To appear in Fritz Allhoff (Ed.) *Philosophy of*
the special sciences. Albany, NY: SUNY Press

Philosophy of the Cognitive Sciences

William Bechtel and Mitchell Herschbach

Abstract: Cognitive science is an interdisciplinary research endeavor focusing on human cognitive phenomena such as memory, language use, and reasoning. It emerged in the second half of the 20th century and is charting new directions at the beginning of the 21st century. This chapter begins by identifying the disciplines that contribute to cognitive science and reviewing the history of the interdisciplinary engagements that characterize it. The second section examines the role that mechanistic explanation plays in cognitive science, while the third focuses on the importance of mental representations in specifically cognitive explanations. The fourth section considers the interdisciplinary nature of cognitive science and explores how multiple disciplines can contribute to explanations that exceed what any single discipline might accomplish. The conclusion sketches some recent developments in cognitive science and their implications for philosophers.

1. What Is Cognitive Science?

Cognitive science comprises a cluster of disciplines including portions of psychology, linguistics, computer science, neuroscience, philosophy, sociology, and anthropology. Its roots lie in the 1950s, it acquired an academic identity in the 1970s, and continues to thrive in the 21st century. It seeks to explain mental activities such as reasoning, remembering, language use, and problem solving, and the explanations it advances commonly involve descriptions of the mechanisms responsible for these activities. Cognitive mechanisms are distinguished from the mechanisms invoked in other domains of biology by involving the processing of information. Many of the philosophical issues discussed in the context of cognitive science involve the nature of information processing, especially the central notion of representation. One of the distinctive features of cognitive science is that it is not a discipline, but a multi-disciplinary research cluster. It draws upon the contributing disciplines for the problems it investigates, the tools it uses to investigate them, and the explanatory strategies invoked, but the results transcend what is typically achieved in any of the contributing disciplines. This gives rise to philosophical questions about the nature of interdisciplinary research.

The term ‘cognitive science’ was only coined in the mid-1970s. In 1975 it was employed in two books. *Explorations in Cognition*, the product of a collaborative research group at the University of California at San Diego (UCSD), concludes with the suggestion: the “concerted efforts of a number of people from . . . linguistics, artificial intelligence, and psychology may be creating a new field: *cognitive science*” (Norman & Rumelhart, 1975, p. 409). Although situated in psychology, the group employed computational models of semantic networks to explain word recognition, analogy, memory, and semantic interpretation of verbs, sentences, and even brief stories. The collaborative book by computer scientist Daniel Bobrow and cognitive psychologist Allan Collins, *Representation and Understanding: Studies in Cognitive Science*, invoked the term in its subtitle. Two years later the Alfred P. Sloan Foundation announced its Particular Program in Cognitive Science and, over ten years, provided $17.4 million to establish and foster interdisciplinary centers at selected research universities. UCSD received one of the early Sloan grants and used a portion of its funding to sponsor the 1979 La Jolla Conference on Cognition, which became the first meeting of the now international Cognitive Science Society. In 1980 it assumed ownership of the journal *Cognitive Science*, which itself had begun publication in 1977.

While it was during the 1970s that cognitive science began to acquire an institutional identity, its roots go back to the middle of the century when a new intellectual perspective began to inspire researchers in psychology and linguistics to reject the strictures that behaviorism had placed on most research in these disciplines in North America since John Watson (1913) issued his manifesto "Psychology as a Behaviorist Views It" and urged a focus on behavior, not hypothetical mental activities. Although B.F. Skinner, who advocated a radical behaviorism that eschewed mental entities, is perhaps the best known behaviorist, the behaviorist tradition was relatively diverse. Not all behaviorists were opposed to positing events inside the head: Clark Hull (1943) appealed to intervening variables such as drive, but stopped short of overtly mentalistic concepts. Edward Tolman (1948) was exceptional among behaviorists in postulating *cognitive maps* to explain navigational abilities of rats. Leonard Bloomfield (1939) carried behaviorism to linguistics where he advanced a strongly empiricist approach to cataloguing and analyzing linguistic forms and rejected mentalistic accounts of these forms.

While behaviorism cast a broad shadow over psychology and linguistics in North America, in Europe a variety of alternative perspectives more favorable to mentalistic characteristics of human beings prospered and would come to influence the development of cognitive science. For example, Jean Piaget proposed cognitive operations in his genetic epistemology, Frederick Bartlett introduced schemas (organizing structures) to account for memory distortions, Gestalt psychology recast perception in terms of self-organizing forms, and Lev Vygotsky and Alexander Luria initiated studies demonstrating cultural influences on language and thought. Even in North America psychophysics and parts of developmental and social psychology were pursued outside of behaviorism's shadow. But for much of psychology, a revolution was required to reverse behaviorism's proscription on appeals to mental phenomena in explaining behavior.

The cognitivism that emerged in the 1950s maintained behaviorism's emphasis on explaining behavioral phenomena and invoking only behavioral evidence. Hence, unlike earlier mentalistic psychology, it rejected introspection as the avenue to the mind. What it required was a way of conceptualizing internal events that construed them as causal processes contributing to the generation of behaviors. This conceptualization was provided by information theory, which developed from formal engineering analyses of communication channels such as telephones conducted at Bell Laboratories between the 1920s and 1960s. This endeavor attempted to quantify the capacity to transmit information across channels that are subject to capacity and rate limits and to noise. Construing information as a reduction of uncertainty at the end of the channel about the message at the beginning of the channel, Claude Shannon (1948) introduced the *bit* as the unit of information: the unit of information required to differentiate between two equally likely messages. Shannon also showed that one could determine the redundancy in a message in terms of the reduction of uncertainty. George Miller drew upon this analysis in his Harvard Ph.D. dissertation that used redundancy in a message to explain how messages in spoken English could be understood in noisy environments.

