

The Philosophy of Cognitive Science

Rick Grush
Department of Philosophy
University of California, San Diego

version: 3.1
date: 02.16.01
word count: 8,513

Contents:

0. Introduction

1. Historical background: behaviorism and the cognitive revolution

2. Current topics.

- 2.1 Cognitive Architecture
 - 2.1.1 Classical cognitive science and artificial intelligence
 - 2.1.2 Connectionism
 - 2.1.3 Philosophical issues in cognitive architecture
- 2.2 Content assignment
 - 2.2.1 Informational-causal accounts
 - 2.2.2 Conceptual role/functional role semantics
 - 2.2.3 Biosemantics
 - 2.2.4 Eliminative materialism
- 2.3 Counter-revolution
 - 2.3.1 Embedded cognition
 - 2.3.1 Dynamical systems theory

3. Future directions

4. Bibliography

0. Introduction

Philosophy interfaces with cognitive science in three distinct but related areas. First, there is the usual set of issues that fall under the heading of philosophy of science (explanation, reduction, etc.), applied to the special case of cognitive science. Second, there is the endeavor of taking results from cognitive science as bearing upon traditional

philosophical questions about the mind, such as the nature of mental representation, consciousness, free will, perception, emotions, memory, etc. Third, there is what might be called *theoretical cognitive science*, which is the attempt to construct the foundational theoretical framework and tools needed to get a science of the physical basis of the mind off the ground – a task which naturally has one foot in cognitive science and the other in philosophy. In this article, I will largely ignore concerns of the first two sorts. As for the first, other entries in this volume cover topics such as explanation and reduction in detail. As for the second, little of interest has emerged from this research agenda, at least so far. I will focus on the third topic, the theoretical foundations of cognitive science. This article will begin with a discussion of behaviorism and the cognitive revolution which overturned it, thus setting the agenda for much of what is now philosophy of cognitive science. My discussion of this will focus on three topics: cognitive architecture, content assignment, and the 'counter-revolution' of embodied/embedded cognition and dynamical systems theoretic approaches to cognitive science. I will close with some more broadly cast speculations about future directions in the field.

1. Historical background: behaviorism and the cognitive revolution

It will be useful to start with the behaviorism of the early part of the 20th century. As part of a broad intellectual movement aimed at making inquiry into the nature of the world systematic and reliable, it was believed that a science should admit only publicly observable entities, states, operations, and theoretical entities that could be readily reduced to, or cashed out in terms of, publicly observable entities, states, or operations. Various names were associated with this movement – empiricism, operationalism, verificationism, logical positivism. This movement clashed with much of what was traditionally believed about the mind, which was long thought to be the repository of states and operations which were, almost by definition, private and hence not publicly observable. Since thoughts, feelings, dreams, and the like were not part of what could be observed by the public, so much the worse for them as legitimate objects of scientific inquiry. Rather, behaviorism, as the then kosher psychological theory, officially recognized only stimuli, responses, and entities which could be readily reduced to them, such as strengths of connections between a stimulus and a response. (For *loci classici* of behaviorism, see Watson, 1913, 1938; Skinner, 1938, 1957.)

On the heels of these scientific biases (what has been called *psychological behaviorism*) came philosophical biases (sometimes called *philosophical* or *analytical behaviorism*) to the effect that mental entities were either fictional, or that, contrary to what might be thought, that their status as private was fictional – putative mental states, such as 'expecting rain' being really no more than complex patterns of overt behavior lacking a private mental cause (see Ryle, 1949). This ontological puritanism covered all

mental entities and states, including i) qualitative states, such as impressions of pain or red; ii) contentful states such as thoughts and desires; and iii) mental operations such as reasoning and planning. As such, behaviorism offered answers to the central questions of philosophy and mind, questions concerning the status of the mental and its relation to the physical. The answers were that there was nothing distinctively mental – such terms and expressions as are to be found in everyday discourse about the mind and mental states either fail to refer, or refer to complex sets or patterns of overt physical states and processes. For example, Quine (1960) argues that the 'meanings' of linguistic expressions, rather than being mental or non-physical entities, are (to a good first approximation) simply sets of stimulus conditions.

Behaviorism itself is perhaps most interesting as an object lesson in just how implausible a view can be adopted by scientists and philosophers and receive the status of orthodoxy. It is of historical interest in that it was the context for the cognitive revolution which overturned it and which provides the current theoretical backdrop of most cognitive science, psychology and philosophy of psychology and cognitive science. The cognitive revolution in essence was the realization that any adequate theory of human and animal mentality would need to posit representational states between sensory stimulus and behavioral response – at least for a great many domains of behavior. These states would be theoretical, and not simply reducible to constructs of observables.

The cognitive revolution brought about a renewed legitimacy of talk and theorizing about some types of mental or cognitive states, specifically, content-bearing states such as beliefs, desires, or more generally states which were about things, or carried information about things, and over which operations (such as inference) could be performed so as to solve problems and plan. The other class of mental states rejected by behaviorism – qualitative sense impressions – did not get resurrected by the cognitive revolution. The revolution was brought about primarily by three influences, the first psychological, the second linguistic, and the third mathematical.

