Behavioral and Brain Sciences* 3: 417-24.

Smolensky, P. (1987). 'The Constituent Structure of Connectionist Mental States'. *Southern Journal of Philosophy Supplement* 26: 137-60.

Smolensky, P. (1988). 'On the Proper Treatment of Connectionism'. *Behavioral and Brain Sciences* 11: 1-74.

Smolensky, P. (1991). 'Connectionism, Constituency and the Language of Thought'. In B. Loewer and G. Rey (eds.), *Meaning in Mind: Fodor and his Critics*. Oxford: Blackwell.

Tarski, A. (1936/56), 'The Concept of Truth in Formalized Languages'. In J. H. Woodger (ed.), *Logic, Semantics, Metamathematics*. Oxford: Oxford University Press.

Tolman, E. (1948). 'Cognitive Maps in Rats and Men'. *The Psychological Review* 55 (4): 189-208.

Pof. Dr. Robert Cummins,
Department of Philosophy
University of Illinois at Urbana-Champaign
Urbana
United States of America