In perhaps his best known research, Miller (1956b) identified comparable capacity limitations in a number of cognitive domains, including short-term memory: humans can hold up to seven, plus or minus two, separate items in memory over a period of minutes (unless they are interrupted earlier). In this research, information is construed as the commodity the mind utilizes and the various tasks it performs (remembering, planning, problem solving) as involving the processing of information. Donald Broadbent (1958) advanced a model in which information about sensory events is held briefly in a short term store, and an attentional filter restricts which gets transmitted along a single, limited-capacity channel for further processing. These ideas were incorporated into a general framework for understanding cognitive activity by Ulric Neisser (1967) in his pathbreaking textbook *Cognitive Psychology*.

The idea of the mind as an information processor was further promoted with the introduction of the digital computer as itself an information processor. Shortly after its creation following World War II, some researchers in the new field of computer science began to explore the possibility of programming a computer to behave intelligently (e.g., perform activities that would be judged intelligent if performed by humans). A pivotal conference at Dartmouth College in the summer of 1956 introduced the name *artificial intelligence* and witnessed the first presentation of a program performing intelligently: Alan Newell and Herbert Simon's Logic Theorist, which developed proofs of theorems in symbolic logic.

These contributions were brought together on September 11, 1956, the second day of the second Symposium on Information Theory. Newell and Simon (1956) again reported on Logic Theorist and George Miller (1956a) presented his research on capacity limitations of short-term memory. In between a young linguist, Noam Chomsky (1956), presented a paper "Three Models of Language" in which he argued that various computational systems, such as finite-state automata, were inadequate to model the grammar of human languages and introduced arguments for what he called transformational grammars. These employed procedures for generating core linguistic structures (trees) and transformations to modify these structures. For Miller, on that day "cognitive science burst from the womb of cybernetics and became a recognizable, interdisciplinary adventure in its own right" (Miller, 1979, p. 4). The interdisciplinary interaction that day between psychology, artificial intelligence, and linguistics became characteristic of cognitive science, although as discussed above, 20 years were to pass before the name was introduced and the field became institutionalized.

From these beginnings, research in cognitive science has burgeoned. We can here note just a few landmarks that provide an indication of the breadth of the field. Newell and Simon (1972) introduced the idea of a production system, consisting of a working memory and operations (productions) designed to respond and alter the contents of that memory and employed it to model the strategies humans use to solve problems. Chomsky (1957) developed the first of several grammatical theories (minimalism is the most recent; see Chomsky, 1995). Chomsky elicited an opposition movement which rejected the autonomous status Chomsky claimed for syntax and interlaced syntax with semantics (resulting in what Harris, 1993, characterizes as the linguistic wars of the 1960s and 1970s). More recently, cognitive linguistics has emphasized how other cognitive processes such as spatial representation (Fauconnier, 1994) or metaphors grounded in the body (Lakoff & Johnson, 1999) serve as the basis for linguistic structures. Research in the field of memory that started with the distinction between short- and long-term memory has expanded as researchers have distinguished different forms of long-term memory and distinctive features of how each are processed (Schacter, 1996; Tulving, 2002).

2. Explanation in Cognitive Science

The practitioners of the various cognitive sciences generally construe themselves as engaged in explaining the behavior of human agents. This raises the question of what sort of explanation suffices to explain behavior. Although a number of humanists have contended that the mind must be understood in different terms than other physical systems, cognitive scientists have tended to view their enterprise as contiguous with those of the other natural sciences, especially biology. Traditional philosophical accounts of explanation have construed laws of nature as central, with explanation involving the demonstration that the event to be explained occurred in accordance with laws. On the deductive-nomological (D-N) model, such demonstration involved the derivation of a description of the event to be explained from a statement of the law and initial conditions (Hempel, 1965). A law on this account is minimally a true universally quantified conditional statement which supports inferences about counterfactuals (e.g., inferences about what would happen if the conditions specified in the antecedent were true in a given situation).

The D-N account, however, fares poorly in characterizing the explanations biologists and cognitive scientists offer. The central problem with applying the D-N account to research in cognitive science is the paucity of acknowledged laws within the fields of cognitive science. Perhaps, though, there are laws without their being referred to as such. Indeed, as Cummins (2000) has maintained, psychologists often speak of effects where other scientists might refer to laws. Thus, one finds references to the spacing effect (Ebbinghaus, 1885), the serial position effect (Glanzer & Cunitz, 1966), and the Garcia effect (Garcia, McGowan, Ervin, & Koelling, 1968), where each of these provides a generalization about what happens under specified conditions. But, as Cummins also shows, these effects do not provide explanations but rather serve to identify the phenomena in need of explanation. Thus, the spacing effect is the phenomenon that retention is greater when learning is spaced out in multiple learning episodes, rather than compressed (as in cramming for an exam)—a feature of memory encoding that calls out for explanation.

In biology, there is a similar paucity of acknowledged specifically biological laws—a textbook or research report might refer to laws (or equations) from physics and chemistry but not ones specific to biology. A few philosophers have recently followed the lead of biologists themselves, who commonly appeal to mechanisms as providing explanations. These philosophers have attempted to explicate the nature of mechanistic explanation and how it figures in biology. Although they vary in the vocabulary used to characterize mechanisms, the basic idea is that a mechanism consists of component parts which perform different operations and that these parts are so organized and the operations orchestrated that the whole mechanism, in the appropriate context, realizes the phenomenon of interest (Bechtel, 2006; Bechtel & Richardson, 1993; Craver, 2007; Darden, 2006; Machamer, Darden, & Craver, 2000). Thus, to explain how a cell makes protein, one identifies the various components of the cell that are involved (DNA, mRNA, RNA polymerase, ribosomes, etc.), the operations each of them performs (RNA polymerase creates an mRNA strand from a DNA template), specifies the organization of the parts and shows how the various operations are orchestrated to produce a protein.