First psychology. In the middle third of the 20th century, Edward Tolman and a great many collaborators and followers demonstrated complex maze navigation behavior in rats that resisted explanation in terms of stimulus-response mechanisms, but seemed, rather, to suggest that the rats built up complex representational states, or *cognitive maps*, while exploring, and then used these representational structures to solve novel navigation problems in novel ways (see, e.g. Tolman, 1948). In the early 1970s, O'Keefe and Dostrovsky (1971) found so-called *place cells* in the rat hippocampus – cells which fired when and only when the rat was in a given location. O'Keefe and Dostrovsky appeared to find Tolman's maps in the brain. (A great irony lies in Tolman's work. Early in his career Tolman took his rat maze navigation investigations to be squarely within the behaviorist tradition. He once, in an attempt to sum up his faith in the behaviorist methodology, wrote "everything important in psychology . . . can be investigated in essence through the continued experimental and theoretical analysis of

the determiners of rat behavior at a choice point in a maze" (1938, p. 34). Given the results that his work led to, one can read an unintended element of prophetic truth in these words.)

In linguistics, Noam Chomsky provided powerful arguments to the effect that no purely stimulus-driven mechanisms could possibly learn the structure of natural language, and that rather, language learning seemed to require at least some innate cognitive representational structures which circumscribed possible grammars that were then selected from by exposure to linguistic data (Chomsky 1957, 1959). In fact, Chomsky (1959) was explicitly directed against Skinner's behaviorist theory of language.

In mathematics, the theory of computation developed by Turing (1936) and others provided a theoretical framework for describing how states and processes interposed between input and output might be organized so as to execute a wide range of tasks and solve a wide range of problems. There were no fairly direct neural correlates found of the entities posited by Chomsky's linguistic theories or the burgeoning computational theory of cognition (as there was in the case of the cognitive maps). The framework of McCulloch and Pitts (1943) attempted to show how neuron-like units acting as and- and or- gates, etc., could be arranged so as to carry out complex computations. And while evidence that real neurons behave in this way was not forthcoming, it at least provided some hope for physiological vindication of such theories.

The cognitive revolution made common currency of the view that complex behavior is, in large part at least, controlled by inner representational states. These representations carry contents – they are about things – and they are operated on by processes in such a way that the system can solve problems or make plans. Both of these topics, the nature of the processes or 'cognitive architecture', and the contents carried by representational states, have attracted the bulk of the interest in the philosophy of cognitive science.

2. Current topics.

2.1 Cognitive Architecture

Perhaps the first philosophical issue broached by the cognitive revolution was the issue of architecture. Now that inner representational states were given a new lease, the questions of what processes operated on them and what they were came to the fore. The following two sections briefly discuss the two primary trends on this topic in the last three or so decades.

2.1.1 'Classical' Cognitive Science and Artificial Intelligence

The rise of computer science made available a way of thinking about the mind which has had great influence. The idea was that the mind is like a program, and the brain is the hardware (or 'wetware') on which this program runs (see, e.g. Turing 1950; Newell and Simon 1976). The computer model provides an architecture according to which the states of the cognitive system are, in the first instance, representational states with conceptual contents corresponding to entities like names, predicates, quantifiers, etc., in natural language (Fodor 1975). Combinations of these yield representations with propositional contents. And the processes which operate over these representations are primarily inferential, and learning conceived of as a matter of hypothesis formulation and testing.

This idea has a number of *prima facie* advantages. First, it renders the mind ontologically unmythical, for the mind is merely a certain *functional organization* of matter. Second, it seems to secure a lasting role for psychology in the face of threats to the effect that neuroscience will tell us all we need to know about the mind, and at the same time to tell us what the correct tools are for theorizing about the mind. This is because if the mind is like a computer program, then the brain is more or less irrelevant to understanding it. In the same way that one can learn all about, e.g., a certain word processing program regardless of what kind of computer it happens to be running on (amount of memory, type of processor, operating system, etc.), so too, the details of the mind are independent of implementation. Studying a computer won't tell you anything about the programs one might run on it. Rather, we study the input-output operation of the mind, how it behaves when it breaks down in various ways, and on the basis of this we learn about the program that the brain is running.

This trend in cognitive psychology and computer science ushered in a trend in philosophy of mind: *functionalism*. According to functionalism, the mind was not some mysterious entity, but was merely a functional organization of matter (see, e.g. various of the essays in Putnam 1975). Not only did functionalism supply the above-mentioned 'software' theory of mind's contentful states such as beliefs, but it also provided tools to give an account of qualitative states – something that the development of the computer model of the mind, which was functionalism's inspiration, was quite unconcerned with. The idea was that a qualitative state, like the state of being in pain, was merely a functionally specified state of the mind, a state with the right sorts of connections to input, output, and other interposed states. And not only qualitative states, but content-bearing states could, it was thought, also be defined in functional terms. This approach to assigning content to cognitive states will be discussed in section 2.2.3.

So the hope was that an account, inspired by computer science, of the mind and its states – both qualitative and content-bearing – would finally solve the perplexing problems of mind that had baffled philosophers for so long.