Appeals to mechanisms to provide explanations are equally ubiquitous in the cognitive sciences, and philosophers have begun to analyze the mechanistic models offered in research on vision and memory (Bechtel, 2008) and emotion (Thagard, 2006). Memory researchers, for example, have both differentiated memory operations and developed accounts of how they are related. For example, through mental rehearsal an individual can retain for short periods a small number of separate items (e.g., a list of names of people). But humans can also retain for long periods knowledge of facts (e.g., the dates of World War I—a semantic memory) and have the ability to re-experience events in their own lives (e.g., arguing with an officer who gave them a traffic citation—an episodic memory). Explanations of memory processes specify what brain areas and mental operations are involved in, for example, encoding new semantic memories and how they are organized. Thus, on one popular account, for several weeks or months after initial learning, information is encoded in the hippocampus, which then causes changes in regions of the cerebral cortex where the information is maintained for long periods (McClelland, McNaughton, & O'Reilly, 1995). A successful mechanistic explanation then explains how it is that humans are able to exhibit the various mental phenomena that they do.

3. The Distinctive Role of Representations in Cognitive Science Explanations

A difference between many biological mechanisms and cognitive mechanisms is that rather than being concerned with the transformation of materials (e.g., putting amino acids together to constitute proteins), cognitive mechanisms are involved in using information to regulate behavior. Thus, cognitive mechanisms are commonly characterized as *information-processing mechanisms*. The core idea is that states in the head stand in for phenomena outside the head and that by operating on those internal states agents coordinate their behavior with events in the outside world. The states in the head are construed as *re-presentations* of the phenomena outside the head. Consider how you are able to cook a meal from a memorized recipe (or, a bit more challenging, how good cooks can modify memorized recipes to create new dishes). The prototypical cognitive approach treats your knowledge of the recipe as a set of representations in your head and explains your behavior by positing causal processes operating on these representations. The challenge for cognitive science is to characterize these representations more precisely and identify the operations performed on them. There are differing views in cognitive science as to how to meet this challenge.

The idea that the mind trades in representations has roots in the history of philosophy. An innovation of the cognitive revolution was its treatment of the brain, a physical system, as a representational system. One inspiration for the crucial idea that the mind uses representation is that human culture has developed a number of systems used to represent phenomena. The one initially most influential in cognitive science was natural language: we use spoken and written words to communicate with each other because words and the sentences composed from them represent things. But humans operate with a variety of non-linguistic representational systems as well: maps, diagrams, pictures, and so on. Using such external representational systems as models, cognitive scientists posited that states in our heads could similarly be understood as representing things outside the head.

It is important to note, however, that these culturally created external representational systems do not function independently of human beings—if a sandstorm left a tracing in sand on the Martian surface with the shape “Stay out,” that would not be a representation as it was neither constructed by human beings nor processed by them. When “Stay out” is printed on a fence here on Earth, it was the fact that it is created and interpreted by human beings that makes it a representation. When cognitive science proposes to incorporate representations in the head as part of the explanation of how we perform cognitive tasks, including the task of interpreting external representations, the question is how states inside the head constitute representations. It would not help to posit a homunculus (i.e., a little person) inside the heads of humans to interpret these internal representations, since that only recreates the problem of explaining the cognitive abilities of the homunculus.

Issues such as this underlie ongoing debates in cognitive science and the philosophy of cognitive science over what makes something a representation and what kinds of representations are required to model cognition. We will present the major accounts of: representational *vehicles*, or the kinds of structures that serve as representations; the types of operations that are performed on these structures; and how the vehicles acquire their *content*, or meaning.

The primary inspiration for one approach to the first two issues emerged from the development of digital computers. As Newell and Simon (1976) put it, computers are “physical symbol systems”: they are machines which process information by producing meaningful changes in representations or symbols. The crucial feature of computers that makes this possible is that structures which count as symbols in the computer are composed and transformed via formal or *syntactic* rules—i.e., rules which only concern the physical form of symbols, rather than their meaning or *semantics*. These rules are themselves embodied in physical states in the computer and the manipulations performed on these states mirror the relations among the objects represented. The inspiration, which played a foundational role in the development of Artificial Intelligence (AI), is that by following purely formal rules, a computer can manipulate symbols in a manner that would count as intelligent reasoning if performed by a human being. Consider a simple addition function: taking the complex input symbol '3+5' (i.e., the concatenation of '3', '+', and '5') and producing the symbol '8' as output. A computer can do this by applying a formal rule indicating that input strings of one physical type should produce outputs of another physical type. The computer need not understand the meaning of the symbols (e.g., that '3' means the number three) or the function being computed (addition); it need only apply the rote procedure characterized by the syntactic rules. By being an information-processing device, the digital computer thus provided a model for how human cognition could be explained in terms of representational processes. The mind was the "program" or "software" running on the "hardware" of the brain.

The physical symbol systems developed by Newell and Simon and other pioneers in AI employed representations modeled on linguistic representations. In applying this model to humans, Fodor (1975) proposed that thinking occurred in a "language of thought" in which, as in natural languages like English, or formal languages like first-order logic, representational vehicles of cognition are sentences constructed from representational atoms (symbols) in accordance with a combinatorial syntax. In these "classical" cognitive architectures, cognitive processes such as planning and reasoning involve the serial manipulation of sentential representations according to syntactic rules, much as in formal logic, proofs are constructed through sequential transformations of sentential representations.

The idea that cognitive activities involve formal operations upon symbols was also developed in other domains of cognitive science. To account for the productivity of language with a finite set of principles, Chomsky (1957) advanced transformational grammars in which sentential structures are created using rewrite rules to which transformations are then applied. For a simple illustration, the rewrite rules S→NP+VP (a sentence can consist of a noun phrase and a verb phrase) and VP→V+DO (a verb phrase can consist of a verb and a direct object) could generate "Susie loves Charlie." A transformational rule could then be applied to replace *Charlie* with *whom*, and then move *whom* to the front, to yield the question "Whom does Susie love?" Psychologists were also attracted to symbol processing models. John Anderson and Gordon Bower (1973), for example, developed a model of human associative memory which provided the basis for Anderson's subsequent attempts to develop a model of the mind that could account for a broad range of cognitive abilities (Anderson, 2007).

In relying on the computer as a model of a physical symbol system, symbolic accounts tended to abstract away from the physical details of the brain. Following the computer metaphor for cognition, these accounts are at the "software" or "program" level of description, rather than at the level of physical implementation. Although in the last twenty years a number of symbolic theorists, including both Anderson and Newell (1990), have tried to render their accounts more neurally plausible, other researchers from the very beginnings of cognitive science were attracted to models inspired by the physical structure of the brain. These cognitive scientists investigated how units that send activation signals to each other, modeled loosely on the neurons and neural pathways of the brain, could process information. Warren McCulloch and Walter Pitts (1943), for example, showed how networks of artificial neurons could implement logical relations while Frank Rosenblatt (1962) explored the capacity of two-layer networks he called *perceptrons* to recognize perceptual patterns. Rosenblatt also introduced a procedure whereby perceptrons could learn to do this. Marvin Minsky and Seymour Papert's (1969) demonstration of the limitations of perceptrons temporarily sidetracked this approach, but the discovery of a learning procedure for multi-layer networks, which do not face the same processing constraints, rejuvenated it in the 1980s. Since it was the weighted connections between artificial neurons that determined the information-processing abilities of such networks, the movement that emerged using such networks to model cognitive processes (Rumelhart & McClelland, 1986; Bechtel & Abrahamsen, 2002) came to be known as *connectionism*.

Whereas language has made it obvious how to construe symbols as representational vehicles, it is less obvious how to identify representations in connectionist networks. One strategy is to treat each unit as playing a representational role, with its degree of activation serving as a measure of the degree to which it is construed as present in the pattern presented. But a far more interesting approach involves "distributed" representations, in which the representational vehicles are the patterns of activation across a set of units. The same units can figure in multiple vehicles and thereby represent the relations between representations. Cognitive processes are then identified with changes in the network's activation patterns as activity spreads through the network, rather than the application of syntactic rules as on the sentential approach. Learning, as noted above, occurs as the network alters the connections between units, rather than the acquisition of rules or programs. The distributed nature of connectionist representations accounts for some of the benefits of connectionist networks over classical architectures, such as their ability to generalize and their gracefully degrading performance in response to noisy input or the loss of units (conditions which typically cause catastrophic failure in classical architectures).

Many connectionists view successful connectionist models as providing reason to reject the idea that cognition involves sentential representational vehicles. Critics of connectionism, on the other hand, argue that there are limits to connectionist models that can only be overcome by invoking syntactic rules operating over sentence-like representations. But not everyone sees connectionist networks and sentential accounts as incompatible. Some theorists propose that connectionist networks implement symbolic architectures: that a network can be described at more abstract level of analysis as a classical architecture operating on sentential representations. This enables researchers to take advantage both of the syntactic operations available in classical architectures and the generalization and graceful degradation of connectionist networks.

Whatever representational vehicle researchers employ in their cognitive models, an account must also be provided of how these structures come to represent things, how they acquire their content or meaning. Otherwise they are meaningless structures (an objection pressed against classical architectures by Searle, 1980, in his Chinese Room argument). This question of how to account for meaning has mainly been addressed by philosophers, rather than cognitive scientists themselves. The major accounts include appeals to causal/informational relations, teleology, functional role, and resemblance.

When talking about representational vehicles, we can distinguish between *types* of vehicles, and concrete instances of a vehicle type, which are called *tokens*. A type of representational vehicle may be defined by, e.g., a certain kind of physical structure, so particular entities exhibiting this structure would count as tokens of that representation. The appearance of a representation token in a particular cognitive system is sometimes described as the vehicle type being "tokened" in the system. The type-token distinction applies to all kinds of cognitive architectures. In classical architectures with sentential vehicles, symbol types may be defined by physical shape, so tokens of a symbol would be physical entities with that shape. In connectionist networks, one can distinguish between a type of activation pattern and particular instances of that pattern. Theories of content thus address how tokens of different vehicle types acquire their meaning.

One possibility is that a vehicle represents what it is caused by—e.g., smoke (the vehicle) means fire because smoke is caused by fire. This is the basic idea behind one construal of information (see Dretske, 1981): a certain type of representational vehicle would represent or carry information about, say, cats if cats reliably cause vehicles of that type to be tokened in the system. But causal/informational relations alone fail to account for some important features of representations: that we can represent non-existent objects (which could not cause representations to be tokened), and can misrepresent things (as when a representation is caused by something that it does not represent). Further, all kinds of things carry information about their causes without representing those causes (e.g., a gun's firing does not represent its trigger being pulled).