One of the problems that immediately beset functionalist accounts of mentality was the observation that if true, then anything which had the correct functional organization would be a mind and be in qualitative and contentful states – anything, including big arrangements of bottles connected by string, tinkertoy constructions, or large water-pipe and valve systems. The proposal that if one arranges a big collection of cans on strings in the right way that the whole mess would feel pain seemed to many philosophers to constitute a *reductio* of the position (for this and other criticisms, see Block, 1978). Another well-known objection comes from John Searle (1980) who argues that a computer program, or an appropriately programmed computer, designed to process natural language – that is, something with the right functional organization – is not sufficient for really understanding language, since someone could manually run through the program and successfully process the linguistic input while having no understanding at all of the language in question. The conclusion reached is that genuine human understanding is not, in fact, just a matter of our mental implementation of the right program; the mind is not just a functional organization of matter.

2.1.2 Connectionism

Connectionism as a method of solving problems and as a theoretical stance in cognitive science has been around in one form or another at least since the middle of the 20th century. But it was clearly the publication of McClelland and Rumelhart's *Parallel Distributed Processing* volumes (McClelland and Rumelhart 1986; for the philosophical *locus classicus*, see Churchland 1989) that thrust the framework into the spotlight in philosophy, cognitive science and neuroscience. The basic idea of connectionism (I here give a brief description of only one, but perhaps the best known, connectionist scheme) is to process information by representing it numerically (as a set of numerical values, aka a *vector*) and passing it through sets of interconnected units in parallel – in particular through the web of connections between these units. In effect, it takes a vector as input, pushes it through a matrix that represents the weights of the connections, and one gets as output another vector. By changing the efficacy of the connections that make up the web, the system can be configured so as to implement a broad range of functions, manifested here as vector-vector mappings. Furthermore, simple schemes exist for getting such systems to learn how to solve problems by trial and error, as opposed to the need for explicit hand-programming present in traditional artificial intelligence.

In addition to the learning aspects of connectionist nets, one of the advantages often claimed by proponents of connectionism is *biological plausibility*. The claim is that the connectionist units function roughly like neurons, the connections between them are analogous to axons and dendrites, and the connection weights are analogous to the efficacy one neuron has in making another fire. These broad analogies aside, assessing

the accuracy of the claim to biological plausibility is not so straight-forward. Many connectionist networks are simply not candidates for biological plausibility at all for any number of reasons. For instance, some assign entire sentential contents to single units, and it is implausible that a single neuron has a propositional content associated with its activity. To take one more example, the most powerful learning algorithms, including back-propagation, by which these nets learn require that signals travel two ways along the same connection – and this seems not to be something that can happen in biological neurons.

On the other hand, other connectionist models have a high degree of biological plausibility, either because they employ learning algorithms that need only mechanisms which real neurons are known to exhibit, or because they are specifically designed to mimic some known neural system, or (often) both.

The modeling successes of connectionism have been impressive, but not complete. Assessing overall merits is difficult because of the range of models and applications in both connectionist and traditional AI modeling, but to a first approximation traditional models do much better at so-called high-level processes, such as planning, reasoning, and language processing, while connectionist models do much better at so-called low-level processes, such as perception and motor control.

2.1.3 Philosophical Issues in Cognitive Architecture

The two hottest philosophical topics in the 90s concerning cognitive architecture centered on language and putatively language-like cognitive states. The first was a revisiting of an issue that was first broached in the debate with the behaviorists – the ability to learn to process certain kinds of linguistic structures on the basis of exposure to linguistic data. It was claimed that since connectionist schemes learned via exposure to data, that they would be subject to the same sorts of limitations that killed off the behaviorists – namely an inability to account for the linguistic competence we in fact have. One prominent example is the ability to process dependency relations that span a dependent clause, such as "The boy who likes the girls runs away" in which the singular 'runs' goes with the singular 'boy', even though it is next to the plural 'girls'. In one of the most cited papers in psycholinguistics, Elman (1991) managed to train a connectionist network to successfully process such embedded clauses, casting into doubt the arguments to the effect that mere exposure to data would be insufficient. The status of this debate is difficult to assess, though, as Elman's model still managed only a rather modest task, and it is not at all clear that similar connectionist models would be able to account for more complex patterns. The jury is out.

The second issue, first voiced by Fodor and Pylyshyn (1988), was the systematicity of cognition. It is an empirical fact, they claimed, that any creature able to entertain the thought *Rab* will ipso facto be able to entertain the thought *Rba* – for instance, being

able to think 'John loves Mary' implies the ability to think 'Mary loves John.' You simply don't, the argument claims, have any cognitive systems that could have thoughts of the first sort without the ability to have thoughts of the second. Given this, they argue, the cognitive architecture must be comprised of symbols that can be recombined in ways analogous to the names and predicates of first order predicate logic. A number of points might be questioned, such as the initial assumption to the effect that this systematicity is actually an empirical fact (but this seems not too implausible), and the inference from recombining to an architecture defined over syntactic tokens analogous to first order predicate logic predicates and names. This inference is surely shaky, as there are kinds of structure other than logical structure. (See Smolensky 1988 for a defense of connectionism.)

2.2 Content assignment

The second of the major topics in the philosophy of cognitive science is content assignment. This question is much more a philosophical enterprise than questions of cognitive architecture, at least judging by the people who publish in the area. The problem is this: We know that there are representational states (this is the core of the cognitive revolution), and that the vehicles of these representational states are presumably neural states. But what is it in virtue of which these physical states carry the content they do?