Some theorists have tried to supplement causal/informational accounts with other factors to provide an adequate account of representational content (Cohen, 2004). For example, teleological theories propose that something is a representation when it has been selected (by evolution or learning) for the function of carrying information about something in the world (Millikan, 1984; Dretske, 1995). This means that if a representation is selected to carry information about cats, then it will still represent cats even if on a particular occasion its tokening is caused by a dog. Jerry Fodor's (1987) asymmetric dependence account offers a different way of supplementing causal/informational relations. It claims that vehicles represent only one of the many things they carry information about—namely, the one which is causally responsible for that vehicle's carrying information about the other things. If a symbol carries information about both cats and dogs-seen-at-night, but it does the latter because it does the former, but not the reverse, then the symbol represents cats.

Critics of the above accounts often contend that there is another factor that figures in determining content that these accounts leave out—the functional role of a representation in a cognitive system. In part this role involves the relations a representation has to other representations (Block, 1986). Insofar as the functional role of one representation depends on relations to other representations, and these to yet other representations, functional role accounts are *holistic*. This has spurred the objection that since all representations are related to others, one cannot acquire representations one at a time (Fodor & Lepore, 1992). In contrast, causal/informational theories are *atomistic*, since each vehicle's content is determined independently of other vehicles, and thus can be learned separately.

So far we have followed the mainstream of the debates, which have treated linguistic representations as the prototypical representational vehicle. Relatively early in the development of cognitive science, however, other theorists focused on mental images as representational structures, where images are viewed as more pictorial in nature. While sentences have a linear order, the spatial properties of the vehicle are not really doing the representational work—this is done by the language's combinatorial syntax. In contrast, pictures are representational vehicles which make use of their spatial properties to represent the spatial layout of objects. Roger Shepard and Jacqueline Metzler (1971) showed that in answering questions about whether one object was a rotated version of another, the time required corresponded to the degree of rotation. This suggested people performed a rotation-like operation on mental images of the objects. Stephen Kosslyn (1980) offered evidence that people can scan, zoom and rotate their representations just as we do pictures in the world. Since clearly we do not have pictures in our brains, these accounts have explained our mental imagery in terms of the processing mechanisms that our brain uses to process sensory information. Thus, in constructing and reasoning with a visual image, on these accounts, we use our visual system, driven not by visual input but by top-down processes, a proposal that has received support from neural-imaging studies (Kosslyn, 1994).

Recently a number of cognitive scientists have appealed to our capacities for sensory representation to ground an account of our conceptual capacities (Barsalou, 1999; Mandler, 2004). On these views, language-like representations are not the primary tools of thought, but rather language is a secondary tool for indexing and manipulating those representations. One particularly intriguing way of developing this idea, adumbrated initially by Kenneth Craik (1943) and developed more recently by Jonathan Waskan (2006), is that our representational vehicles are like scale models of things in the world. Just as we use physical models of airplanes in wind tunnels as representations of real airplanes, the brain is thought to operate with scale models structurally isomorphic to what they represent. Whereas the sentential representations used in classical models require separate data structures explicitly indicating how they can be manipulated so as to maintain the semantic relation to what they represent (i.e., syntactic rules), images and scale models are claimed to be structured appropriately such that changes in these representational vehicles automatically mirror changes in the represented system.

Images and scale models introduce a different sort of vehicle than found in classical symbolic models. While there are plausible ways to implement images and scale models in connectionist models, they represent a specific way of employing the connectionist framework—just as in implementing a classical architecture in a connectionist network, researchers need to constrain their networks to implement vehicles that serve as images or scale-models. However images or scale-models are implemented, they provide a distinctive way of approaching the content issue: resemblance relations or isomorphisms between vehicles and content (Cummins, 1996; Waskan, 2006). The intuitive appeal of resemblance accounts can be seen in the case of pictures. Pictures seem to represent because they share some of the physical properties of what they picture, such as color. Appealing to such “first-order” isomorphisms between individual objects and individual representations is, however, quite limited: the brain does not share, for example, the color and shape of objects it represents. Appealing to “second-order” isomorphisms—i.e., relations between the relations among various worldly objects and the relations among the associated representations—are a much better option for resemblance theories. Consider maps: although a point on a map bears little resemblance to a location in the world, the distance relations between the points on a map do resemble the distance relations between locations in the world. Such second-order isomorphisms have been found to be a common way brain areas are organized—e.g., the spatial topology of primary visual cortex resembles the spatial topology of the visual field.

Currently there is no consensus about which if any of these accounts of vehicles, processing, and content are correct and vigorous discussions are continuing among both cognitive scientists and philosophers. At the same time, though, a radically alternative perspective has emerged that calls into question the reliance on representations in cognitive science. Some antirepresentationalists instead advocate characterizing cognition in terms of the mathematics of dynamical systems theory (Port & van Gelder, 1995). Others emphasize the coupling of our brains with our bodies and our world in ways that do not depend upon building up internal models (Clark, 1997). This is to view brains not as representing the world, but as dynamically coupled with the body and extra-bodily environment. It is controversial, however, whether dynamical and situated accounts are incompatible with the mental system invoking representations in its engagement with the world (although what is represented may be different when one focuses on how cognitive agents couple with things in their world). Among those advocating a dynamical or situated perspective, some have proposed treating the brain, body and world as extended cognitive systems, with representations propagating across these various representational media (Clark & Chalmers, 1998; Hutchins, 1995).