2.2.1 Informational, causal, covariational accounts

The first answer to this question we will consider is the one often implicitly assumed by most people working in cognitive science and neuroscience. On this view, a physical state P means or represents some content C (a is F , say) iff P co-varies with C (the F -ness of a), and hence carries information about C . For instance, neuroscientists believe themselves to be finding cells which represent faces or specific shapes when they record the activity of such cells and find some which fire strongly when, and only when, the stimulus (face, shape, color, etc.) is present in the visual field. Since a typical situation in which two things covary is when one of them is the exclusive cause of the other, the three descriptions *informational*, *covariational*, *causal*, are closely related, though they are not synonymous. (See Dretske 1981, 1988; Fodor 1987, 1990)

So the basic idea is that neural or cognitive states represent those things that cause them. This seems innocent enough, but problems arise almost immediately. One is the problem of distality – it is true that a certain shaped stimulus in my visual field will cause a certain set of neurons to fire in my visual cortex. But it is also true that in this case it is a pattern of ganglion cell activity in my retina that is causing that pattern of activity in my

visual cortex. It is also the case that the experimenter who pushed the button making the shape appear on the screen in front of me caused those cells in my visual cortex to fire. The causal chain back from those cells firing in my brain is continuous and long, and it is not clear how we can single out one element in that chain as being the one that determines content.

Another problem is the disjunction problem (there are actually a few different things that go under the heading of the disjunction problem, but I will discuss only one). Suppose that my brain is such that when a horse is in front of me, a certain cell fires. We might say then, if we can solve the distality problem, that the cell means 'horse'. But now suppose that, on a dark night or in fog, a cow is in front of me, and this cell fires. We might think that I misidentified the cow as a horse. But on the causal account, since this neural firing can be caused by either horses or cows, it would have a disjunctive meaning: 'horse or cow'. And hence error is impossible. I will not, on this account, have misidentified the horse as a cow, but correctly identified the cow as a member of the disjunctive type *horse or cow*.

These problems have been addressed, with questionable success, by a variety of means – these include, but are not limited to, appeal to ideal conditions or learning conditions, in order to determine which causes are the ones that really set the content and which are spurious.

2.2.2 Functional role/conceptual role semantics

The idea here, inspired by the functionalist accounts of mind and mental states, is that a state's content, or meaning, is its *conceptual role* (see Field, 1977; Block, 1987). For example, a state, call it #, which is such that, when it interacts with states whose meaning is 5 and 7, the state meaning 12 is reliably produced, and when interacting with states meaning 2 and 6 a state meaning 8 is reliably produced, would have the functional role of *addition*, and hence would mean *addition*. Thus, the state's meaning as an addition operator is supplied by the function that that state has in the system. Of course, the same is true for all such states, including the ones just now labeled as meaning 5 and 7, etc. So in fact, the functional-role meaning of all the states of a system are co-determined in a holistic manner. In its basics, conceptual role accounts are similar to functional role accounts, but place more emphasis on roles in inferences specifically rather than any functional relations.

A major problem for such accounts is that there were purported to be proofs to the effect that for any finite functional system, there would be an infinite number of incompatible yet internally consistent interpretations of its states (I say purported, because the proofs tell us about very delimited types of system, and it is not clear that all functional role type systems are such that these proofs apply to them). So to take the

example above to the next level of sophistication, if the states '\$ # %' yield '!', while '* # @' yield '&', then perhaps \$ means 5, # means *addition*, % means 7, and ! means 12, while * means 2, @ means 6 and & means 8. Alternately, # could mean *multiplication*, while ! means 35 and & means 12. That is, one will always be able to find an infinite number of interpretations for all the states in a system which are internally consistent, but which are inconsistent with each other. Functional role, or so it seemed, did not determine a (single, determinate) meaning for functional states after all.

One attempted rescue maneuver was to combine the functional role and causal accounts (so-called *two-factor theories*), by allowing causal links between some of the system's states and objects or properties in the world to fix the interpretation of these states in such a way as to anchor the interpretation of the other interposed states. So the idea is that if the state \$ is reliably produced when and only when exactly three objects are in view, then that state will mean 3, and if some other state is reliably produced when and only when a horse is in view, that that state means *horse*. With the meaning of many such states fixed, it will be possible to eliminate many or all of the alternate interpretations and fix just one.

The hope is that by combining functional role semantics and causal accounts, it might be possible to solve the alternate interpretation problem faced by the former, as well as the distality and disjunction problem faced by the latter (the state's functional role will determine a content as being just one of the items along the causal chain, or as being just one of the disjunctively sufficient causes, etc.). The stratagem of going two-factor, or more generally of including as relevant factors things outside the cognitive system proper, is also aimed at solving other problems, such as the fact that in different contexts, states with the same conceptual role might differ in referent.

2.2.3 Biosemantics

A theory of content that is currently very popular (indeed, the spirit of her proposal is now embraced by the original major exponent of the informational theory, Dretske) was first introduced by Ruth Garrett Millikan (1984, 1989). Her biosemantic proposal is that we can fix the content of a state by appeal to the evolutionary history of the mechanisms that support that state. The hope is that this can solve the problems facing the bare causal/informational accounts. The idea is that a neural state means C (even though it might be caused by C or D or E), if the reason for that state's evolutionary selection (or more adequately, for the selection of the mechanisms which support that state and its operation) is that it carried information about C. So for example, while it might be the case that either flies or random retinal ganglion cell firings can get the neurons in the frog's brain that control the tongue to become active, the explanation for why that mechanism was selected would need to appeal to flies, and not to random retinal ganglion cell firings. We would not have explained why that mechanism was selected for if we mentioned that it became active during random retinal activity, but we

would explain why it was selected by appealing to the fact that it carried information about flies.