4. Relations between Disciplines

Insofar as cognitive science presents itself as interdisciplinary, it is important to consider how disciplines can integrate. Much of the philosophical discussion of interdisciplinary relations has focused on the question of reduction and the particular model of reduction advanced by philosophers in the mid-20th century. Using this model, philosophers have debated whether the theories of one cognitive discipline, psychology, reduce to those of another, neuroscience (Churchland, 1986; Fodor, 1974). On the theory reduction model (Nagel, 1961), theories are construed as linguistic statements, ideally organized in an axiomatic form, and reduction involves the derivation of the theories of the reduced science from those of the reducing science. One of the requirements of a valid reduction is that common vocabulary is used in the premises and the conclusion. Since this is typically not the case in the relation between neuroscience and psychology, additional premises are required, bridge principles that relate the vocabulary of one discipline to that of the other. Much of the controversy over the reducibility of psychology to neuroscience has turned on the issue of whether appropriate bridge principles can be generated. Fodor (1974) argues that bridge principles are not possible since the concepts used in neuroscience group phenomena in very different ways than those employed in psychology. One version of the argument appeals to the claim that psychological states can be realized in very different types of brains, in which very different states realize the same psychological predicates (Putnam, 1967; a claim questioned by Bechtel & Mundale, 1999).

The theory reduction account fails, in many ways, to characterize the interactions between disciplines that are characteristic of cognitive science, such as between psychology and AI, linguistics, and philosophy. Although theories are sometimes the point of engagement, many of the engagements go beyond theories, and involve the utilization of techniques of inquiry from different disciplines and the combining of explanatory approaches (e.g., computational modeling from AI in psychology). Finally, it even fails to capture the sort of relations one actually finds between psychology and neuroscience that are characterized as reductionist, since the focus is seldom on deriving one body of theory from another, but rather on working out a mechanism responsible for a phenomenon. Insofar as a mechanism involves both an account of the parts and what they do, and of the organization in the whole mechanism and how it confronts its environment, a mechanistic account is inherently an interlevel account to which research on both lower and higher levels contributes.

Lindley Darden and Nancy Maull (1977) introduced the notion of an interfield theory as an alternative account of how theories can relate the results of inquiries in different fields. On this account, the product of interactions between fields or disciplines is not the logical derivation of the theories of one field from those of another, but the integration of information procured in multiple fields to address a common problem. Interfield theories typically develop when investigators in different fields possess different tools which each provide part of the information needed to address the phenomenon of interest. The quest to draw and coordinate resources from contributing disciplines to explain phenomena of shared interest is characteristic of cognitive science. For example, while linguists have focused on developing grammars that account for the productive features of languages, psychologists have been concerned with the mental representations and psychological processes involved in language production and comprehension (Abrahamsen, 1987). Different tools are needed and employed to formulate and test grammars than to propose and evaluate psychological mechanisms. In addition, AI researchers contribute to trying to understand language by developing computational models, while neuroscience researchers offer evidence of the brain processes employed and the operations each can perform.

From the 1950s until the 1980s, neuroscience has played only a marginal role in cognitive science. Neuroscientists were actively involved in the early interdisciplinary discussions that prefigured cognitive science, but their primary investigatory tools at the time, such as recording from single neurons, could not be employed on human subjects engaged in cognitive tasks. But beginning in the 1980s and especially through the 1990s, new non-invasive neural imaging techniques that could measure brain activity (using blood flow as a proxy) have provided an avenue for linking psychological studies of behavior with information about brain activity. This research is often characterized as *cognitive neuroscience* and differentiated from cognitive science proper, but increasingly these tools are being invoked in cognitive science itself.

Much of cognitive science has focused on cognitive operations detached from affect (reflecting a long differentiation in philosophy between reason and emotion). Recent research in various cognitive science disciplines has challenged this segregation. For example, evidence has amassed that effective moral reasoning requires a proper integration of emotion and reason (Damasio, 1995). Similarly, pathological conditions such as autism appear to involve deficits in both reasoning and emotion. A brief exploration of this research illustrates some of the most exciting interdisciplinary engagements in contemporary cognitive science.

Autism is a developmental disorder characterized by impairments in social interaction and communication, as well as repetitive and stereotyped patterns of behavior (think of the preoccupation with *The People's Court* displayed by Dustin Hoffman's character in the 1988 movie *Rain Man*). While first under the purview of developmental psychology, autism has become the focus of study for a number of disciplines (see Volkmar, Paul, Klin, & Cohen, 2005). One issue has been to characterize more precisely autism's symptoms, particularly its social deficits. Developmental psychologists and neuropsychologists work to construct improved behavioral measures to capture the spectrum of deficits found in people with autism. Given more precise descriptions of the behavioral phenomena, researchers offer theories of the cognitive operations impaired in people with autism. Some propose that autism involves deficits in "executive functions" (skills such as planning, inhibition, and cognitive flexible); others point to problems of "weak central coherence" (the general ability to integrate pieces of information into coherent or meaningful wholes). One of the most prominent theories explains autism's social deficits in terms of an impaired "theory of mind": that autistic individuals cannot conceive of people's mental states, and thus cannot use this information to guide their social interactions.

While behavioral experiments are used to investigate the predictions of these explanatory proposals, researchers have increasingly turned to neuroscience to determine how the brains of autistic people differ (anatomically, functionally, developmentally) from those of unimpaired people and people with other psychological disorders. A popular neurobiological theory claims autism involves a dysfunctioning "mirror neuron system" (Williams, Whiten, Suddendorf, & Perrett, 2001). Mirror neurons, which were first discovered in primates, fire both when an agent acts and when they observe other agents' actions. An analogous neural system has been found in humans in the pars opercularis of the inferior frontal gyrus, and is proposed to account for, among other things, our understanding of people's intentions, emotions, and other mental states. For example, Mirella Dapretto et al. (2006) argue that we normally understand the emotional significance of people's facial expressions because the mirror neuron system, in concert with the limbic system, causes this emotion to be "mirrored" in us; in this way we “feel” and accordingly understand the perceived person’s emotion. Based on an fMRI study showing little activity in the mirror neuron systems of autistic children when imitating or perceiving emotional facial expressions, Dapretto et al. suggest they do not experience the emotions of perceived others in the way unimpaired people do. This proposal thus explains autism’s deficits in social understanding and interaction in terms of the inability to automatically experience the emotions and other mental states of social interactants.