Despite its current popularity, biosemantics has been the subject of a number of criticisms, including the charge that it depends on questionable evolutionary explanations. Another objection is that biosemantics entails that if there is some natural biological mechanism that can represent C, and we construct an exact physical duplicate, that the duplicate will not be able to represent C. This flies in the face of most cognitive neuroscience, which assumes that an explanation for the representational properties of the brain are a function of its physical constitution. If biosemantics is correct, then only those mechanisms which came about through the right sort of evolutionary processes represent anything. The objection is a serious one and often misunderstood. Consider the following analogy. Structural engineering claims that the weight-supporting properties of a bridge are a function of its physical constitution – the materials involved and their configuration. It doesn't matter if the bridge was built by Smith or Jones: if they have the same physical parts in the same configuration, their weight bearing properties will be the same. It would be an odd sort of claim, one clearly incompatible with what is known in physics, to maintain that Smith and Jones could build physically identical bridges, but that Smith's would carry a large load and Jones' would crumble immediately – as though something magical, beyond the explanatory reach of physics, is transmitted through Smith's fingertips. This is precisely what biosemantics wants us to believe about the representational properties of the brain. Somehow, genuine representational properties are like some mysterious ether, without physical effect (any physical effects could be duplicated without the aid of evolution, after all), that presumably moves with DNA. The objection is not fatal – one could after all just bite the bullet and hold that the physical constitution of the brain does not determine its representational properties –, but the objection shows that there is a serious tension between biosemantics and materialism as normally conceived.

2.2.4 Eliminative Materialism

For some sort of completeness I will now discuss eliminative materialism (EM). This position is often misunderstood to be one which argues against there being any contents at all, claiming that notions of content and representation are merely entities posited by a bad theory of mentality (though perhaps one proponent of EM, Stephen Stich (1983), has defended a position which is fairly close to this). For the most part, those philosophers who identify themselves as eliminative materialists, such as Paul Churchland (1981) and Richard Rorty (1965), have only held that certain kinds of mental contents or mental states are fictional, not that the notion of content or representation or mental state are ill-conceived *tout court*. For instance, Rorty argued against the idea that there were anything like essentially private inner sensations which posed insurmountable obstacles to the explanation of behavior in physical terms. But he never argued against the notion of a mental states *simpliciter*. And Churchland's

eliminative materialism is directed not only at such things as private mental qualia, but also against representations with propositional content, such as beliefs. He does not argue against the notion of content *per se*, and in fact has provided a number of positive views on what non-propositional content is and how it is carried by physical states.

2.3 Counter-revolution

The 1990s witnessed a resurgence of what might be called a counter-revolution to the cognitive revolution. Though there has always been resistance to various of the dogmas of the cognitive revolution, this resistance never became a serious challenge to the orthodoxy. One notable example of this movement from the 1960s is the work of J.J. Gibson (1966), whose theory of ecological perception attempted to show how much of what might have been thought to require sophisticated information processing and memory in the cognitive system could in fact be carried out by simpler mechanisms which i) exploited information made available in the environment by various invariances and ii) were tuned to organismically relevant affordances.

Though Gibson's work was widely read, it did not have the systematic effect on cognitive science that it might have. It did, however, remain salient enough to exert a heavy influence on the current counter-revolution, and make Gibson one of its heroes. This current trend is perhaps most centrally expressed in the two related movements of embedded/embodied cognition (E/E), and dynamical systems theoretic approach to cognition (DST).

2.3.1 Embedded/embodied cognition

This movement starts by providing a caricature of the traditional view of cognition. According to this caricature, organisms have sense organs that act as transducers, turning peripheral sensory information into symbols which are passed to a central processor. This central processor then manipulates these symbols together with symbols from stored data structures, and forms a plan or settles on some solution to a problem. At this point the central processor sends a bolus of symbols to output transducers, which control effectors so as to produce some sort of movement or other effect on the body or environment.

Proponents of the E/E movement then argue for, and provide examples to support, the idea that many problems can be solved by simple non-representational mechanisms operating in embodied interaction in a structured environment in which the organism is embedded. For example, rather than maintain sophisticated cognitive maps of its

environment for use in navigation, a bee might simply have mechanisms which guide it in certain directions with respect to the clearly visible sun. In conjunction with a simple internal clock, such a humble mechanism can be very powerful, and solve many navigation and homing problems that might have otherwise been thought to require sophisticated internal representational structures.

One of the *loci classici* of this movement is Brooks (1990; see also Beer, 1995; Clark, 1997), in which he describes two robots, Alan and Herbert, which have the task of tooling around the hallways of the lab looking for empty soda cans, and when finding one, clearing it away. They have, however, no powerful central processor which takes in symbolic representations of sensor data and then plans routes or executes can collecting maneuvers. Rather, the robots have a 'subsumption architecture' in which the bulk of the work is done by a number of independent systems with close links to their own sensors and with little or most often no manipulation of representations. The simple independent systems often interact closely with the world itself rather than with representations of it – prompting the slogan that 'the world is its own best representation'. These robots are claimed to execute their task in a manner much more robust than that of other robots using more traditional methods.

2.3.2 Dynamical systems theory

At about the same time the digital-computer-inspired cognitive revolution got going, one of the contenders in the game of cognitive architecture was the cybernetics camp (Ashby, 1952; Weiner, 1948; for the contemporary revival, see Port and van Gelder, 1995). These researchers were very much inspired by mathematical and technological advances in control theory and dynamical systems theory, one of whose main applications was the autonomous control of vehicles and guided weapons systems (the term 'cybernetics' derives from the Greek term for the pilot of a ship). Simple feedback control systems were the prime theoretical tool. To see the appeal, note that a thermostat (a feedback controller) does a very good job of regulating the temperature in a room without any sophisticated inner representations about the thermal properties of the room or the power of the heating and cooling systems. Rather, it simply subtracts a current measure of the temperature from a goal temperature, and does one of three actions depending on whether the result is positive, negative, or zero. Similarly, an autopilot can keep a plane flying straight by making simple comparisons between a few numerically specified goal values, and a few instrument readings, and executing one of a small number of actions based on the mismatch, if any.

The cyberneticists' tools of choice for describing these feedback control mechanisms was the growing mathematical apparatus of dynamical systems theory. Dynamical systems theory is a mathematical apparatus for representing systems and their evolution over time. The systems is represented by a set of state variables. The set of

state variables establishes a state space: an abstract space in which each point represents one possible state of the system, and the set of all points represents all possible states of the system. The rules of evolution for the system (the *dynamic*) specify how the system will evolve in time – that is, which point it will move to as a function of which point it is at now. The dynamic thus establishes a set of paths through state space (trajectories) that the system will traverse. Dynamical systems theory supplies tools for discussing such systems and their behavior over time.

Note that feedback control mechanisms work because they are in continuous interaction with (i.e. they get continuous feedback from) the environment in which they are embedded. This continual feedback can, in many situations, make complicated internal mechanisms unnecessary. This is the conceptual link between feedback control and the E/E movement. The connection to the DST movement is simply the fact that the tools of dynamical systems theory are well suited to describing feedback control systems. In fact, almost all of the examples used by proponents of DST are feedback control systems – a small subclass of the possible dynamical systems.

In any case, to the extent that feedback control mechanisms can solve complex problems, two things seem to follow. First, symbolic systems of the sort posited by Newell and Simon-inspired artificial intelligence are in fact not necessary for solving such problems. Second, closely coupled interaction between the agent's body and the environment may, contrary to the encapsulated central processor view of classical AI, be needed to solve many problems.

The most salient feature of this debate is the extent to which the two sides talk past each other, because each of the two sides adopts a different paradigm for what counts as 'cognitive'. Classicists take reasoning, playing chess, and processing language as paradigm cases, and the E/E and DST camps take sensorimotor tasks as central. Each research program, to all outward appearances, seems to do the better job of accounting for its preferred 'cognitive' tasks.

2.3.3 Objections to the counter-revolution

Assessing the merit of the counter-revolution is not straight-forward, but a first gloss on what is right and wrong about it is this: the counter-revolution is right that representations understood as symbols structured along something like first-order predicate logic and manipulated via something like inference rules probably have very limited application in understanding the various aspects of cognition; but the counter-revolution is quite wrong to try to exorcise the notion of representation altogether. Representation is here to stay. How to correctly understand its various manifestations is

what is up for grabs. I will say more about this in the final section, but for now, some objections to the counter-revolution.

The most serious objection is that the bulk of the abilities studied by cognitive science are abilities executed, or executable, without any dynamic, embedded interactions with the environment. The Watt governor (a favorite example of the DST camp), or Brooks' robots, do nothing if not hooked up in the right way with the right things to interact with. Human cognition, on the other hand, seems to chug along fine in silent contemplation. Chess players try out moves in their heads before trying them on the board, people plan routes to drive home before getting in their car, and people dream of France while silent and motionless in their beds at night. All of these things, and many more, require a representational story for their explanation. How these representations are best understood is of course another story. (For an account that combines the core insights of the DST/EE camp while providing for genuine representations, see Grush 1995, 1997.)

3. Future directions

As this sections requires guesses as to what the future holds, it will of course reflect my biases to a degree even greater than the previous sections. Reader beware.

First, some bare predictions as to how current issues are likely to resolve. For starters, the cognitive revolution is here to stay because it is, in its essentials, right. Insofar as it pushes anti-representationalism, the counter-revolution is misguided and will be washed out in time. On the other hand, the counter-revolution is right to stress the role of the body, the environment, and real-time activity in cognition and problem solving. The solution will involve rethinking the nature of cognition and representation in such a way as to move away from the idea of the disembodied central processor and toward the idea of representations and processes that are more closely tied to agent-environment interactions, but without denouncing representations. The tools of dynamical systems theory are unlikely to have much lasting impact on our understanding of central features of cognition such as language, thought and reasoning.

As far as topics in cognitive architecture go, it is likely that different tasks such as memory, perception, reasoning, will turn out to involve different sorts of processing at an architectural level. But for the more central cognitive systems, they most certainly involve structured representations which can recombine in ways at least analogous to the behavior of the symbols posited by classical computational cognitive science. However, these representations are most likely operated on by processes which are not at all well-described by the formalism of first-order predicate logic or its extensions,

exactly because cognitive representations will be found to have their structural features because of their semantic features – syntax being merely a shadow cast by semantics.