As autism research shows, explanations in cognitive science proceed using tools from a variety of disciplines. These tools are brought to characterize the parts, operations, and organization of the cognitive mechanisms underlying our mental abilities.

5. Conclusion

Cognitive science is not static. The interdisciplinary project continually identifies new problems and develops solutions for solving them. At the same time, the scope of cognitive science has been expanding. We briefly note a few of these developments and some of the implications they have for philosophical discussions of cognitive science.

As we have noted, as new tools for studying brain processing have developed, cognitive scientists have become increasingly concerned with how cognitive operations are implemented in the brain. How to incorporate information about neural processing poses challenges for both classical and connectionist modeling in cognitive science. Insofar as cognitive models focus on the mental activity, they must to some degree abstract from the neural detail. This frames a philosophical question about how much can they can abstract from details of neural processing and still claim to provide accounts of how humans process information. A related issue involves cognitive science’s traditional reliance on logical and heuristic techniques to model reasoning. Increasingly, both cognitive and neuroscience researchers are advancing probabilistic models, (Chater, Tenenbaum, & Yuille, 2006). For philosophers, this raises the question of whether such models should replace more traditional cognitive models, or the two kinds of models can be constructively related.

Another major new direction in cognitive science is the concern with the embodied, situated nature of cognition. While traditional cognitive science focused solely on what is going on in the heads of cognizers, recent theorists have argued that the non-neural body and environmental context are not merely inputs to the cognitive system but play a constitutive role in cognition (e.g., Clark, 1997). Some of those championing more attention to the organism’s body and environment have appealed to a previously untapped philosophical tradition known as phenomenology (comprising writers such as Husserl, Heidegger, and Merleau-Ponty) for insight about these issues (e.g., Wheeler, 2005)). When the focus is on the real-time responses of organisms to their environment, the temporal dynamics of cognitive processes is obviously important and has been emphasized by those advocating use of dynamical systems tools for understanding cognition (Port & van Gelder, 1995). Many advocates of applying dynamical systems theory to cognition have, as we noted, also argued against the reliance on representations in cognitive models. Others, such as Rick Grush (2004) have tried to show how control theory, a dynamical approach, employs neural representations in accounting for motor control, and to extend this approach to other cognitive processes. These and other current debates in cognitive science provide rich opportunities for continued philosophical engagement with cognitive science.

References

Abrahamsen, A. A. (1987). Bridging boundaries versus breaking boundaries: Psycholinguistics in perspective. *Synthese*, *72*(3), 355-388.

Anderson, J. R. (2007). *How can the human mind occur in the physical universe?* Oxford: Oxford University Press.

Anderson, J. R., & Bower, G. H. (1973). *Human associative memory*. New York: John Wiley and Sons.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577-660.

Bechtel, W. (2006). *Discovering cell mechanisms: The creation of modern cell biology*. Cambridge: Cambridge University Press.

Bechtel, W. (2008). *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. London: Routledge.

Bechtel, W., & Abrahamsen, A. (2002). *Connectionism and the mind: Parallel processing, dynamics, and evolution in networks* (Second ed.). Oxford: Blackwell.

Bechtel, W., & Mundale, J. (1999). Multiple realizability revisited: Linking cognitive and neural states. *Philosophy of Science, 66*, 175-207.

Bechtel, W., & Richardson, R. C. (1993). *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton, NJ: Princeton University Press.

Block, N. (1986). Advertisement for a semantics for psychology. In P. French, T. Uehling & H. Wettstein (Eds.), *Midwest Studies in Philosophy* (Vol. 10, pp. 615-678). Minneapolis, MN: University of Minnesota Press.

Bloomfield, L. (1939). *Linguistic aspects of science*. Chicago: University of Chicago Press.

Broadbent, D. (1958). *Perception and communication*. London: Pergamon Press.

Chater, N., Tenenbaum, J. B., & Yuille, A. (2006). Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences, 10*(7), 287-291.

Chomsky, N. (1956). Three models for the description of language. *Transactions on Information Theory, 2*(3), 113-124.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1995). *The minimalist program*. Cambridge: MIT Press.

Churchland, P. S. (1986). *Neurophilosophy: Toward a unified theory of mind-brain*. Cambridge, MA: MIT Press.

Clark, A. (1997). *Being there: Putting brain, body, and world together again*. Cambridge, MA: MIT Press.

Clark, A., & Chalmers, D. (1998). The extended mind. *Analysis, 58*, 10-23.

Cohen, J. (2004). Information and content. In L. Floridi (Ed.), *Blackwell guide to the philosophy of information and computing* (pp. 215-227). Oxford: Blackwell.

Craik, K. (1943). *The nature of explanation*. Cambridge: Cambridge University Press.

Craver, C. (2007). *Explaining the brain: What a science of the mind-brain could be*. New York: Oxford University Press.

Cummins, R. (2000). "How does it work?" versus "what are the laws?": Two conceptions of psychological explanation. In F. Keil & R. Wilson (Eds.), *Explanation and cognition* (pp. 117-144). Cambridge, MA: MIT Press.

Damasio, A. R. (1995). *Descartes' Error*. New York: G. P. Putnam.