Along these lines, there will be growing appreciation for the theories of cognitive linguistics, and especially Ronald Langacker's Cognitive Grammar (see also Talmy, 2000; Fauconnier et al 1996), not only for their provision of the correct tools for understanding human linguistic competence, but also because of the light they shed on cognitive representation and processing in general. Cognitive Grammar takes the view that linguistic expressions are meaningful in virtue of their parasitism on the meaningfulness of representational structures whose home is in perception and action. I will say a bit more about this below.

As for content assignment, two points can be made with some confidence. The first is that for the purposes of cognitive science and neuroscience, at least for the foreseeable future, the off-the-shelf causal/informational theories will be fine. The second is that a philosophically adequate account of the content-bearing properties of physical/neural states is a long way off, and will almost certainly have no resemblance to any causal/informational or biosemantic account. As to what form the eventual correct account will take, only a few general features can be discerned. It must be an account that explains how a given arrangement of physical/neural entities can create its own representational endowment or potential, without recourse to external objects or states of affairs (such as evolutionary history or causal antecedents), or outside interpreters. And relatedly, it must be an account which explains the objectivity of contents – that is, which does not simply take it as unproblematic that the content carried by a cognitive system is of objects and states of affairs which are understood to be independent of their being represented (this will be discussed more below).

Now to some predictions of a positive nature. A growing trend among the newer generation of those who identify themselves as philosophers of cognitive science is a growing appreciation for traditional topics in philosophy, especially topics whose genesis was in Kant, but which have more current expression in the work of philosophers such as Peter Strawson and Gareth Evans. A major goal of this trend is to understand representational structures – such as the representation of space (both egocentric and allocentric), of oneself as an agent in space, and of objects as permanent denizens of the world which are represented as being independent of being represented; that is, as objective. Such representational schemes take as basic the cognitive system's representation of itself as an embodied agent actively engaged in an environment populated with temporally extended objective processes. For example, Bermudez (1998) provides an account of self-consciousness (understood as self-representation) that relies on non-conceptual representational machinery whose home is in perception and action; Metzinger (1993) provides an account of consciousness and qualia that relies on the brain's own self-representation; Grush (2000, in preparation)

provides an account of the neural mechanisms of spatial representation, self-representation and objectivity.

Such accounts, if they succeed, could represent a radical rethinking of the nature of representation. The new notion will maintain that truth-conditions are crucial for representational content, but that rather than taking truth-conditions to be the satisfaction of an n -place predicate by n objects, truth-conditions will be reconceived as located, structured, objective processes. *Located* in the sense that they are conceived as being spatially and temporally related to the conceiver (unlocated processes being a degenerate case); *structured* in the sense that the processes are conceived as involving the interaction of multiple entities (objects and properties being a limiting case); *objective* in the sense that these processes are conceived as being independent entities in a world that is independent of the representer (see Strawson, 1959, ch. 2); and *processes* being temporally extended (temporally punctate states of affairs, such as the possession of a property by an object, being again a degenerate case). Such representational structures will be largely learned from actual embodied interaction in an environment, and will reflect this perspective in their content – that is, it will be the environment-as-interacted-with that is represented in the first instance, not the environment as it is in itself, whatever that might mean.

In addition to this trend, a number of others can be discerned. I will mention only two. First, the study of emotions, their nature, their connection to reasoning, and their neural substrate, is a growing area of research (see Damasio 1994, Griffiths 1997). Second, the nature of the 'theory of mind' issue, while it has been the focus of attention for some small groups, appears to be taking on more currency, in philosophy, psychology and neuroscience generally (see e.g. Caruthers et al 1996; its original home is in developmental psychology). The 'theory of mind' phenomenon can be illustrated as follows. Imagine a child watching a person A hide something S in the kitchen and then leave. B then comes in, and moves S from the kitchen to the living room. A returns, and the child is asked where A will look for S. Before a certain age, children answer that A will look in the living room. This is, after all, where S is. After a certain age, children realize that A's actions are mediated by A's representation of the environment and not the environment itself – they realize that other people have representational minds. This issue is of import for a number of philosophical questions (including the philosophy of mind and action), psychology, and for understanding a number of phenomena, such as autism and various psychopathologies.

4. Bibliography

Ashby, W.R. (1952). *Design for a brain*. New York: Wiley.

Beer, Randall (1995). A dynamical systems perspective on agent-environment interaction. *Artificial Intelligence* 72:173-215.

Bermudez, Jose Luis (1998). The paradox of self-consciousness. Cambridge, MA: MIT Press.

Block, Ned (1978). Troubles with functionalism. In Wade Savage, ed., *Minnesota Studies in the Philosophy of Science: Volume IX*. Minneapolis: University of Minnesota Press.

Block, N. (1987). Functional Role and Truth Conditions. *Proceedings of the Aristotelian Society* LXI:157-181.

Brooks, Rodney (1991). Intelligence without representation. *Artificial Intelligence* 47:139-159.

Carruthers, Peter, and Peter Smith (eds., 1996). *Theories of theories of mind*. Cambridge: Cambridge University Press.

Chomsky, Noam (1957). *Syntactic structures*. Paris: Mouton.

Chomsky, Noam (1959) review of Skinner's *Verbal Behavior*, in *Language* 35:26-58.

Churchland, Paul (1981). Eliminative materialism and the propositional attitudes. *Journal of Philosophy* 78(2).

Churchland, Paul (1989). *A neurocomputational perspective: the nature of mind and the structure of science*. Cambridge MA: MIT Press.

Clark, Andy (1997). Being there: putting brain, body and world back together again. Cambridge MA: MIT Press.

Damasio, Antonio (1994). *Descartes' Error*. New York: Avon Books.

Dretske, Fred (1981). Knowledge and the flow of information. Cambridge MA: MIT Press.

Dretske, Fred (1988). *Explaining behavior*. Cambridge MA: MIT Press.

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7:195-225.

Fauconnier, Gilles, Eve Sweetser and Eileen Smith Sweet (eds, 1996). *Spaces, worlds and grammar*. University of Chicago Press.

Field, H. 1977, "Logic, Meaning and conceptual role" *Journal of Philosophy* 69, 379-408.

Fodor, Jerry (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, Jerry (1987). *Psychosemantics*. Cambridge MA: MIT Press.

Fodor, Jerry (1990). *A theory of content and other essays*. Cambridge MA: MIT Press.

Fodor, Jerry and Zenon Pylyshyn (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition* 28:3-71.

Gibson, J.J. (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.

Griffiths, Paul (1997). *What the emotions really are*. University of Chicago Press.

Grush, Rick (1995). Emulation and cognition. PhD Thesis, UC San Diego. UMI.

Grush, Rick (1997) The architecture of representation. *Philosophical Psychology* 10(1):5-25. Reprinted in *Philosophy and the Neurosciences: A Reader*. Bechtel, W., Mandik, P., Mundale, J., and Stufflebeam, R. (Eds.) Oxford: Basil Blackwell.

Grush, Rick (2000). Self, world and space: the meaning and mechanisms of ego- and allocentric spatial representation. *Brain and Mind* 1(1):59-92.

Grush, Rick (in preparation). *The Machinery of Mindedness*.

Langacker, Ronald (1987, 1991). *Foundations of Cognitive Grammar* (2 volumes). Stanford: Stanford University Press.

Langacker, Ronald (2000). *Grammar and Conceptualization*. The Hague: Walter de Gruyter.

McClelland, Jay, and David Rumelhart (eds, 1986). *Parallel distributed processing: explorations in the microstructure of cognition* (2 vols). Cambridge MA: MIT Press.

McCulloch, W.S. and W.H. Pitts (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics* 5:115-133.

Metzinger, Thomas (1993). *Subjekt und Selbstmodel*. Paderborn: Mentis.

Millikan, Ruth Garrett (1984). *Language, thought, and other biological categories*. Cambridge MA: MIT Press.

Millikan, Ruth Garrett (1989). Biosemantics. *Journal of Philosophy* 86(6):281-297.

Newell, Alan, and Herbert Simon (1976). Computer science as empirical inquiry: symbols and search. *Communications of the Association for Computing Machinery* 19:113-126.

O'Keefe, John and J. Dostrovsky (1971). The hippocampus as a spatial map: preliminary evidence from unit activity in the freely moving rat. *Brain Research* 34:171-5.

Port, Robert, and Timothy van Gelder (1995). *Mind as motion: Explorations in the dynamics of cognition*. Cambridge MA: MIT Press.

Putnam, Hilary (1975). *Mind, Language and Reality: Philosophical Papers, Volume 2*. Cambridge: Cambridge University Press.

Quine, Willard van Orman (1960). *Word and object*. Cambridge MA: MIT Press.

Rorty, Richard (1965). Mind-body identity, privacy, and categories. *Review of Metaphysics* 19(1):24-54.

Ryle, Gilbert (1949). *The concept of mind*. London: Hutchinson and Company, Ltd.

Searle, John (1980). Minds, brains and programs. *Behavioral and Brain Sciences* 3:417-424.

Skinner, B.F. (1938). *The behavior of organisms: an experimental analysis*. New York: Appleton-Century.

Skinner, B.F. (1957). *Verbal behavior*. Englewood Cliffs, NJ: Prentice-Hall.

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

Stich, Stephen (1983). *From folk psychology to cognitive science: the case against belief*. Cambridge MA: MIT Press.

Strawson, Peter F. (1959). *Individuals*. New York: Routledge.

Talkmy, Leonard (2000). *Toward a cognitive semantics* (2 vols). Cambridge, MA: MIT Press.

Tolman, E.C. (1938). The Determiners of Behavior at a Choice Point. *Psychological Review*, 45:1-41.

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

Turing, A.M. (1936). On computable numbers, with an application to the Entscheidungs-Problem. *Proceedings of the London Mathematical Society 2nd Series*, 42:230-65.

Turing, A.M. (1950) Computing machinery and intelligence. *Mind* 59:433-460.

Watson, J. (1913) Psychology as a behaviorist views it. *Psychological Review* 20:158-77.

Watson, J. (1924) *Behaviorism*. New York: Norton.

Weiner, N. (1948). *Cybernetics: or, control and communication in the animal machine*. New York: Wiley.