Dapretto, M., Davies, M. S., Pfeifer, J. H., Scott, A. A., Sigman, M., Bookheimer, S. Y., et al. (2006). Understanding emotions in others: Mirror neuron dysfunction in children with autism spectrum disorders. *Nature Neuroscience, 9*, 28-30.

Darden, L. (2006). *Reasoning in biological discoveries: Essays on mechanisms, interfield relations, and anomaly resolution*. Cambridge: Cambridge University Press.

Darden, L., & Maull, N. (1977). Interfield theories. *Philosophy of Science, 43*, 44-64.

Dretske, F. I. (1981). *Knowledge and the flow of information*. Cambridge, MA: MIT Press/Bradford Books.

Dretske, F. I. (1995). *Naturalizing the mind*. Cambridge, MA: MIT Press.

Ebbinghaus, H. (1885). *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Leipzig: Duncker & Humblot.

Fauconnier, G. (1994). *Mental spaces: Aspects of meaning construction in natural language*. Cambridge: Cambridge University Press.

Fodor, J. A. (1974). Special sciences (or: the disunity of science as a working hypothesis). *Synthese, 28*, 97-115.

Fodor, J. A. (1987). *Psychosemantics: The problem of meaning in the philosophy of mind*. Cambridge, MA: MIT Press.

Fodor, J. A., & Lepore, E. (1992). *Holism: A shopper's guide*. Oxford: Blackwell.

Garcia, J., McGowan, B. K., Ervin, F. R., & Koelling, R. A. (1968). Cues: Their relative effectiveness as a function of the reinforcer. *Science, 160*, 794-795.

Glanzer, M., & Cunitz, A. R. (1966). Two storage mechanisms in free recall. *Journal of Verbal Learning and Verbal Behavior, 5,* 351-360.

Grush, R. (2004). The emulation theory of representation: Motor control, imagery, and perception. *Behavioral and Brain Sciences, 27*, 377-396.

Harris, R. A. (1993). *The linguistics wars*. New York: Oxford.

Hempel, C. G. (1965). Aspects of scientific explanation. In C. G. Hempel (Ed.), *Aspects of scientific explanation and other essays in the philosophy of science* (pp. 331-496). New York: Macmillan.

Hull, C. L. (1943). *Principles of behavior*. New York: Appleton-Century-Crofts.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Kosslyn, S. M. (1980). *Image and mind*. Cambridge, MA: Harvard University Press.

Kosslyn, S. M. (1994). *Image and brain: The resolution of the imagery debate*. Cambridge, MA: MIT Press.

Lakoff, G., & Johnson, M. H. (1999). *Philosophy in the flesh: The embodied mind and its challenge to Western thought*. New York: Basic Books.

Machamer, P., Darden, L., & Craver, C. (2000). Thinking about mechanisms. *Philosophy of Science, 67*, 1-25.

Mandler, J. M. (2004). *The foundation of mind: Origins of conceptual thought*. Oxford: Oxford University Press.

McClelland, J. L., McNaughton, B., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*(3), 419-457.

McCulloch, W. S., & Pitts, W. H. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 7*, 115-133.

Miller, G. A. (1956a). Human memory and the storage of information. *Transactions on Information Theory, 2*(3), 129-137.

Miller, G. A. (1956b). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review, 63*, 81-97.

Miller, G. A. (1979). *A very personal history* (Occasional paper No. 1). Cambridge, MA: Center for Cognitive Science.

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

Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press.

Nagel, E. (1961). *The structure of science*. New York: Harcourt, Brace.

Neisser, U. (1967). *Cognitive psychology*. New York: Appleton-Century-Crofts.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Newell, A., & Simon, H. A. (1956). The logic theory machine: A complete information processing system. *Transactions on Information Theory, 'IT-2*(#3), 61-79.

Newell, A., & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Newell, A., & Simon, H. A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM, 19*, 113-126.

Norman, D. A., & Rumelhart, D. E. (1975). *Explorations in cognition*. San Francisco: Freeman.

Port, R., & van Gelder, T. (1995). *It's about time*. Cambridge, MA: MIT Press.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan & D. D. Merrill (Eds.), *Art, Mind and Religion* (pp. 37-48). Pittsburgh: University of Pittsburgh Press.

Rosenblatt, F. (1962). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. Washington: Spartan Books.

Rumelhart, D. E., & McClelland, J. L. (1986). *Explorations in the microstructure of cognition. Volume 1. Foundations*. Cambridge, MA: Bradford Books, MIT Press.

Schacter, D. L. (1996). *Searching for memory: The brain, the mind, and the past*. New York: Basic Books.

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

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27*, 379-423, 623-656.

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science, 171*, 701-703.

Thagard, P. (2006). *Hot thought: Mechanisms and applications of emotional cognition*. Cambridge, MA: MIT Press.

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

Tulving, E. (2002). Episodic memory: From mind to brain. *Annual Review of Psychology, 53*, 1-25.

Volkmar, F. R., Paul, R., Klin, A., & Cohen, D. J. (Eds.). (2005). *Handbook of autism and pervasive developmental disorders*. Hoboken, NJ: John Wiley.

Waskan, J. (2006). *Models and cognition*. Cambridge, MA: MIT Press.

Watson, J. B. (1913). Psychology as the behaviorist views it. *Psychological Review, 20*, 158-177.

Wheeler, M. (2005). *Reconstructing the cognitive world: The next step*. Cambridge, MA: MIT Press.

Williams, J. H., Whiten, A., Suddendorf, T., & Perrett, D. I. (2001). Imitation, mirror neurons and autism. *Neuroscience and Biobehavioral Reviews*, *25*, 287-295.